KNOWLEDGE EXTRACTION AND ANALYSIS OF MEDICAL TEXT WITH PARTICULAR EMPHASIS ON MEDICAL GUIDELINES

by

Hossein Hematialam

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computer Science

Charlotte

2021

Approved by:

Dr. Wlodek Zadrozny

Dr. Yaorong Ge

Dr. Xi Niu

Dr. Albert Park

Dr. Reza Mousavi

ABSTRACT

HOSSEIN HEMATIALAM. Knowledge extraction and analysis of medical text with particular emphasis on medical guidelines. (Under the direction of DR. WLODEK ZADROZNY)

In this dissertation document, we describe the potential for Information Extraction, Information Retrieval, and Machine Learning methods to improve the process of analyzing medical texts and, in particular, Clinical Practice Guidelines (CPGs). We present the results of three in-depth studies consisting of dozens of experiments on finding condition-action and other conditional sentences in guideline documents. We are improving the state-of-the-art results (up to 25%) and showing for the first time the applicability of domain adaptation and transfer learning to this problem.

We also present new methods for identifying inconsistencies in disagreements between medical guidelines, and for analyzing them using a combination of machine learning, information retrieval, and text mining methods. We show the need for a formal distinction between contradictions and disagreements in natural language texts to formally reason between contradictory medical guidelines.

We introduce new representations for collections of guideline documents and an algorithm for comparing collections of documents. We use these to investigate conceptual distances between guidelines for the same conditions. Throughout this process, we prove the hypothesis that the difference in recommendations largely (by 69% to 86%) correlates with the differences in concepts used by the medical bodies authoring the guidelines.

Finally, we show the applicability of text analysis methods to practical problems of analyzing textual information in electronic health records. We achieved 83% accuracy in matching medical records with a list of pre-defined conditions in an EHR system, resulting in clinical system support changes in one of the leading US hospitals.

ACKNOWLEDGEMENTS

I would like to express appreciation and gratitude to my advisor Dr. Wlodek Zadrozny for his support and mentorship during my Ph.D. study. I am also very grateful to Dr. Yaorong Ge, Dr. Xi (Sunshine) Niu, Dr. Reza Mousavi, and Dr. Albert Park for accepting to be on my doctoral committee and sharing their knowledge with me. I would also like to express my gratitude to Dr. William Tolone, the graduate school at UNC Charlotte, and the Department of Computer Science for their support during my Ph.D. study.

TABLE OF CONTENTS

LIST OF TABLES x

LIST OF FIGURES xiv

LIST OF ABBREVIATIONS xvii

CHAPTER 1: INTRODUCTION 1

1.1. Overview of the Problem Space of this Dissertation 1

1.2. Dissertation Structure 3

CHAPTER 2: PRELIMINARIES: CLINICAL PRACTICE GUIDELINE AND TEXT ANALYSIS METHODS 5

2.1. Clinical Practice Guidelines — the Focus of This Dissertation 5

2.2. Information Extraction and Other Text Analysis Tasks 7

2.2.1. Named Entity Recognition 8

2.2.2. Part of Speech Tagging 8

2.2.3. Parsing 9

2.2.4. Coreference Resolution 10

2.2.5. Discourse/Text Segmentation 11

2.2.6. Sentiment Analysis 12

2.2.7. Terminology Extraction 13

2.2.8. Language Representation Models 14

2.3. Information Retrieval 16

2.3.1. Indexing Process 17

2.3.2. Document and Query Modeling 18

2.3.3. Query Expansion 20

2.3.4. Measures for Information Retrieval 21

2.4. Summary 22

CHAPTER 3: SEMANTIC MODELING OF CONTRADICTIONS AND DISAGREEMENT: A CASE STUDY OF MEDICAL GUIDELINES 23

3.1. Introduction 23

3.2. Formal Representation of Disagreement and Contradiction 24

3.3. Finding Contradictions and Disagreements 28

3.4. Conclusion and Discussion 30

CHAPTER 4: CONCEPTUAL DISTANCES BETWEEN MEDICAL RECOMMENDATIONS: EXPERIMENTS IN MODELING MEDICAL DISAGREEMENT 31

4.1. The Problem and the Method 32

4.1.1. Motivation 32

4.1.2. Brief Description of the Proposed Method 34

4.1.3. Summary of Contributions 37

4.1.4. Organization of the Chapter 37

4.2. Discussion of Prior Art 38

4.2.1. Text Analysis of Medical Guidelines 39

4.2.2. Vector Representations of Documents Using Word Embeddings 40

4.2.3. Other Work on Disagreements and Contradictions 40

4.3. From Recommendations to Vectors of Differences and a Graph 41

4.3.1. Computing the Differences in Recommendations 43

4.3.2. From Differences to Distances and a Graph 43

4.4. Transforming Full Guidelines Documents into Vectors and Graphs 45

4.4.1. Data Preparation for All Experiments 46

4.4.2. Measuring Distances between Full Documents 46

4.4.3. Building Vector Representations of Full Documents 48

4.4.4. Our Best Model: Using BioASQ Embeddings and Word Mover's Distance 48

4.5. Graph-Based Method for Comparing Collections of Documents 49

4.6. Details of Experiments and Their Results 52

4.6.1. Steps Used in All Our Experiments and Evaluation 52

4.6.2. Results of the Experiments 53

4.7. Additional Experiments 55

4.7.1. Experimenting with the Full Texts of the Guidelines 55

4.7.2. Lower Back Pain Management Guidelines 60

4.7.3. Comparing Hypertension Management Guidelines 62

4.8. Discussion 64

4.9. Conclusions 65

CHAPTER 5: IDENTIFYING CONDITIONAL AND CONDITION-ACTION STATEMENTS IN MEDICAL GUIDELINES 67

5.1. Introduction 67

5.1.1. Motivation 68

5.1.2. Organization of this Chapter and Brief Description of the Studies 69

5.2. Preliminaries and Related Work 70

5.2.1. Five Decades of Automated Analysis of Medical Texts 71

5.2.2. Analysis of Medical Guidelines 72

5.2.3. Deep Learning Methods, Domain Adaptation and Transfer Learning 73

5.3. The Data 74

5.3.1. The Dataset of Three Annotated Guidelines 75

5.3.2. The Data from the Perspective of Domain Adaptation and Transfer Learning 77

5.4. Using Syntactic and Semantic Features (Studies 1 and 2) 79

5.4.1. The Feature Set 79

5.4.2. Evaluation Measures and Baseline Results 81

5.4.3. Identifying Conditional Statements Using Semantic Types (Study 2) 82

5.5. Deep Learning and Transfer Methods (Study 3) 85

5.5.1. Deep Learning. Study 3, Experiment 1 86

5.5.2. Deep learning. Study 3, Experiment 2 88

5.5.3. Extracting Conditional Statements using Transfer Learning 90

5.6. Discussion 91

5.7. Conclusions 94

CHAPTER 6: FROM KNOWLEDGE EXTRACTION TO INFORMATION RETRIEVAL AND INSIGHTS 96

6.1. PubMed – a Dominant Paradigm in Medical Search 97

6.2. Related Works 98

6.3. Semantic Medical Guideline Information Retrieval System 99

6.3.1. Data Acquisition and Preparation 100

6.3.2. Processing Tables 101

6.3.3. Training Word Embedding Models 103

6.3.4. Implementing the Search Engine 104

6.4. Applications 105

6.4.1. Semantic Search 105

6.4.2. Case Study: Asthma Medicine 109

6.5. Discussion 114

CHAPTER 7: IMPROVING BLOOD TRANSFUSION MEDICAL RECORDS USING TEXT ANALYSIS 116

7.1. Introduction 116

7.2. Overview of Prior Art 119

7.3. Classification objectives and Process 121

7.3.1. Mapping to Predefined Reasons 122

7.3.2. Detecting Repetitive Reasons 125

7.4. Results 126

7.4.1. Mapping to a Predefined Reason 126

7.4.2. Repetitive Reasons 128

7.5. Discussion and Conclusion 129

CHAPTER 8: SUMMARY OF DISSERTATION, OPEN PROBLEMS, AND FUTURE DIRECTIONS 131

REFERENCES 134

LIST OF TABLES

TABLE 4.1: The table shows recommendations as follows: N—no recommendation; b—both patient and doctor, shared decision; r—recommending mammography. 35

TABLE 4.2: This table shows the number of differing feature values for pairs of guidelines, based on Table 4.1. The Jaccard distances between the documents are obtained by dividing the value in the table by five (the number of features). 36

TABLE 4.3: Normalized distances between the summarized guidelines computed using Jaccard distances from Tables 4.1 and 4.2. 43

TABLE 4.4: Guidelines with references. All the sources were last retrieved in summer 2020. 46

TABLE 4.5: This table shows the word mover's distances between the guidelines using BioASQ embeddings. This model also performed very well on the datasets in the section 4.7. 49

TABLE 4.6: This table shows the values obtained in multiple experiments. Column 2, `Distortion`, shows the distortions of graphs produced using corresponding models from Column 1. Average distortions per permutation are shown in Column 3. `STD` is the standard deviation of the distortion per permutation of vertices. Note that the distortion is somewhat depended on how we measure distances; however, the shapes of the distributions are very similar. (The cosine measures are capitalized for readability). 54

TABLE 4.7: **Using sentences in recommendations and minimum mutual concepts**. This table shows the values obtained in additional experiments, where full document guidelines were modified by attending to concepts in sentences (see above). Column 1 refers to the number of concepts overlapping with summaries. `Distortion` shows the distortions of graphs produced using corresponding models from Column 1. As before, in Table 4.6, the distortion depends on how we measure the distances; however, the shapes of the distributions are very similar. 57

TABLE 4.8: **Using the whole summary recommendations and minimum mutual concepts**. This table shows the values obtained in additional experiments, where the whole CDC summary was used to obtain sets of mutual concepts (see above). Column 1 refers to the number of concepts overlapping with the summary. `Distortion` shows the distortions of graphs produced using corresponding models from Column 1. As before, in Tables 4.6 and 4.7 the distortion is somewhat depended on how we measure distances; however, the shapes of the distributions are very similar. 59

TABLE 4.9: Jaccard distances based on the combined Tables 1 and 2 from [1]. The guidelines are about the management of non-specific lower back pain. 61

TABLE 4.10: The performance of the algorithms on the combined Tables 1 and 2 from [1] is in line with the results in Section 4.6.2, except for the weaker showing of the Conceptualized_WMD model. 62

TABLE 4.11: We see the robustness of the proposed method when comparing the abstracts' conceptual distances and the full documents of guidelines. 64

TABLE 5.1: Examples of classified sentences and their classes/types. 76

TABLE 5.2: Statistical information about annotated guidelines. `Words` – the total number of words in the document. `Avg Length` – average number of words per sentence (applies to all sentences). `CA` condition-action (recommendation); `CC` condition-consequence; `A` action; `NC` no condition (nor action). 76

TABLE 5.3: Classification results on annotated guidelines using only POS tags and their combinations, focusing on the detection of condition-action sentences. These baseline results are the core of Study 1. 81

TABLE 5.4: Summary of classification results on annotated guidelines, using domain independent syntactic features (Study 1), and based on [2]. The CA and CC classes are combined and shown as CCA. 82

TABLE 5.5: Classification results on annotated guidelines, focusing on condition-action sentences and using all features (semantic and syntactic); Study 2. 84

TABLE 5.6: This table shows the classification results on combined condition-consequence and condition-action classes using all features (semantic and syntactic); Study 2. 85

TABLE 5.7: This table illustrates the classification results on identifying condition-action statements (CA) using transformer embeddings as features (Study 3, Experiment 1). In this table, we only report results from the logistic regression classifier, but use different vectorized representations of sentences coming from the models in the first column. 87

TABLE 5.8: This table illustrates the classification results on identifying conditional statements (CCA) using transformer embeddings as features (Study 3, Experiment 1). In this table, we only report results from Logistic Regression classifier. 87

TABLE 5.9: This table illustrates the classification results on identifyng condition-action statements (type CA) using different features (Study 3, Experiment 2). In this table, we only report results from the Logistic Regression classifier. 89

TABLE 5.10: This table illustrates the classification results on identifying conditional statements (type CCA) using different features (Study 3, Experiment 2). In this table, we only report results from the Logistic Regression classifier. 89

TABLE 5.11: Study 3. Experiment 3. On the class of conditional sentence (CCA), 72% F1 and 87% accuracy (A) shows applicability of machine learning transfer; it beats results of Study 2 Table 5.6 of 65%. Syntactic and semantic features from Study 2 were used in the first and third experiments. 90

TABLE 5.12: Study 3. Experiment 3. On the class of condition-action (CA) sentences the 67% F1 score shows the applicability of transfer learning to this class, closely matching the 68% F1 score of Table 5.5. Syntactic and semantic features from Study 2 were used in first and third experiments. 91

TABLE 5.13: This table illustrates the improvements in classification results on identifying condition-action statements. 92

TABLE 5.14: This table illustrates the classification results on identifying conditional statements. 92

TABLE 5.15: This table illustrates the classification results on identifying recommendations, defined in [3] and [4], as CCA+A . This experiment uses as features the embeddings from the transformer models, as previously shown in Study 3, Example 1. 94

TABLE 6.1: Asthma Medicine Products Approved by the U.S. FDA [5] 110

TABLE 7.1: Blood Transfusion Indications from the CPOE. The "Other – ..." fields are the source of free-text data used in our analysis. 118

TABLE 7.2: Classification results on free-text reasons. *Conceptual Match*: How the concepts in a reason match with the concepts from the assigned predifned reason. *Numerical Match*: whether a reason and its assigned predefined reason match or not. *Has a Numerical Condition*: Whether a reason has a numerical condition or not. *Frequency*: reports of the reasons classified based on *Conceptual Match* and *Has a Numerical Condition*. *Mapped Correctly*: The number of reasons from that class which our domain expert agrees that the mapped reason is the preferable choice for a reason. 128

TABLE 7.3: Frequencies of the repetitive concepts for different products in the blood management system. 129

LIST OF FIGURES

FIGURE 2.1: A parsed tree for the following sentence: "If bp is greater than 50, follow the instruction." 10

FIGURE 2.2: Architecture of a general IR system 17

FIGURE 3.1: The architecture used to evaluate extraction of contradictions in medical guidelines. 29

FIGURE 4.1: Note the contradictory recommendations in green and blue boxes. The colors in the table come from [6], but the original table comes from the CDC [7]. Only a part of the table is reproduced here. 33

FIGURE 4.2: The method of comparing concepts in full documents and recommendations contained in summaries. Note the difference in representations: the documents are represented by a large number of high-dimensional (200) vectors with real valued features, whereas the disagreement representations can be low-dimensional vectors with discrete features (e.g., five-dimensional for the breast cancer screening guidelines). Our exposition will roughly follow the left-to-right order of this figure, using the breast cancer screening guidelines as the motivating example. 35

FIGURE 4.3: Similarities and disagreements in summarized recommendations. The yellow coloring shows patient making decisions, the blue coloring shows explicit screening recommendations. The concentric circles show different age groups. Red marks—physician recommends, green marks—patient decides. 42

FIGURE 4.4: In panel (**a**) we see a pictorial representation of the numbers of differing features, per Tables 4.2 and 4.3. These differences between recommendations are converted into distances (using the Jaccard measure), resulting in panel (**b**). Can we replicate the geometric structure of panel (**b**) using automated tools? See Section 4.6 for an answer. 42

FIGURE 4.5: Visual comparison of the similarity/distance graphs based on human analysis is shown in panel (**a**), and computer generated comparison from Table 4.5 is shown in panel (**b**), which suggest a similar geometry. As we rigorously show in Section 4.6, this 69% *similarity is not accidental*; the distortion is about 31%. Notice that we are not pointing to the actual locations of similarities and differences in the guideline documents. Instead, we are pointing to global (latent) differences stemming from concepts appearing in them. 45

FIGURE 4.6: For the graphs of the eight hypertension guidelines and their abstracts a visual comparison is more difficult than it was earlier in Figure 4.5. Therefore, we need a quantitative comparison, which is given in Table 4.11. 63

FIGURE 5.1: Top panel: the frequency distribution on 445 words-in-common in the training data (hypertension+rhinosinusitis; Series 1) and test data (asthma; Series 2). The bottom panel shows the large difference between the two distributions. 78

FIGURE 5.2: Example of the parse tree showing the part-of-speech (POS) features used in Study 1, and syntactic structures used in Study 2. 80

FIGURE 6.1: An example table from (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5409140/) 102

FIGURE 6.2: A simple search engine UI. *Query* can be used for keyword searches. UMLS semantic types and concepts are designed for semantic search. *MeSH* can filter guidelines based on their indexed MeSH type in PubMed. *Query Suggestion* can be used to translate users' queries into UMLS concepts and MeSH terms. 106

FIGURE 6.3: First five statements retrieved for breast cancer screening from evidence-based guidelines. The first retrieved statement is a formulated table information from table caption, first row, first column, and the cell. 107

FIGURE 6.4: Word cloud of *Pharmacologic Substances* based on their frequencies in guidelines tagges as *Asthma* guidelines. 109

FIGURE 6.5: We show a heat map of the frequencies of the drugs in each guideline. Each guideline is represented by "Year::PMID". For all frequencies higher than 10, we used dark green in order to emphasize the appearances of concepts better. 111

FIGURE 6.6: A heatmap of frequencies of a *Pharmacologic Substances* and a disease or a symptom in *asthma* guidelines in our repository. Values more than 10 are being shown in dark green. Each guideline is represented by "Year::PMID". 113

FIGURE 7.1: A proposed architecture for mapping reasons to standardized options (predefined reasons). 122

FIGURE 7.2: A proposed architecture to generate candidates for new standardized options. 126

LIST OF ABBREVIATIONS

AAFP An acronym for the American Academy of Family Physicians.

ACOG An acronym for the American College of Obstetrics and Gynecology.

ACP An acronym for the American College of Physicians.

ACR An acronym for the American college of Radiology.

ACS An acronym for the American Cancer Soceity.

AQE An acronym for Automatic Query Expansion.

Bio-NER An acronym for Biomedical Named Entity Recognition.

BIR An acronym for Binary Independence Retrieval.

CBOW An acronym for Continuous Bag-of-Words.

CDSS An acronym for Clinical Decision Support System.

CPG An acronym for Clinical Practice Guideline.

EHR An acronym for Electronic Health Record.

EMR An acronym for Electronic Medical Record.

FOL An acronym for First Order Logic.

IARC An acronym for the International Agency for Research on Cancer.

IE An acronym for Information Extraction.

IQE An acronym for Interactive Query Expansion.

IR An acronym for Information Retrieval.

MeSH An acronym for Medical Subject Headings.

NCBI An acronym for National Center for Biotechnology Information.

NERC An acronym for Named Entity Recognition and Classification.

NLM An acronym for National Library of Medicine.

PMC An acronym for PubMed Central.

POS An acronym for Part of Speech.

TF-IDF An acronym for Term Frequency-Inverse Document Frequency.

USPSTF An acronym for the United States Preventive services Task Force.

WSD An acronym for Word Sense Disambiguation.

CHAPTER 1: INTRODUCTION

1.1 Overview of the Problem Space of this Dissertation

Natural language understanding (NLU) is one of the great unsolved problems of artificial intelligence. We, humans, have not (yet?) constructed machines capable of engaging in intelligent conversation or in-depth reading of a novel. Even for technical jargons, which do not contain metaphors or poetic comparisons, the NLU problem is still unsolved.

In this dissertation, we look into the problem of understanding medical texts from different perspectives. At this point, there is no single approach that would allow us to create practical models of such texts. For example, being able to do entity and relationship extraction (which is the focus of Chapters 6 and 7) is not sufficient to understand the conceptual relations between full texts of medical documents, such as clinical guidelines (Chapter 4).

We devote a large portion of our research to medical guidelines for several reasons. They directly affect the patients, but they often contradict each other because different professional medical societies focus on different aspects of patient health. Moreover, not much research has been done on understanding the texts of the guidelines, one reason being, the lack of annotated corpora.

Therefore, we look at medical guidelines from several complementary perspectives. As with any issue, we can take top-down and bottom-up approaches, and we do both. We compute conceptual distances between full documents to verify the hypothesis that different training contributes to differences in recommendations (Section 4.3), and we attempt to find the specifics of contradictions between sentences in different guidelines (Section 3.2).

In our experiments, we employ different types of tools. We use Information Extraction tools to address a practical problem of understanding free-text notes from transfusions in a Duke University hospital (Chapter 7) in order to understand adherence to the appropriate clinical guidelines and to modify the structure of the electronic health records system. On the other end of the spectrum, we use very recently developed deep learning techniques to find condition-action sentences in medical guidelines (Chapter 5). We also experiment with applying information retrieval techniques both for fine-grained detection of contradictions (Section 3.3), and to better understand the space of medical guidelines (Chapter 6).

Overall, this dissertation provides a comprehensive look at extraction of different types of knowledge from medical texts, and in particular, from medical guidelines. It also discusses most of the modern machine learning techniques and their applicability to natural language processing in this domain. The dissertation also introduces practical and already implemented methods (Chapter 7), as well as more speculative ones (Chapters 3, 4, and 6), which we hope will eventually influence the field and, among other things, reduce treatment variabilities.

The presented contributions include:

- Three in-depth studies of finding condition-action and other conditional sentences in medical guidelines. Here we improved state of the art up to 25% (Chapter 5).

  As part of these studies, we developed two new annotated guideline documents, which gained the acceptance of other researchers [4, 3]. The newest results are currently in process of being submitted to *BMC Bioinformatics*.

- A novel model of extraction of contradictions and disagreements from medical guidelines (Section 3.3) and a formal model allowing logical reasoning about degrees of disagreement (Section 3.2) (Published in *IWCS 2017—12th Interna-*

tional Conference on Computational Semantics).

- A preliminary study of a semantic search engine (Chapter 6) which should enable non-specialist finding guidelines, and snippets of interest relevant to answering both medical and public health questions, such as changes in drug recommendation or diffusion of medical innovation (on-going work).

- A study, with a Duke University researcher, of text notes from transfusion procedures showing high accuracy of a specially-developed data extraction and classification model, which achieved 95% accuracy in extracting numerical condition and 83% in matching medical records with a list of standardized conditions.

  Furthermore, this study's results have already resulted in changes to the EHR system in a Duke University hospital. (And an article about it is under review in *Journal of Biomedical Informatics*).

- A very novel (and somewhat speculative) model, developed with a researcher at University of Central Florida medical school showing that the authors of medical guidelines should be viewed as epistemological near peers. That is, the difference between guidelines can, to a large extent (69%–86%), be attributed to the difference between the concepts used by different medical specialties. (Published in *Applied Sciences 11, no. 5 (2021)*).

1.2 Dissertation Structure

The remainder of this dissertation is organized as follows:

In Chapter 2, we review the definition of Clinical Practice Guidelines and related systems. We provide a summary of available text analysis methods.

In Chapter 3, we propose a formal representation of contradictions and disagreements in text. We also introduce an architecture for identifying contradictions and disagreements in medical guidelines. We report an experiment of adapting the pro-

posed architecture on finding agreements and disagreements between medical guidelines.

In Chapter 4, we introduce a natural language processing approach to represent medical guidelines as embeddings and a novel graph-based similarity model for comparing collections of documents. We report the evaluation of our approach on three sets of medical guidelines: breast cancer screening, lower back pain management guidelines, and hypertension management guidelines.

Motivated by [8], we address the problem of identifying condition-action statements in Chapter 5. We propose an automated process to identify conditional and condition-action statements from medical guidelines using classical machine learning models and deep learning language representation models. We report the evaluation of the proposed methods on extracting conditional and condition-action statements from three sets of medical guidelines: asthma, hypertension, and rhinosinusitis guidelines.

In Chapter 6, we present a process of creating and indexing medical statements. Our proposed process provides semantic indexing and can handle different formats of textual information like narrative text and tables. We used the proposed process to create a corpus of medical guidelines. We provide some examples on the capabilities of our system.

In Chapter 7, we introduce a semi-automated method for matching free-text elements of medical records in a patient blood management system. We also report our analysis of the repetition of the conditions appeared in the records. This analysis helps to identify new treatment services and new thresholds for blood product orders.

In Chapter 8, we conclude the dissertation by highlighting our contributions and possible future directions.

CHAPTER 2: PRELIMINARIES: CLINICAL PRACTICE GUIDELINE AND TEXT ANALYSIS METHODS

In this chapter, we are going to review medical guideline and NLP analysis and computational linguistic techniques which help us in the process of retrieving and extracting knowledge from medical texts.

Since 2010, on average, more than 1,200 new practice guidelines have been indexed in PubMed [1] each year. There is a considerable number of medical guidelines for each disease. For example, more than 200 guidelines are indexed in PubMed for diagnosis of breast cancer. Besides the large number of available guidelines from different organizations, each guideline is being updated regularly. An automated process of analyzing guidelines is necessary for studying and controlling the impact of medical guidelines on evidence-based medicine and public health, due to the enormous number of medical guidelines available.

We first review the definition of Clinical Practice Guidelines and its role in the decision making process. After that, we will review available NLP analysis and computational linguistic techniques to retrieve and extract knowledge from medical guidelines.

2.1 Clinical Practice Guidelines — the Focus of This Dissertation

The Institute of Medicine (IoM) defined practice guidelines as "systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances" [9]. In 2011, the institute's committee updated the Clinical Practice Guideline(CPG) definition to reflect a better current consensus on what constitutes CPG. In a new definition, CPGs are "statements that include recommendations, intended to optimize patient care, that are informed by a

[1] `https://pubmed.ncbi.nlm.nih.gov/`

systematic review of evidence and an assessment of the benefits and harms of alternative care options" [10]. The Institute of Medicine introduced five dimensions to categorize CPGs [11]:

- Clinical orientation: the main focus of a CPG can be a clinical condition [12], a technology (broadly defined) [13], or a process [14].
- Clinical purpose: CPGs may advise about screening and primary prevention [15, 16], diagnosis [17, 18], treatment and management (including secondary prevention) [19], or more discrete aspects of health care.
- Complexity: CPGs may be relatively straightforward in presentation [15] and discussion or be full of lengthy narrative and documentation, considerable detail, or complex logic [19].
- Format: CPGs can be presented in free-text, tables, charts, or by other means.
- Intended users: CPGs are intended to be used by practitioners [19, 15], patients [20], or others.

Different medical societies develop CPGs. Since they employ experts with different specialties and sub-specialties, different methods, and different evidence, we might see disagreement between guidelines. For example, the American College of Radiology recommends that women Aged 50 to 75 get a mammography annually while the American College of Physicians recommends mammography once every two years for that group. These disagreements might contribute to overdiagnosis or overtreatment since they raise uncertainty.

CPGs are the primary knowledge source for computer-based clinical decision support systems (CDSSs) [21]. CDSSs are "software that is designed to be a direct aid to clinical decision-making in which the characteristics of an individual patient are matched to a computerized clinical knowledge base, and patient-specific assessments

or recommendations are then presented to the clinician and/or the patient for a decision" [22]. Deciding which questions to ask, tests to order, procedures to perform, treatment to indicate, or which alternative medical care to try, are examples of clinical decisions which CDSSs try to answer. CDSSs generally fall into two categories [23] :

- "Determining *what is true* about a patient (usually what the correct diagnosis is)."
- "Determining *what to do* for the patient (usually what test to order, whether to treat, or what therapy plan to institute)."

Most of the questions physicians need to consult with CDSSs about are from the latter category. CPGs are most useful at the point of care [24] and answering *what to do* questions with recommendations.

2.2 Information Extraction and Other Text Analysis Tasks

In this section, we review Information Extraction (IE) as a powerful tool in extracting knowledge from medical texts. Any process that selectively structures and combines data that is found, explicitly stated or implied, in one or more texts is called Information Extraction [25]. IE is different from Information Retrieval (IR) because the IR goal is to retrieve relevant documents, but IE's goal is to extract relevant information. Different goals result in different techniques adopted by these technologies. The combining of IR and IE provides more powerful tools for users to seek information. While IR systems rely on information theory, probability theory, and statistics, IE systems apply NLP analysis and computation linguistic techniques and theories to extract desired information from texts. The input for an IE system can be unstructured data (e.g., free-text), semi-structured data (e.g., tables), or structured data (e.g., HTML pages).

Message Understanding Conferences (MUC) inspired researchers to focus on the development of the IE systems. MUCs introduced several important IE tasks [26]:

named entity recognition, coreference resolution, template element construction, template relation construction, and scenario template production. Named entity recognition and coreference resolution grab more attention from the NLP community.

2.2.1 Named Entity Recognition

Named Entity Recognition and Classification (NERC) is a subtask of IE which identifies references to the entities in text. For example, the sentence "Apple Inc. was founded by Steve Jobs, Steve Wozniak, and Ronald Wayne in 1976." is annotated as " Inc.]$_{organization}$ was founded by [Steve Jobs]$_{person}$, [Steve Wozniak]$_{person}$, and [Ronald Wayne]$_{persom}$ in [1976]$_{time}$." by a NERC tagger. The task of recognition of named entities was added to the Sixth Message Understanding Conference MUC (MUC-6). Early systems [27, 28, 29] used handcrafted rule-based algorithms to perform the task. More studies [30, 31, 32] illustrate that machine learning techniques, especially supervised learning ones, can be used to perform the task and induce rule-based systems. [33, 34] reviewed early NERC studies. Recently, researchers applied deep learning techniques on NER tasks. li et al. [35] provided a survey on deep learning for NER.

Biomedical named entity recognition (Bio-NER) is a key element in extracting knowledge from medical guidelines. Bio-NER systems extract different biomedical entity types (e.g., disease and gene) from text. MetaMap [36] and cTAKES [37] are two well-known Bio-NER tools which extract biomedical entities and normalize them to the unified medical language system (UMLS) [38] concepts.

2.2.2 Part of Speech Tagging

Part of speech tagging (POS) is the process of labeling each word of a sentence, with its grammatical speech role such as verb. POS taggers usually take advantage of the grammatical or semantic context of the terms [39]. There are many available datasets of part-of-speech labels. A popular one in the modern English language is

Penn Treebank where the tagset size is 48 [40].

An example of part of speech tagging for the following sentence [41] is computed using Python NLTK [42]:

"Many studies comparing different inhaled steroids are of inadequate design and have been omitted from further assessment."

[('Many', 'JJ'), ('studies', 'NNS'), ('comparing', 'VBG'), ('different', 'JJ'), ('inhaled', 'JJ'), ('steroids', 'NNS'), ('are', 'VBP'), ('of', 'IN'), ('inadequate', 'JJ'), ('design', 'NN'), ('and', 'CC'), ('have', 'VBP'), ('been', 'VBN'), ('omitted', 'VBN'), ('from', 'IN'), ('further', 'JJ'), ('assessment', 'NN'), ('.', '.')]

In this example JJ stands for *adjective*, NNS for *noun plural*, VBG for *verb gerund*, VBP for *verb present*, IN for *preposition*, *NN* for *noun singular*, CC for *coordinating conjunction*, and VBN for *verb, past participle*.

2.2.3 Parsing

In natural language processing, parsing refers to syntax analysis of a sentence. Intuitively, a tree diagram can represent the syntax where every word is tagged with the part of speech. The primary step of a parser is lexical analysis to extract the tokens, which can be considered as a word-level segmentation in discourse analysis. The next step is the syntax analysis of the sentence and extraction of its constituent components based on a specific grammar. In practice, many of the parsers take advantage of some statistical methods to hire an existing trained corpus. Here, a trained corpus contains large numbers of annotated words - e.g., Penn Treebank [40] which includes 4.5 million tagged English words. When dealing with a large corpus, parsing is costly; therefore, in practice, different parsing algorithms should consider a trade-off between the computational cost and the accuracy [43].

The parse tree output for the following sentence [41] is using Stanford NLP Parser is shown in Figure2.1:

"If bp is greater than 50, follow the instruction."

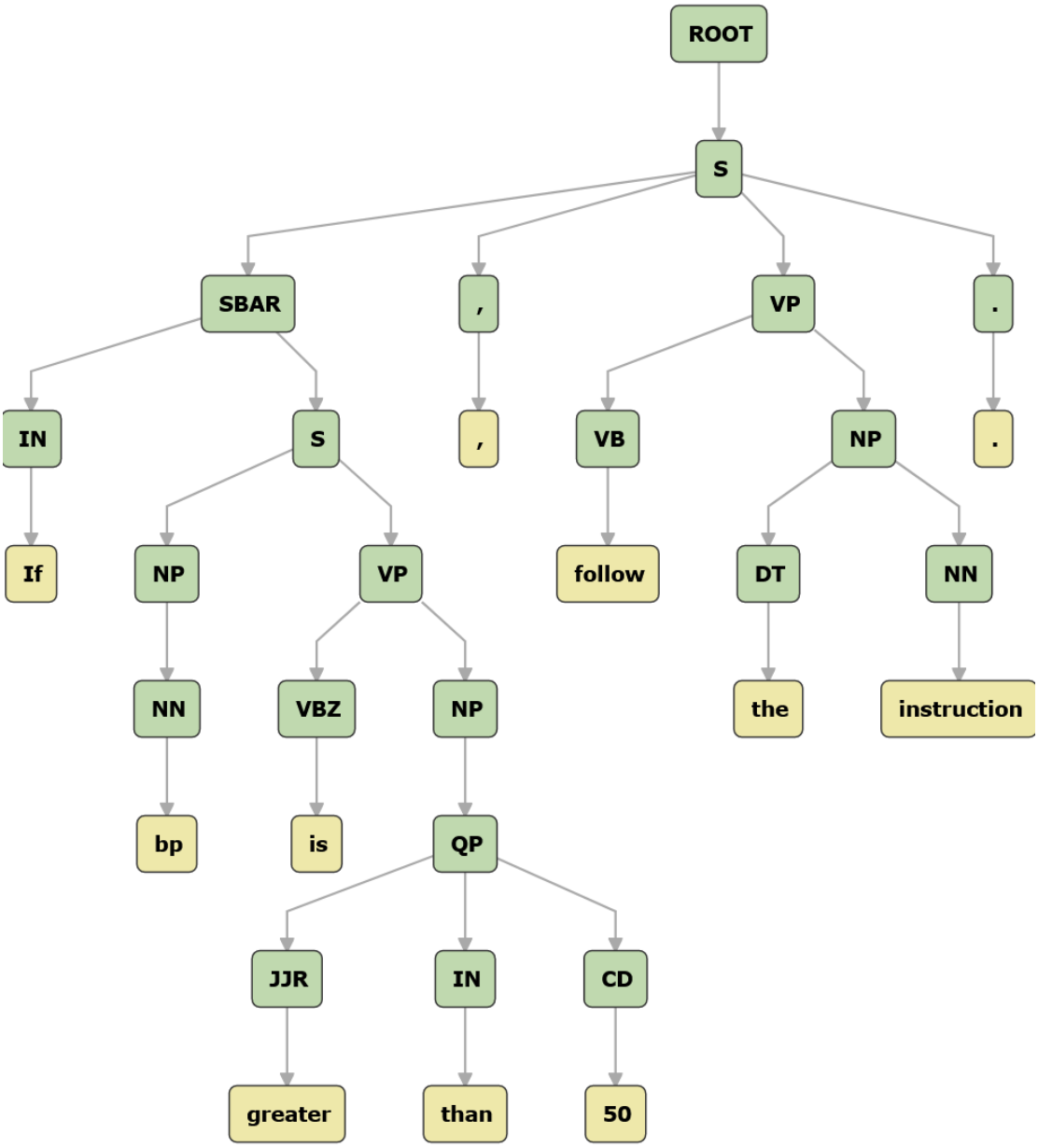


Figure 2.1: A parsed tree for the following sentence: "If bp is greater than 50, follow the instruction."

2.2.4 Coreference Resolution

The task of resolving noun phrases to the entities that they refer to is called coreference resolution. For example, in the sentence "John wants to marry Maria because he is in love with her", *he* refers to John, and *her* refers to Maria. Coreference resolution helps NLP researchers in areas such as NERC and question answering. The early

studies [44, 45] focused on pronoun resolution and used linguistic-based approaches. Machine learning approaches such as Naive Bayes', clustering, and decision trees were adopted to perform the task by [46, 47, 48].

2.2.5 Discourse/Text Segmentation

In order to analyze and understand the syntactic and semantic structure of a discourse, a text is required to be partitioned into meaningful units. This process is referred to as text or discourse segmentation. Segmentation can be applied to different levels of abstraction such as word, sentence, intent, topic, and conversation. The result of a segmentation task is boundaries of the segments of the discourse. Some written languages have explicit boundary cues such as white space in English for word-level segmentation. However, these explicit boundaries do not exist equally for every level of abstractions and in all languages. This is one of the problems that makes the task non-trivial.

Focusing on the English language, both syntactic and semantic cues have been used for segmentation tasks. In lower-level segmentations such as word-level, white space, and some other delimiters like hyphens give a simple but relatively accurate result. However, moving to the higher level of abstraction, these explicit cues are less accurate, and the task becomes more ambiguous. Thus, more complicated methods might be required in this case. For instance, sentence segmentation relies primarily on punctuations, specifically full stop, question, or exclamation marks. But the ambiguity of compound sentences as well as using the same punctuation in other parts of the sentence will increase the need for the other complementary rules and learnings. Moving upward to topic segmentation, the problem becomes even more complicated as the human readers also might have different ideas about the segments and boundaries. Generally, the rule-based approaches are accurate enough for word and sentence segmentation, but topic segmentation methods are mostly probabilistic and rely on machine learning approaches [49, 50].

As an example, consider the following paragraph [41]:

"Many studies comparing different inhaled steroids are of inadequate design and have been omitted from further assessment. In view of the clear differences between normal volunteers and asthma patients in the absorption of inhaled steroids, data from normal volunteers have not been taken into account. Only studies in which more than one dose of at least one of the inhaled steroids or both safety and efficacy had been studied together in the same trial were evaluated. Non-blinded studies also had to be considered because of the problems of obtaining competitors' delivery devices. All comparisons used BDP-CFC (chlorofluorocarbons) as the reference."

The sentence segmentation using Python NLTK would be: [Many studies comparing different inhaled steroids are of inadequate design and have been omitted from further assessment.] [In view of the clear differences between normal volunteers and asthma patients in the absorption of inhaled steroids, data from normal volunteers have not been taken into account.][Only studies in which more than one dose of at least one of the inhaled steroids or both safety and efficacy had been studied together in the same trial were evaluated.][Non-blinded studies also had to be considered because of the problems of obtaining competitors' delivery devices.][All comparisons used BDP-CFC (chlorofluorocarbons) as the reference.]

And considering the first sentence, the word segmentation, known as tokenizer, using the same library would be: ['Many', 'studies', 'comparing', 'different', 'inhaled', 'steroids', 'are', 'of', 'inadequate', 'design', 'and', 'have', 'been', 'omitted', 'from', 'further', 'assessment', '.']

2.2.6 Sentiment Analysis

A major aspect of understanding a discourse is analyzing it based on the attitude of the author toward a subject or entity. Sentiment analysis and opinion mining are the computational approaches to capture people's opinions, attitudes, and sentiments. While these two concepts are sometimes being used interchangeably, some researchers

believe their goals are different [51]. Based on their argument, in opinion analysis, we are looking for the reason as well as the attitude and sentiment. Generally, the target of sentiment analysis is extracting people's opinion towards an entity, identifying the expressed sentiment for that, and classification of the polarity of these sentiments [52]. The popular approaches in text mining used to be mainly lexicon-based [53]. Machine learning [54] approaches are becoming more popular in recent years. A lexicon-based approach takes advantage of available dictionaries and statistical methods, while a machine learning-based approach utilizes machine learning algorithms that are trained on a labeled corpus to extract the polarity features. According to IBM Watson analyzer [55], the following sentence [41] holds a negative sentiment of -0.56 on the scale of -1 to +1.

"Many studies comparing different inhaled steroids are of inadequate design and have been omitted from further assessment."

2.2.7 Terminology Extraction

Terminology extraction refers to the methods that retrieve a basic vocabulary for a domain-specific corpus [56] [57]. Providing a list of possible candidate terms, the algorithm needs to validate those terms in different ways. It is necessary to review the relationships among those terms that represent any domain concept. In the path to refine the candidate terms, one may need to answer a few questions. To what extent is a (short) list of terms covering the domain? To what extent do these terms refer to a similar (but not the same) domain? To what extent might they refer to a completely different domain? Answering these questions, we may come up with a measure of quality for the terminology extraction process. In other words, a reliable design of the terminology extraction process should provide precise answers to these questions. There are many linguistic and statistical algorithms for terminology extraction. In linguistic approaches, good candidate terms are being considered based on their syntactic characteristics. Thus, this approach requires a step of parsing the

corpus. On the other hand, the statistical approach, as its name suggests, relies on term frequencies. In the simplest example, the n-grams will be filtered based on term frequency-inverse document frequency (TF-IDF). Note that in many cases, the statistical and linguistic algorithms are being combined in hybrid approaches. Considering the following two paragraphs [41]:

"Many studies comparing different inhaled steroids are of inadequate design and have been omitted from further assessment. In view of the clear differences between normal volunteers and asthma patients in the absorption of inhaled steroids, data from normal volunteers have not been taken into account. Only studies in which more than one dose of at least one of the inhaled steroids or both safety and efficacy had been studied together in the same trial were evaluated. Non-blinded studies also had to be considered because of the problems of obtaining competitors' delivery devices. All comparisons used BDP-CFC (chlorofluorocarbons) as the reference.

BDP and budesonide are approximately equivalent in clinical practice, although there may be variations with different delivery devices. There is limited evidence from two open studies of less than ideal design that budesonide via the turbohaler is more clinically effective. However, at present a 1:1 ratio should be assumed when changing between BDP and budesonide."

The suggested terminology using IBM Watson Analyzer [55] top terminology would be:

"different inhaled steroids" with a score of 0.94, "normal volunteers" with a score of 0.74, "delivery devices" with a score of 0.72, and "comparing different inhaled steroids" with a score of 0.67.

2.2.8 Language Representation Models

Word Embeddings are distributional semantic models which aim to map words, n-grams, phrases, or other units of meaning in a text to vectors. In an early study [58], authors introduced a new method for indexing and retrieval to overcome the deficien-

cies of term-matching retrieval systems. In this subsection, we present two state-of-the-art models for word embedding and a recent language representation model.

Mikolov et al. [59] proposed two model architectures, Continuous Bag-of-Words (CBOW) and Continuous Skip-gram, for computing vector representations of words from a corpus. In CBOW, the model predicts the current word based on the surrounding words. The skip-gram model predicts a surrounding window of context words. CBOW represents frequent words better, and skip-gram can predict rare word representation with higher accuracy [59]. Word2Vec models can be used for NLP tasks such as word similarity tasks.

Global Vectors for Word Representation (GLoVe) [60] is an unsupervised learning algorithm that aims to represent words with vectors. The authors proposed a weighted least squares regression model which uses word-word co-occurrence statistics to produce a word vector space. The model was evaluated with word analogies [59], word similarity, and named entity recognition tasks.

BERT [61] is a language representation model developed by the Google AI language group. Unlike most of the word embedding models, BERT is not a feature-based model. Feature-based models provide pre-trained representation as additional features for task-specific architectures [61]. The pre-trained BERT representation is designed to be fine-tuned with just one additional output layer to create models for NLP tasks.

BERT introduced the "masked language model" (MLM) to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers [61]. Since it aims to extract long contiguous sequences, BERT was trained on English Wikipedia and the BooksCorpus [62].

BERT outperformed previous state-of-the-art models, such as OpenAI GPT[63] and ELMO [64], on all General Language Understanding Evaluation(GLUE) [65] benchmark datasets. The GLUE benchmark includes datasets for the evaluation of

various NLP tasks such as question answering, sentiment analysis, recognizing textual entailment, and semantic textual similarity.

NLP community provides various software and services to perform IE tasks. NLTK [42], OpenNLP [66], GATE [67], Google Cloud Natural Language, and AllenNLP [68] are some examples of those systems.

2.3 Information Retrieval

In this section, we present different aspects of Information Retrieval systems. Information Retrieval (IR) is the activity of searching for documents, information within documents, information within relational databases, text, metadata, multimedia files, or any information space [69]. The idea of using machines to access large amounts of stored knowledge was introduced by Vannevar Bush in 1945 [70]. The term "information retrieval" was defined by Mooers [71] in 1950 as follows: "Information retrieval is the name of the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him."

Figure2.2 illustrates the architecture of a general IR system. In this architecture, the user submits a query to the retrieval system. The retrieval system uses the indexed data to retrieve documents that are probably relevant to the query and compute a relevance score for each retrieved document. The documents will be presented to the user. An IR system implements in 3 processes:

1. Indexing process: representing documents in a summarized content form.

2. Query formulation process: representing the user information need.

3. Matching process: retrieval of relevant documents that satisfies user information needs.

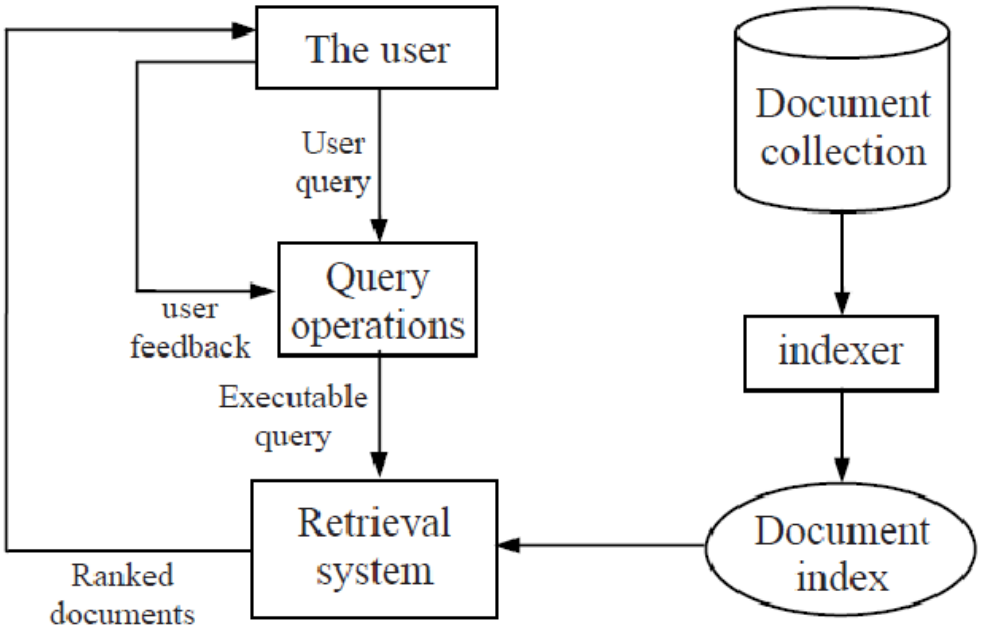


Figure 2.2: Architecture of a general IR system

2.3.1 Indexing Process

Salton [72] introduced a blueprint for automatic indexing method:

1. Identify individual word in the document (Tokenizing)
2. Remove stop words
3. Stemming and reducing the terms to their root
4. Index multi-word phrases
5. Replace low-frequency terms with thesaurus classes
6. Replace high-frequency terms by phrases
7. Compute IDF measure for all terms
8. Assign to each document the corresponding single terms, phrases, and classes with IDF weights

2.3.2 Document and Query Modeling

A document or query should be represented in a form that is understandable by an IR system. IR systems can be classified based on underlying conceptual models. Boolean model, the vector-space model, and probabilistic models are some of the commonly used models [73]. The inference network model [74] is another model which can implement most of the techniques used by IR systems.

2.3.2.1 Boolean Model

In a Boolean model, the presence and absence of a term in a document is represented by 1 and 0 respectively [75]. Users define queries by using a combination of Boolean ANDs, ORs, and NOTs over a set of keywords. IR systems label documents as either relevant or irrelevant. Even though Boolean systems are efficient, easy to implement, and yield good performance in certain situations, they have several shortcomings, e.g., there is no inherent notion of the document ranking.

2.3.2.2 Vector Space Model

The vector space model was introduced by Salton et al. [76] in 1975. This model was created to overcome the weakness of the Boolean model. The Vector Space model represents documents and queries by vectors of terms in $|V|$-dimensional space (V is the set of all terms in the documents and queries). If a text contains a term, the term gets a non-zero value in the vector that represents that text. The model ranks documents based on their similarity to the user query. The similarity score can be calculated as the angle between two vectors representing the query and the document. For calculating the angle, one can use the cosine function:

$$d_i = (w_{1,i}, w_{2,i}, ..., w_{t,i})$$

$$q = (w_{1,q}, w_{2,q}, ..., w_{t,q})$$

$$cosine\_similarity(d_i, q) = \frac{d_i.q}{||d_i|| \times ||q||} = \frac{\Sigma w_{n,i} w_{n,q}}{\sqrt{\Sigma w_{n,i}^2}\sqrt{\Sigma w_{n,q}^2}}$$

Salton et al. [76] adapted a term weighting procedure [77] to introduce term weight scheme know as tf-idf. A term weighting system is proportional to multiplying standard term frequency f_i^k by a factor inversely related to document frequency d_k of the term k ($f_i^k.IDF_k$). IDF_k was defined as:

$$IDF_k = \lceil \log_2 n \rceil - \lceil \log_2 d_k \rceil + 1.$$

Even though the Vector Space model is simple and fast, it has its own weaknesses (e.g., the assumption of independence between terms.)

2.3.2.3 Probabilistic Model

Maron and Kuhns proposed a technique called "Probabilistic Indexing" [78] in 1960. The technique allows a machine to make statistical inferences and ranks documents based on the probability that the document will satisfy the given query. Similar to a vector space model, documents and queries are represented by vectors. Introducing binary independence retrieval (BIR) will cover some basic concepts of probabilistic IR. The probability of relevance for document d_i w.r.t to query q_k is $P(R|q_k, d_i)$. The basic assumption for this model is that terms are distributed differently within relevant and non-relevant documents [79]. By applying Bayes' theorem and using odds instead of probabilities, the relevance measure will be:

$$O(R|q_k, d_i) = \frac{P(R|q_k, d_i)}{P(\overline{R}|q_k, d_i)} = \frac{P(R|q_k)}{P(\overline{R}|q_k)} . \frac{P(d_i|R, q_k)}{P(d_i|\overline{R}, q_k)}$$

Fuhr discussed assumptions, probabilistic parameters, and learning strategies of some other probabilistic IR models in [79]. Some probabilistic models can be found in [80, 81, 82, 83].

2.3.3 Query Expansion

User queries are usually short and do not necessarily use the same words as indexed document terms. Furnas et al. [84] called this issue the *vocabulary problem*. Synonymy (the same word with different meanings) and polysemy (different words with the same meaning) are two components of the vocabulary problem, which may result in a decrease in recall and precision, respectively [85]. Several approaches have been proposed to deal with the vocabulary problem: automatic query expansion, interactive query refinement, relevance feedback, word sense disambiguation, and search results clustering [85].

Automatic query expansion (AQE) automatically expands the original query by using other terms that would represent the user information needed to better retrieve more relevant documents. Query expansion improves the effectiveness measures of the IR system (e.g., F-measure) [86, 85]; however, it may decrease precision [87]. Carpineto and Giovanni [85] reviewed different aspects of AQE.

Interactive query expansion (IQE), like AQE, generates some features to be used in reformulating the user query, but the user in this approach makes the final decision. This interaction gives users more control over the query processing.

Some users may not be able to formulate their information need into a query, but they would quickly recognize that the retrieved document is relevant to their need. These users can provide *Relevance Feedback* (RF) to IR systems to improve the results in quantity and quality. The IR system uses the information provided by RF to justify the query by adding terms to it or adjusting the weight of terms. Harman explored some methods for ranking terms by using RF in [88, 89]. Ruthven and Lalmas [90] reviewed RF methods and use of them in IR systems.

Ambiguity is a problem when a computer wants to understand natural language. A query, which is formulated by a user, may create the same problem. Word sense disambiguation (WSD) is the ability to identify which sense of a word is used in

context in a computational manner [91]. Weiss [92] started using disambiguators to resolve word sense in IR systems. Navigli [91] overviewed supervised, unsupervised, and knowledge-based approaches in WSD.

Clustering engines consider clustering as a postprocessing step to rerank the retrieved documents [93, 86] or provide an interactive interface for users to choose the area of interest [94, 95]. The latter approach is called search result clustering. Search results clustering (SRC) combines query-based and category-based search to provide clustered results for users with weakly specified or ambiguous queries. Carpineto et al. [96] mentioned the search aspects where SRC can be most useful as follow: fast subtopic retrieval; topic exploration; and alleviating information overlook. Reviews of SRCs are found in [96, 97, 98, 99].

2.3.4 Measures for Information Retrieval

The most common measures for evaluating IR systems are the Information Extraction performance metrics of precision, recall, and F-measure.

Precision is the proportion of the set of documents that are both relevant to a query and retrieved by the system out of all documents which are retrieved:

$$Precision = \frac{\text{Relevant} \cap \text{Retrieved}}{\text{Retrieved}}$$

Recall is the proportion of the set of documents that are both relevant to a query and retrieved by the system out of all documents which are relevant to that query:

$$Recall = \frac{\text{Relevant} \cap \text{Retrieved}}{\text{Relevant}}$$

A common aim of every IR system would be to maximize both precision and recall. Any system can be tuned to focus on one metric more than the other. The importance of a metric depends on the system goal. For example, in a patent retrieval system, recall is crucial for users.

F-measure is the weighted harmonic mean of recall and precision:

$$F_\beta = (1 + \beta^2)\frac{\text{precision} \times \text{recall}}{(\beta^2 \times \text{precision}) + \text{recall}}$$

β allows one to weigh either precision or recall more heavily. (F_1 score) is the harmonic mean of precision and recall where β is 1.

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Some other measures for evaluating information Retrieval systems [100]:

- Average precision: a single-valued measure to evaluate ranked retrieval. It is computed by measuring precision at different recall points and averaging.
- Mean Average Precision: mean of the average precision for each query of a set of queries.
- Precision at k: the number of relevant results on the first k retrieved documents.

2.4 Summary

This chapter reviewed the Clinical Practice Guidelines, Information Extraction tools, and Information Retrieval systems as the three main components of this dissertation. IE and IR provide us powerful tools to extract and retrieve knowledge from medical texts automatically. As mentioned earlier, automatic processes of analyzing CPGs are inevitable due to the growth rate of guidelines and disagreements between them. In the following chapters, we will report our works in different aspects of analyzing medical texts, mainly medical guidelines, using Information Extraction, Information Retrieval, and Machine Learning methods.

CHAPTER 3: SEMANTIC MODELING OF CONTRADICTIONS AND DISAGREEMENT: A CASE STUDY OF MEDICAL GUIDELINES

3.1 Introduction

In this chapter, we take the perspective of building a natural language understanding system that can adequately represent disagreements. This work is motivated by the challenge of automatically identifying and representing contradictions in medical guidelines. On the practical side, we expect this research to eventually result in a more prominent solution that can provide decision support for patients and physicians and help identify and reason with contradictory advice in their specific cases. However, the proposed solution applies more generally to natural language semantics.

Disagreements in medical guidelines raise uncertainty in disease screening and treatment. Uncertainty derived from the lack of guidelines consistency among different expert groups is confusing for patients and contributes to overdiagnosis. For example, the ACOG recommends that women over age 40 get mammography annually but the USPSTF recommends clinicians base screening decisions for women aged 40 to 49 on the women's individual risk profile and preferences.

Marneffe et al. [101] and Kloetzer et al. [102] contributed on representing contradictions in NLP. The former proposed a taxonomy of linguistic expressions of contradiction, potentially useful when dealing with the linguistic diversity of the guidelines. The latter shows methods for large-scale acquisition of contradictory patterns. We believe such distributional methods might add coverage to our approach and complement the IR method we are currently using. Neither of these works makes a formal distinction between contradictions and disagreements. On the formal side, clearly, there is a large body of work on contextualizing the truth of propositions, for exam-

ple, in modal and para-consistent logics [1].

In this chapter, we propose a novel formal analysis of types of contradictions in texts. Namely, we introduce and formally characterize the distinction between *contradictions* and *disagreements*. This distinction is generally applicable to all semantic processing of natural language text and is orthogonal to other typologies of contradictions, e.g., [101]. We also propose an architecture and a method for identifying contradictions and disagreements in medical guidelines. Results from an implemented system show the feasibility of the proposed approach.

Dealing with multiple guidelines for the same condition can be reduced to analyzing the guidelines pairwise. Thus with two texts of such guidelines, we propose the following:

1. Identify candidate sentences related to the same condition or the same action;

2. Compute candidate contradictions and disagreements (using techniques of information retrieval and statistical language modeling);

3. Identify the specific contradictions and disagreements computationally, using different, deeper modes of analysis based on semantic representation informed by *formal representations of disagreements and contradictions*;

4. A method for automated reasoning with disagreements and contradictions in computational settings focused on the identification areas of agreement and disagreement, including their origins.

3.2 Formal Representation of Disagreement and Contradiction

We need a formal representation of contradictory guidelines to be able to reason about them. This section proposes a way to reason with partially contradictory

[1] `https://plato.stanford.edu/entries/possible-worlds/`, and `https://plato.stanford.edu/entries/logic-paraconsistent/`

information based on a formal distinction between disagreements and contradictions and formalized using a combination of propositional calculus and lattice theory.

Let's consider a few examples of actual sentences containing disagreements or contradictions. For clarity of exposition, we will always present contradictions between pairs of documents, such as guidelines.

Example 1. We will use an example from a CDC table comparing "Breast Cancer Screening Guidelines for Women" provided by seven different accredited medical bodies. [2] There we find contradictory recommendations for "women aged 50 to 74 with average risk" coming from two (of the seven) different organizations):

(a) *Screening with mammography and clinical breast exam annually.*

(b) *Biennial screening mammography is recommended.*

Example 2. Consider the question about the recommended number of minutes of physical activity. Again, the guidelines might differ: One organization is recommending a minimum of 150 minutes per week, and another one recommends 150-300 minutes per week. Someone exercising 30 min per day, six days a week, satisfies both guidelines. The guidelines don't agree 100%, but intuitively they are not 100% contradictory either.

Disagreements vs Contradictions:

To capture the intuitive distinction between Examples 1 and 2, we say that two guidelines are *contradictory* if it is impossible for both guidelines to be followed. Two guidelines are in *disagreement* if there are patients where the two guidelines are possible to be followed and patients for which this is impossible. As it turns out, we can represent this distinction formally, in logic, making it broadly applicable in semantics, using the following idea:

[2] `https://www.cdc.gov/cancer/breast/pdf/BreastCancerScreeningGuidelines.pdf`

- *Contradiction* is present if there is no model for the joint theory expressed in two text segments (coming from different guidelines).
- *Disagreement* is present if the sets of models, for the predicates present in both text segments, are different for each segment, but a model can be created satisfying both segments.

Formalization: We start by assuming that, at least initially, we do not need the full power of first order logic (FOL) or a stronger logical system. So, the basis of our representation will be a formal language of propositions. Thus, we do not have variables or quantifiers. However, to be able to reflect the disagreements, we need to augment it with a representation of parameters. For example, we would like to be able to distinguish between a recommendation of a minimum of 30 minutes of daily exercise and another one of 20 minutes. At the same time, we need to be able to notice that both recommendations pertain to the recommended dose of exercise.

To this end, we assume that our representation language contains propositional symbols $p, q, r, ..., p_1,\ p_2, ...$ and symbols representing parameters $a, b, c, a_1, a_2,$ etc. We have special parameters $o_1, o_2, and...$, which will later represent the provenance of recommendations. This will allow us to find the sources of contradictions and disagreements.

We assume the standard inference rules of propositional logic. (The added parameters don't extend the power of the system beyond propositional calculus). To reason about disagreements, we will need to introduce additional rules of inference.

Example 1 continued: Let p stand for *screening mammography is recommended*; o_1, o_2 represent the provenance of the recommendations (a) and (b) respectively; and a, b stand for *annually* and *biennally*. We then have the formal representation of the respective guidelines as $p(a, o_1)$ and $p(b, o_2)$.

We need means to represent the fact that these guidelines are formally contra-

dictory. This cannot simply be done due to having different constants/parameters inside the parentheses (and ignoring the o's). To see that, consider a similar representation of doses of daily recommended exercise. Here, we would also have two distinct provenances and two distinct values; however, intuitively, we could recognize a disagreement and not a contradiction since anyone exercising 30 min or more is also exercising 20 min or more.

To proceed, we need to make two additional assumptions, namely that no particular guidelines document can have internal contradictions. That is, the set of all $p_i(a^i_j, o)$ for a particular o is never contradictory (viewed as statements in classical propositional logic).

And the second assumption is that the parameters come in different sorts, which we will represent by capital letters followed by a colon, e.g., $A : a_1$. More importantly, we assume elements of any particular sort form a *lattice* (or at least *meet semi-lattice*). For any set of parameters of a particular sort (e.g., time, duration, dosage, etc.) $a_1 \wedge a_2$ is defined, and every such lattice has a minimal element $\bot$.

In the example representations of mammography the meet of *biennial* and *annual* is $\bot$. However, the meet of "20 min or more" and "30 min or more" is the latter. This mechanism allows us to make a formal distinction between contradictions and disagreements.

$p(A\colon a_1, o_1)$ and $p(A\colon a_2, o_2)$ are *contradictory* if $a_1 \wedge a_2 = \bot$

$p(A\colon a_1, o_1)$ and $p(A\colon a_2, o_2)$ *disagree* if taking $a_1 \wedge a_2 = a$, we have $a \neq \bot$ and either $a \neq a_1$ or $a \neq a_2$.

These definitions are naturally extended to multi-parameter cases by defining the contradiction as a situation, where the meet of at least one type of parameter is $\bot$; and the disagreement, when there's no contradiction and at least one type of parameter contains a disagreement.

To do some elementary reasoning about disagreement, we need an inference rule capable of relating two formulas with different parameters. To keep track of provenances, we need to allow propositions with multiple labels, e.g., $\{o_1, o_2\}$. This is nicely combined in a single inference rule:

$$\frac{p(A\colon a_1, o_1), p(A\colon a_2, o_2)}{p(A\colon a_1 \wedge a_2, \{o_1, o_2\})}\ \text{Lattice}\ \wedge$$

If the formula p has more than one parameter, we apply this rule for each parameter separately.

With this inference rule, we are getting the following:

- The set of derivable (using the above rule) contradictory propositions corresponds to the ones that have $\perp$ as at least one parameter.
- The set of disagreements corresponds to the derivable propositions with two or more provenance parameters and a disagreement for one or more sorts.
- For any fixed provenance o, we have the full power of inference rules of propositional calculus applied to sentences of the form $p(a, o)$ where a stands for a collection of parameters of different sorts.

3.3 Finding Contradictions and Disagreements

Having solved the problem of formally representing contradictions and disagreements and having created a formal method of keeping track of their provenances, we now focus on the language understanding part.

The results presented in this section are preliminary in two ways: First, we have not completed a translation from a semantic representation produced by NLP tools to a logical form amenable to reasoning with the parameterized propositional logic of the previous section. We assume this can be done using existing methods, as described in publications ranging from standard textbooks [103] to complex NLP architectures [104].

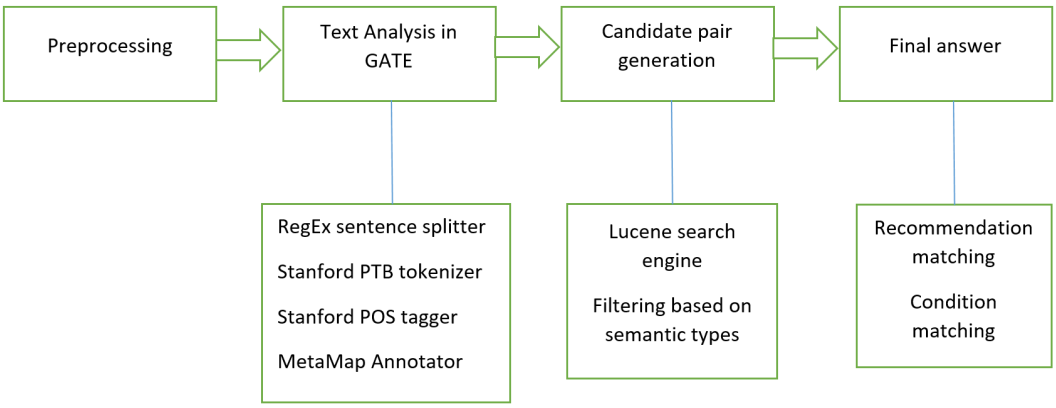


Figure 3.1: The architecture used to evaluate extraction of contradictions in medical guidelines.

Second, our methods for finding contradictions and disagreements, even though not trivial, very likely can be improved. Nevertheless, the results are promising.

Figure 3.1 shows a novel architecture consisting of several well-known components. We follow the approach presented earlier: We use the text analysis tools for feature generation and concept identification. The Lucene search engine was used for finding similar sentences based on indexed semantic features (and words). For example, given the query "mammography is recommended for women age 40-49" we search multiple guidelines and identify sentences for further analysis. This analysis was done through recommendation matching ("mammography recommended"), and condition matching ("age 40-49" or "age") allows the system to decide if the given guidelines document recommends a procedure or not. Similarly, we can identify partial matches, e.g., "age 40-49" and "age over 40".

Evaluation: At this point, we only evaluated this method on finding agreement and disagreement on twelve example recommendations sentences and breast cancer screening guidelines produced by seven different medical organizations. [3] This gives us only 84 data points. However, the results are promising: The system produced only four errors (two false positives and two false negatives), thus on this – admittedly,

[3] `https://www.cdc.gov/cancer/breast/pdf/BreastCancerScreeningGuidelines.pdf`

simple – data set achieved an impressive accuracy of 95%.

3.4 Conclusion and Discussion

Motivated by analysis of medical guidelines, we introduced the formal distinction between disagreements and contradictions. We presented a new system for finding both and results of a preliminary evaluation. The new formal representation and the general architecture of the system are potentially broadly applicable to NLP, for example, to question answering, where an answer can be extracted from texts that disagree on details but broadly provide the same answer or recommend the same action.

We need to address some limitations we faced in our experiment. In evaluation, we used simple sentences, but texts might contain information in multiple sentences, and thus increasing the difficulty of matching. We do not have reliable ways of converting longer texts intended for human reading into a semi-structured representation suitable for text mining (for example, dealing with tables). While solutions to these problems exist, they are not perfect and will likely decrease the system's accuracy.

CHAPTER 4: CONCEPTUAL DISTANCES BETWEEN MEDICAL RECOMMENDATIONS: EXPERIMENTS IN MODELING MEDICAL DISAGREEMENT

Using natural language processing tools, we investigate the semantic differences in medical guidelines for three decision problems: breast cancer screening, lower back pain, and hypertension management. The recommendation differences may cause undue variability in patient treatments and outcomes. Therefore, having a better understanding of their causes can contribute to a discussion on possible remedies. We show that these differences in recommendations are highly correlated with the knowledge brought to the problem by different medical societies, as reflected in the conceptual vocabularies used by the different groups of authors. While this chapter is a case study using three sets of guidelines, the proposed methodology is broadly applicable. Technically, our method combines word embeddings and a novel graph-based similarity model for comparing collections of documents. For our main case study, we use the CDC summaries of the recommendations (concise documents) and full (long) texts of guidelines represented as bags of concepts. For the other case studies, we compare the full text of guidelines with their abstracts and tables, summarizing the differences between recommendations. The proposed approach is evaluated using different language models and different distance measures. In all the experiments, the results are highly statistically significant. We discuss the significance of the results, their possible extensions, and connections to other domains of knowledge. We conclude that automated methods, although not perfect, can be applied to conceptual comparisons of different medical guidelines and can enable their analysis at scale.

4.1 The Problem and the Method

This work investigates a natural question. We are asking whether differences in medical recommendations arise from differences in knowledge brought to the problem by different medical societies. To answer this question at scale, we need an automated method to measure such differences. This work aims to present such a computational method and use a collection of case studies to evaluate its performance.

Our method uses the standard natural language processing approach to represent words and documents as embeddings and combines it with a graph comparison algorithm. We evaluate our approach on three sets of medical guidelines: breast cancer screening, lower back pain management guidelines, and hypertension management guidelines.

The answer to this question matters because physicians with different specialties follow different guidelines. This results in the undue variability of treatment. Therefore, understanding what drives the differences in recommendation should contribute to its reduction and to better patient outcomes [105, 106, 7].

4.1.1 Motivation

There are over twenty thousand clinical practice guidelines indexed by PubMed [1], with over 1,500 appearing every year [107]. Since clinical practice guidelines are developed by different medical associations, which count on experts with different specialties and sub-specialties, there is a high possibility that there may be disagreement in the guidelines. Indeed, as noted by [7], and discussed in [108, 6], breast cancer screening guidelines contradict each other. Besides breast cancer screening disagreements, which we model in this chapter, controversies over PSA screening, hypertension, and other treatment and prevention guidelines are also well-known.

Figure 4.1 illustrates our point. We see disagreements in seven breast cancer screen-

[1] `https://pubmed.ncbi.nlm.nih.gov/`

ing recommendations produced by seven different medical organizations. We investigate the hypothesis that the contradictory recommendations reflect the specialized knowledge brought to bear on the problem by different societies.

Notice that the dominant view is to see expertise as a shared body of information, and experts as *epistemic peers* [109] with identical levels of competence. Under this paradigm of shared knowledge and inferential abilities, the medical bodies should not differ in their recommendations. What they do is interesting and worth investigating. Thus, this research is also motivated by the idea that epistemology of disagreement [110, 109, 111] can be modeled computationally. On the abstract level, medical disagreements are viewed as "near-peer" disagreement [112, 113, 114], where we see expert groups as having partly overlapping knowledge. This work shows that such more realistic and fine-grained models can also be studied computationally, quantitatively, and at scale.

	U.S. Preventive Services Task Force[1] 2016	American Cancer Society[2] 2015	American College of Obstetricians and Gynecologists[3] 2011	International Agency for Research on Cancer[4] 2015	American College of Radiology[5] 2010	American College of Physicians[6]	American Academy of Family Physicians[7] 2016
Women aged 40 to 49 with average risk	The decision to start screening mammography in women prior to age 50 years should be an individual one. Women who place a higher value on the potential benefit than the potential harms may choose to begin biennial screening between the ages of 40 and 49 years.	*Women aged 40 to 44 years* should have the choice to start annual breast cancer screening with mammograms if they wish to do so. The risks of screening as well as the potential benefits should be considered. *Women aged 45 to 49 years* should get mammograms every year.	Screening with mammography and clinical breast exams annually.	Insufficient evidence to recommend for or against screening.	Screening with mammography annually.	Discuss benefits and harms with women in good health and order screening with mammography every two years if a woman requests it.	The decision to start screening mammography should be an individual one. Women who place a higher value on the potential benefit than the potential harms may choose to begin screening.
Women aged 50 to 74 with average risk	Biennial screening mammography is recommended.	*Women aged 50 to 54 years* should get mammograms every year. *Women aged 55 years and older* should switch to mammograms every 2 years, or have the choice to continue yearly screening.	Screening with mammography and clinical breast exam annually.	*For women aged 50 to 69 years*, screening with mammography is recommended. *For women aged 70 to 74 years*, evidence suggests that screening with mammography substantially reduces the risk of death from breast cancer, but it is not currently recommended.	Screening with mammography annually.	Physicians should encourage mammography screening every two years in average-risk women.	Biennial screening with mammography.

Figure 4.1: Note the contradictory recommendations in green and blue boxes. The colors in the table come from [6], but the original table comes from the CDC [7]. Only a part of the table is reproduced here.

4.1.2 Brief Description of the Proposed Method

In this study, we investigate the question of whether differences in medical recommendations come from differences in specialized medical knowledge applied to specific classes of patients, and whether such differences in specialties can be modeled computationally.

Our idea is to model "specialized medical knowledge", which we cannot easily observe, by the differences in the vocabulary used in medical guidelines. We then show that these vocabularies, assembled in vector representations of these documents, produce the differences in recommendations. We evaluate our method using three case studies: breast cancer screening guidelines, lower back pain management guidelines, and hypertension management guidelines. In this study's main track, we use the breast cancer screening guidelines to present our approach and the evaluation, and the additional evaluations on the other two sets of guidelines are presented.

More specifically, we computationally compare the *full* texts of guidelines with the their *recommendation summaries*. For breast cancer screening, the summaries come from the CDC [7]; for lower back pain management, they come from a summary article [1]; and, for hypertension management, where we lack a tabular comparison, we used the abstracts of the documents.

We see if the semantic similarities between the full documents follow the same pattern as semantic similarities between the summaries. Note that each computational comparison was made between two *sets* of documents and not individual documents.

This process involves several steps and is shown in Figure 4.2, for the breast cancer screening guidelines. Thus, the vector representations of full texts of the guidelines model the vocabularies as bags of concepts, and therefore cannot model specific recommendations: the concepts in the recommendations, such as "mammography" and "recommend", appear in *all* full texts, but specific societies may be either for mammography or against it. The vector representations of recommendations model the

differences in prescribed procedures, but not the vocabularies (see Tables 4.1 and 4.2 below).

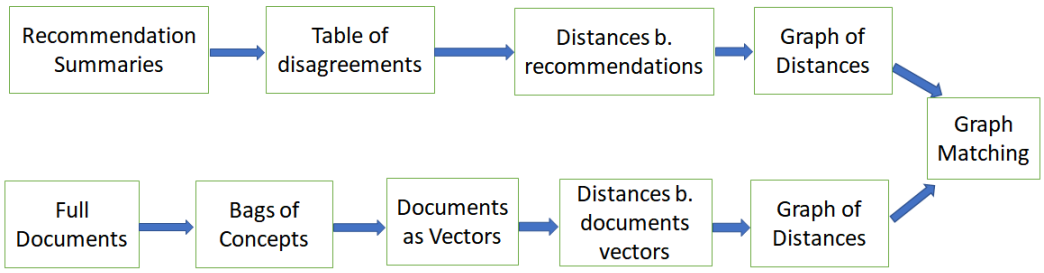


Figure 4.2: The method of comparing concepts in full documents and recommendations contained in summaries. Note the difference in representations: the documents are represented by a large number of high-dimensional (200) vectors with real valued features, whereas the disagreement representations can be low-dimensional vectors with discrete features (e.g., five-dimensional for the breast cancer screening guidelines). Our exposition will roughly follow the left-to-right order of this figure, using the breast cancer screening guidelines as the motivating example.

Table 4.1: The table shows recommendations as follows: N—no recommendation; b—both patient and doctor, shared decision; r—recommending mammography.

Guideline	40–49	50–74	75+	Dense Breast	Higher Than Average Risk
AAFP	b	r	b	b	N
ACOG	r	r	b	b	r
ACP	b	r	r	N	N
ACR	r	r	r	b	r
ACS	b	r	r	b	b
IARC	b	r	N	b	r
USPSTF	b	r	b	b	r

Table 4.2: This table shows the number of differing feature values for pairs of guidelines, based on Table 4.1. The Jaccard distances between the documents are obtained by dividing the value in the table by five (the number of features).

	AAFP	**ACOG**	**ACP**	**ACR**	**ACS**	**IARC**	**USPSTF**
AAFP	0	2	3	3	2	2	1
ACOG	2	0	4	1	2	2	1
ACP	3	4	0	3	2	3	3
ACR	3	1	3	0	1	2	2
ACS	2	2	2	1	0	1	1
IARC	2	2	3	2	1	0	1
USPSTF	1	1	3	2	1	1	0

How do we know if vocabularies determine recommendations? We compute pairwise distances (cosine or word mover's distance) between the full text vectors. In parallel, we compute pairwise distances between the recommendation vectors. We thus get two graphs, and their shapes can be compared. We show that the resulting geometries are very similar and could not have been produced by chance.

This process is slightly modified for lower back pain management, where we start with the tables of disagreement from the summary article [1]. For the hypertension management guidelines, we use the graph of summaries that is generated from the abstracts of full documents, because we do not have any tabular sets of comparisons similar to [7, 1]. However, even with this change, the proposed method performs very well. Notice that we use a large number of high-dimensional (200) real-valued vectors to model full documents. By contrast, the vectors representing the recommendations only have a smaller number of discrete-valued features (five for the breast cancer screening and 12, 59, and 71 for lower back pain management).

4.1.3 Summary of Contributions

The main contribution of this work is in proposing an automated and relatively straightforward method of text analysis that (1) computes conceptual differences between documents addressing the same topic (for example, breast cancer screening) and (2) these automated judgments have a high correlation with recommendations extracted from these documents by a panel of experts. We test the approach on the already mentioned breast cancer screening recommendations, as well as in other sets of experiments on lower back pain management and hypertension management guidelines. As such, these results open the possibility of large-scale analysis of medical guidelines using automated tools.

Another contribution is the articulation of a very natural graph clique-based algorithm/method for comparing the similarity of two *collections* of documents. Given two sets of documents, each of the same cardinality, and a mapping between nodes, we compute the percent of similarity (or, equivalently, the distortion between the shapes of the two cliques), and the chances that the mapping arose from a random process.

We also document all steps of the process and provide the data and the code to facilitate both extensions of this work and its replication (the GitHub link is provided in Section 4.9).

4.1.4 Organization of the Chapter

In Section 4.2, we provide a brief overview of applications of natural language processing to texts of medical guidelines, word embedding, and some relevant work on disagreement. Afterward, we follow the left-to-right order of Figure 4.2 using the breast cancer screening guidelines as the motivating example (other experiments are described in the Section 4.7). Thus, Sections 4.3 and 4.4 explain our example data sources: a CDC summary table of breast cancer screening guidelines and the corre-

sponding full text documents. In these two sections, we also discuss the steps in the conceptual analysis of the table. First, the creation of a graph of conceptual distances between the columns of the table, and then the encoding of full documents as vectors, using two standard vectorization procedures. Our method of comparing summarized recommendations and full guideline documents is presented in three algorithms and discussed in Section 4.5.

After observing a roughly 70% similarity between the distances in the summaries and the distances in the full documents, we prove in Section 4.6 that this similarity is not accidental. We conclude in Sections 4.6 and 4.9 that this case study shows that NLP Methods are capable of approximate conceptual analysis in this space (using the section 4.7 for additional support). This opens the possibility of deepening this exploration using more sophisticated tools such as relationship extraction, other graph models, and automated formal analysis (as discussed in Sections 4.8 and 4.9).

In the section 4.7, we provide information about additional experiments we performed to validate the proposed method. There, we first discuss a few variants of the main experiment, where we filtered out some sentences from the full guidelines' texts. Then, we apply our method to two other collections of guidelines: namely, to hypertension and low back pain management guidelines. All of these experiments confirm the robustness of the proposed method and the system's ability to computationally relate background knowledge to actual recommendations.

4.2 Discussion of Prior Art

We are not aware of any work directly addressing the issue we are tackling in this study; namely, the automated conceptual analysis of medical screening recommendations. However, there is a body of knowledge addressing similar issues individually, which we summarize in this section.

4.2.1 Text Analysis of Medical Guidelines

An overview article [115], from a few years ago, states that different types of analysis of medical guidelines are both a central theme in applications of artificial intelligence to medicine and a domain of research with many challenges. The latter include building formal, computational representations of guidelines and a wider application of natural language processing. From this perspective, our work is relevant to these central and general themes.

A more recent and more technical work [116] focuses on finding and resolving conflicting recommendations using a formal model and automated proof systems—it relies on a manual translation into a formal language, Labelled Event Structure. This is a very interesting work, somewhat in the spirit of our own attempts, using a combination of NLP and information retrieval tools [6]. Another article [117], dealing with contradictory recommendations, focuses on the semi-automatic detection of inconsistencies in guidelines; these tools are applied to antibiotherapy in primary care. Another recent application of natural language processing [118, 119] shows that one can accurately measure adherence to best practice guidelines in the context of palliative care, as well as try to assess the quality of care from discharge summaries.

More broadly, modern NLP methods have been applied to clinical decision support, e.g., [120], with ontologies and semantic webs for concept representation; to clinical trials [121]; and to automatic extraction of adverse drug events and drug-related entities, e.g., using a neural networks model [122]. For document processing, we have, e.g., a knowledge-based technique for inter-document similarity computation [123], and a successful application of conceptual representations to document retrieval [124].

These show that the state-of-the-art systems are capable of performing statistical analysis of sets of documents and a semantic analysis fitting the need for a particular application. Our work extends both of these in a new direction and connects statistics with semantics to analyze medical guidelines.

4.2.2 Vector Representations of Documents Using Word Embeddings

Over the last ten years, we have witnessed a new era in automated semantic analysis of textual documents [125]. While no system can claim to "really" understand natural language, in several domains, such as data extraction, classification, and question answering, automated systems dramatically improved their performance. In some cases, they performed better than humans due to the unmatched pattern recognition and memorization capabilities of deep neural networks (see, e.g., [126] for an overview).

Some of the simplest, easiest to use, and effective of these new methods are different types of word and concept embeddings [127, 60, 128, 129]. Embeddings represent words and concepts as dense vectors (i.e., a few hundred-dimensional real-valued vectors). They are a preferred tool to make similarity judgments on the level of words, phrases, sentences, and whole documents. They have been applied to medical texts—see [130] for a survey.

Word embeddings have been widely used to compare documents, and in particular, to compute their degree of similarity [131, 132]. Other methods proposed to compute document similarity are based on using background knowledge [123].

This work uses both methods, namely human knowledge encoded in the CDC table (Figure 4.1), and embeddings. For the former, we use five-dimensional feature vectors representing differences in recommendations (Section 4.3). For the latter, we use (several versions of) 200-dimensional embeddings of full documents (Section 4.4).

4.2.3 Other Work on Disagreements and Contradictions

A comprehensive review of medical disagreement with a focus on intervention risks and the standards of care can be found in [133]. Once medical experts express their disagreements, what happens next? Observations from disagreement adjudication are analyzed in [134, 135], where the authors observe (among other things) that the differences in experts' backgrounds increase the degree of disagreement.

If we broaden the context beyond medical disagreements to artificial intelligence, there is a substantial amount of work on contradictory knowledge bases, as exemplified by [136, 137, 138]. Of particular interest may be proposals for real-valued measures of contradictions in knowledge bases [139, 138]. However, in that particular research avenue, the starting points are collections of facts and not recommendations; moreover, natural language texts are not mentioned. We believe this type of work will become more relevant as our capabilities to extract knowledge from text improve.

4.3 From Recommendations to Vectors of Differences and a Graph

We start with the simpler task of transforming the screening recommendations (referenced above in Figure 4.1) to vectors of differences, representing the disagreements in the recommendations, and then to a graph of their conceptual distances, where, intuitively, the larger the number of recommendation differences, the bigger the distance.

We will proceed in three steps: First, using a diagram (Figure 4.3) and a table (Table 4.1) we make explicit the difference in recommendations in Figure 4.1. Second, we transform the table into a count of differences (Table 4.2), and from that, we derive distances between pairs of recommendations (Table 4.3). The graph representing the recommendations will have nodes named after each organization (e.g., AAFP, ACOG, etc.) and edges labeled and drawn with distances (Figure 4.4).

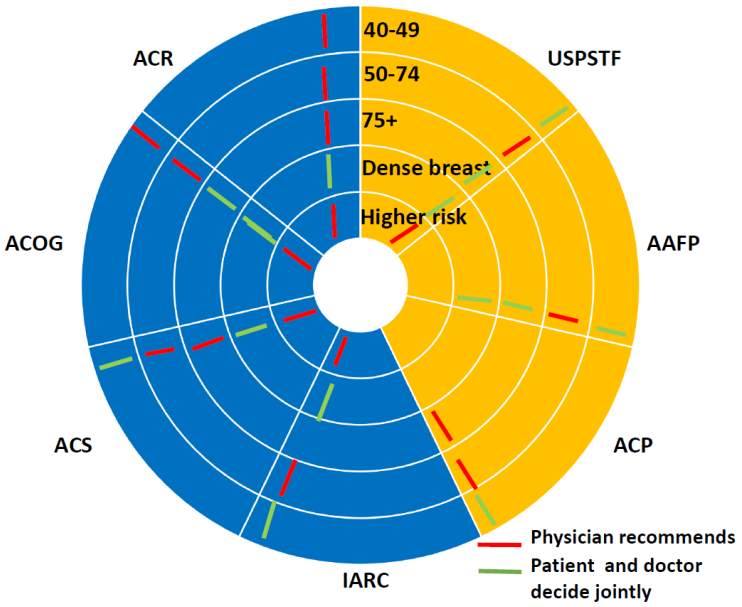
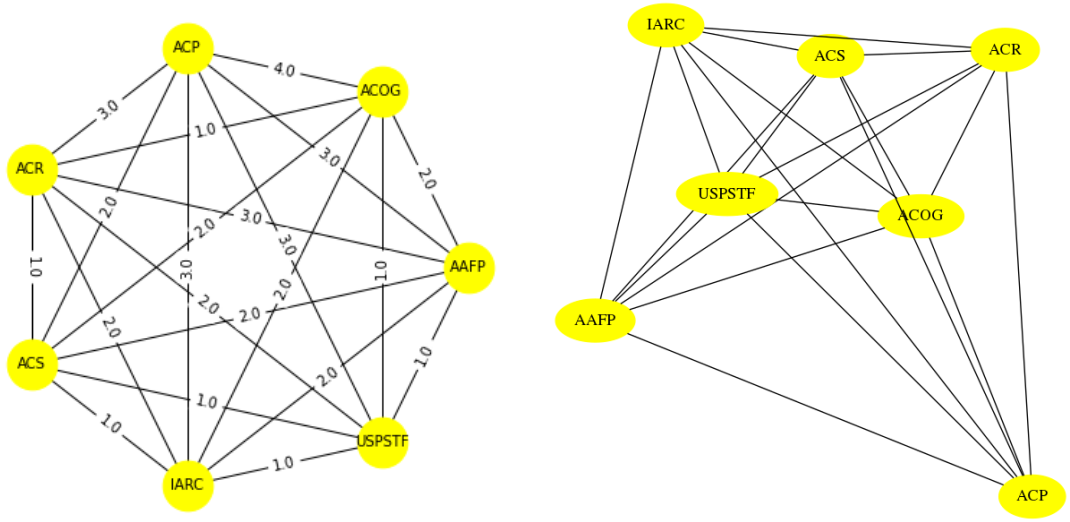


Figure 4.3: Similarities and disagreements in summarized recommendations. The yellow coloring shows patient making decisions, the blue coloring shows explicit screening recommendations. The concentric circles show different age groups. Red marks—physician recommends, green marks—patient decides.



(a) Number of differing features.

(b) Jaccard distances between summary recommendations, as per Table 4.1.

Figure 4.4: In panel (**a**) we see a pictorial representation of the numbers of differing features, per Tables 4.2 and 4.3. These differences between recommendations are converted into distances (using the Jaccard measure), resulting in panel (**b**). Can we replicate the geometric structure of panel (**b**) using automated tools? See Section 4.6 for an answer.

Table 4.3: Normalized distances between the summarized guidelines computed using Jaccard distances from Tables 4.1 and 4.2.

	AAFP	**ACOG**	**ACP**	**ACR**	**ACS**	**IARC**	**USPSTF**
AAFP	0	0.0238	0.0357	0.0357	0.0238	0.0238	0.0119
ACOG	0.0238	0	0.0476	0.0119	0.0238	0.0238	0.0119
ACP	0.0357	0.0476	0	0.0357	0.0238	0.0357	0.0357
ACR	0.0357	0.0119	0.0357	0	0.0119	0.0238	0.0238
ACS	0.0238	0.0238	0.0238	0.0119	0	0.0119	0.0119
IARC	0.0238	0.0238	0.0357	0.0238	0.0119	0	0.0119
USPSTF	0.0119	0.0119	0.0357	0.0238	0.0119	0.0119	0

4.3.1 Computing the Differences in Recommendations

Figure 4.3 is another representation of the information in the CDC comparison of the recommendations [7] earlier presented in Figure 4.1. It clearly shows the differences between the guidelines (and it comes from [140]). As we can see, there are two sides to the circle. The yellow side indicates the scenario where patients will likely decide when breast cancer screening should be done. The purple color side specifies the situation where breast cancer guideline providers most likely will demand screening interventions. White radial lines indicate boundaries between the different societies. The red color marks indicate that the physician decides. Green color marks indicate patients' decisions.

4.3.2 From Differences to Distances and a Graph

Table 4.1 represent the content of this analysis as a collection of features. Table 4.2 encodes these differences in recommendations as numbers of differing features between pairs of recommendations. Then, Table 4.3 shows the distances between the guidelines derived from Tables 4.1 and 4.2 using the *Jaccard distance* (the percentage of different

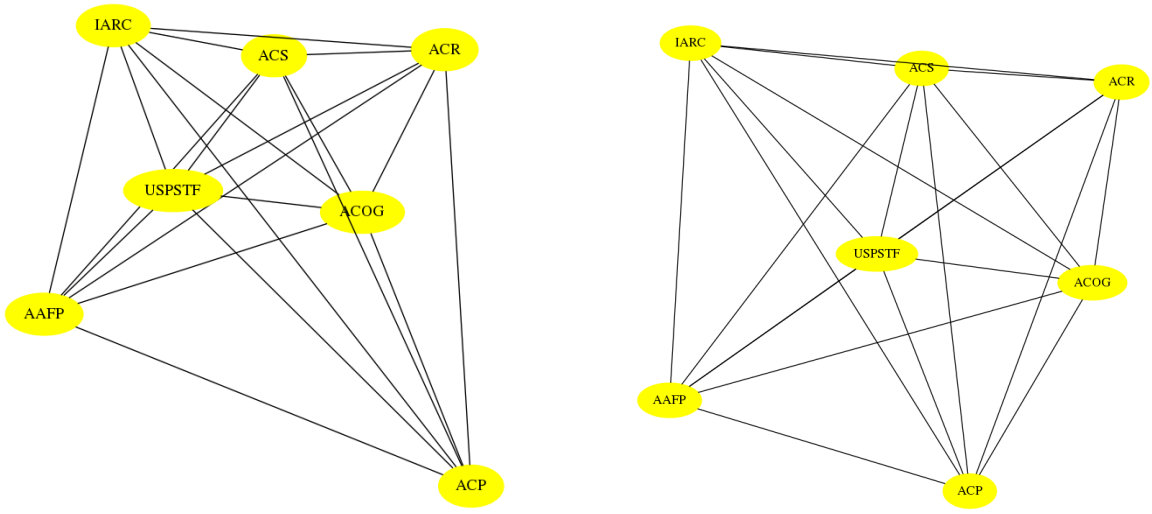
elements in two sets):

$$d_j(A, B) = 1 - \frac{\mid A \cap B \mid}{\mid A \cup B \mid}$$

Given two recommendation summaries A and B, we compute the number of the differing feature values from Table 4.2 and divide it by five. For example, for the pair (AFP, ACR), we get 3/5. All these distances were normalized to sum to 1 and shown in Table 4.3 (we are not assuming that distances are always symmetric. The normalization does not change the relative distances, and in the comparisons with the geometry of full documents, we only care about the relative distances.

Tables 4.1–4.3 represent the process of converting the information in Figure 4.3 into a set of distances. These distances are depicted graphically in Figure 4.4, where we display both Jaccard distances between the recommendations and the number of differing features as per Table 4.2.

In the following section, we will create a graph representation for the full documents (Figure 4.5b). We will present our graph comparison method in Section 4.5. In Section 4.6, we will assign numerical values to the distance between the two graphs, and show that this similarity cannot be the result of chance.



(a) Distances between the seven summary recommendation guidelines.

(b) Distances between full document guidelines using WM distance and BioASQ embeddings with concepts (see text for explanations) .

Figure 4.5: Visual comparison of the similarity/distance graphs based on human analysis is shown in panel (**a**), and computer generated comparison from Table 4.5 is shown in panel (**b**), which suggest a similar geometry. As we rigorously show in Section 4.6, this 69% *similarity is not accidental*; the distortion is about 31%. Notice that we are not pointing to the actual locations of similarities and differences in the guideline documents. Instead, we are pointing to global (latent) differences stemming from concepts appearing in them.

4.4 Transforming Full Guidelines Documents into Vectors and Graphs

In this study, we use both the CDC summaries ([7], reproduced and labeled in Figures 4.1 and 4.3), and the full text of the guidelines used by the CDC to create the summaries. The focus of this section is on the full guideline documents. The detailed information about these guidelines is shown in Table 4.4.

Note that we are using the same acronyms (of medical societies) to refer to full guideline documents in this section. This will not lead to confusion, as in this section, we are only discussing full documents.

Table 4.4: Guidelines with references. All the sources were last retrieved in summer 2020.

Guideline Abbreviation	Full Name of the Organization	Document Citation
ACOG	The American College of Obstetrics and Gynecology	[141]
AAFP	American Academy of Family Physicians	[142]
ACP	American College of Physicians	[143]
ACR	American college of Radiology	[144]
ACS	American Cancer Soceity	[145]
IARC	International Agency for Research on Cancer	[146]
USPSTF	United States Preventive services Task Force	[147]

4.4.1 Data Preparation for All Experiments

From the breast cancer screening guidelines listed in the CDC summary document [7], the texts of the USPSTF, ACS, ACP, and ACR guidelines were extracted from their HTML format. We used Adobe Acrobat Reader to obtain the texts from the pdf format of the AAFP, ACOG, and IARC guidelines. Since the AAFP documents also included preventive service recommendations for other diseases (such as other types of cancers), we added a preprocess step to remove those recommendations, leaving the parts matching "breast cancer".

4.4.2 Measuring Distances between Full Documents

When creating embedding representation of text, we replace each word or term with its embedding representation. Thus, the full guideline document texts are represented as a set of vectors. Our objective is to create a graph of conceptual distances between the documents.

The two most commonly used measures of distance, *cosine distance* and *word mover's distance*, operate on different representations. The former operates on pairs of vectors, and the latter on sets of vectors. Thus, we need to create two types of

representations.

Given a document, the first representation takes the average of all its word (term) embeddings. This creates a vector representing the guideline text. The second representation simply keeps the set of all its embedding vectors.

The *cosine distance* between two vectors v and w is defined as:

$$cosd(v, w) = 1 - cos(v, w)$$

We will also use the following variant of cosine distance to argue that the geometries we obtain in our experiments are similar irrespective of distance measures (see Section 4.6):

$$cosd'(v, w) = 1/cos(v, w) - 1$$

The *word mover's distance* (WMD, WM distance), introduced in [148], is a variant of the classic concept of "earth mover distance" from the transportation theory [149]. Sometimes, the term "Wasserstein distance" is also used. The intuition encoded in this metric is as follows. Given two documents represented by their set of vectors, each vector is viewed as a divisible object. We are allowed to "move" fractions of each vector in the first set to the other set. The WM distance is the minimal total distance accomplishing the transfer of all vector masses to the other set. More formally [148], WM distance minimizes:

$$\min_{T \geq 0} \sum_{i,j=1}^{n} T_{ij} c(i, j)$$

$$Subject\ to: \sum_{j=1}^{n} T_{ij} = d_i\ \forall i \in \{1, ..., n\}$$

$$\sum_{i=1}^{n} T_{ij} = d'_j\ \forall j \in \{1, ..., n\}$$

T_{ij} is the fraction of word i in document d traveling to word j in document d';

$c(i, j)$ denotes the cost "traveling" from word i in document d to word j in document d'; here the cost is the Euclidean distance between two words in the embedding space. Finally, d_i is the normalized frequency of word i in document d (and same for d'):

$$d_i = \frac{c_i}{\sum_{j=1}^{n} c_j}$$

We used the `n_similarity` and `wmdistance` functions from Gensim [150] as a tool for generating vectors and calculating similarities/distances in our experiments.

4.4.3 Building Vector Representations of Full Documents

However, there is more than one way to create word embeddings; we experimented with several methods. We used three language models of medical guidelines' disagreement: "no concept", conceptualized, and BioASQ. (The details of these experiments appear later in Table 4.6). The first two were Word2Vec embedding models trained using the PubMed articles as described in Section . We will describe these models' training process in Section 6.3.3. The third one used pre-trained BioASQ word embeddings created for the BioASQ competitions [151][2].

Our first model, trained on PubMed, included only words, and no additional conceptual analysis with MeSH[3] was done. In the second, which was a more complex model, MeSH terms were replaced with n-grams. For example, if `breast` and `cancer` appeared next to each other in the text, they were replaced with `breast-neoplasms` and treated as a concept.

4.4.4 Our Best Model: Using BioASQ Embeddings and Word Mover's Distance

Table 4.5 shows (unnormalized) WM distances between the seven guidelines using BioASQ embeddings. Figure 4.5 shows side by side the geometries of the two graphs: one generated from the summary of full documents, using features derived from the

[2] `http://BioASQ.org/news/BioASQ-releases-continuous-space-word-vectors-obtained-applying-word2vec-pubmed-abstracts`

[3] `https://www.nlm.nih.gov/mesh/meshhome.html`

CDC summaries, and the second one based on the machine-generated representations of the full guideline documents. To create Figure 4.5, for each metric, a diagram representing the distance between the nodes (guidelines) and a diagram with the labeled edges were drawn using the networkx library[4]. All values were normalized to the same scale to allow visual comparison.

Table 4.5: This table shows the word mover's distances between the guidelines using BioASQ embeddings. This model also performed very well on the datasets in the section 4.7.

	AAFP	**ACOG**	**ACP**	**ACR**	**ACS**	**IARC**	**USPSTF**
AAFP	0.0	1.833953	1.903064	1.994837	1.866007	2.153458	1.681802
ACOG	1.833953	0.0	1.649276	1.290215	1.333061	1.773604	1.286168
ACP	1.903064	1.649276	0.0	1.856171	1.667579	1.956002	1.674375
ACR	1.994837	1.290215	1.856171	0.0	1.41020	1.873691	1.385404
ACS	1.866007	1.333061	1.667579	1.41020	0.0	1.676928	1.163601
IARC	2.153458	1.773604	1.956002	1.873691	1.676928	0.0	1.753758
USPSTF	1.681802	1.286168	1.674375	1.385404	1.163601	1.753758	0.0

The similarity is visible in a visual inspection, and will be quantified in Section 4.6 to be about 70%. However, before we provide the details of the experiments, we will also answer two questions:

— How do we measure the distortion/similarity between the two graphs?

— Could this similarity of shapes be accidental? How do we measure such probability?

4.5 Graph-Based Method for Comparing Collections of Documents

At this point, we have created two graphs, one showing the distances between summary recommendations, and the other representing conceptual distances between

[4] `https://networkx.github.io/`

documents. The procedure we used so far can be concisely expressed as Algorithm 1, where given a set of documents, after specifying the `Model` (type of embeddings) and a `distance` metric, we get an adjacency matrix containing the distances between the nodes representing the documents. An example output of Algorithm 1 is shown in Figure 4.4 above.

What remains to be done is to quantify the difference in shapes of these two graphs, and then to show that the similarity we observe is not accidental. The methods used in these two steps are described in Algorithms 2 and 3. The experiments and the details of the performed computations will be presented in Section 4.6.

Algorithm 1 Computing Graph of Distances Between Documents.

Input: `Guidelines`: a set of guideline documents in textual format.

`Model`: a model to compute distances between two documents.

Output: $\mathcal{A}_G$ — Adjacency matrix of distances between document guidelines.

1: **for** each pair of documents in `Guidelines` **do**
2: Compute the `distance` between the documents according to `Model`
3: Put the `distance` in $\mathcal{A}_G$
4: **end for**
5: **return** $\mathcal{A}_G$

We use a very natural, graph clique-based method for comparing the similarity of two *collections* of documents. Given two sets of documents represented by graphs, and a one-to-one mapping between nodes, in Algorithm 2, we compute the percent distortion between the shapes of the two cliques—this is perhaps the most natural similarity measure (`similarity = 1 − distortion`) for comparing the shapes of two cliques of identical cardinality.

Algorithm 2 Distance or Percentage Distortion between Two Complete Graphs (cliques of the same size).
Note. For example, the distance between the two graphs in Figure 4.5 is 0.31, equivalent to 31% distortion

Input: Adjacency Matrices $\mathcal{A}_1$, $\mathcal{A}_2$ of equal dimensions

Output: Graph distance/distortion $\mathcal{D}(\mathcal{A}_1, \mathcal{A}_2)$, as a value between 0 and 1.

1: Normalize the distances in $\mathcal{A}_1$ (by dividing each distance by the sum of distances in the graph) to produce a new adjacency matrix $\mathcal{AN}_1$

2: Normalize the distances in $\mathcal{A}_2$ to produce a new adjacency matrix $\mathcal{AN}_2$

3: Set the value of *graph_distance* to 0.

4: **for** each `edge` in $\mathcal{AN}_1$ **do**

5: Add the absolute value of the difference between the `edge` length and its counterpart in $\mathcal{AN}_2$ to the *graph_distance*

6: **end for**

7: **return** $\mathcal{D}(\mathcal{A}_1, \mathcal{A}_2) = graph\_distance$

Next, we need to compute the chance that the mapping arose from a random process. This is because if the chances of the similarity arising from a random process are small, we can conclude that a full document's conceptual vocabulary determines the type of recommendation given by a particular organization. In our case, the nodes of both graphs have the same names (the names of the medical societies), but the shapes of the graphs are different, one coming from human summaries and comparison (Figure 4.1, Table 4.1) and the other from a machine produced conceptual distances. Thus, the randomization can be viewed as a permutation on the nodes. When such permutations do not produce similar structures, we can conclude the similarity of the two graphs in Figure 4.5 is not accidental.

Next, in Algorithm 3, we compute the average distortion, and the standard deviation of distortions, under the permutation of nodes. The input consists of two cliques of the same cardinality. The distance measure comes from Algorithm 2.

Algorithm 3 Computing Graph Distortion Statistics.
The input is two cliques of the same cardinality.

Input: Normalized Adjacency Matrices $\mathcal{N}_1$, $\mathcal{N}_2$ of equal dimensions

Output: Baseline for the graph distance, standard deviation of graph distances under permutations of computed distances.

1: Set the value of *graph_distances* to an empty list.

We are permuting the labels of graph, leaving the lengths of the edges intact.

2: **for** each permutation $\mathcal{N}_2 p$ of the nodes of $\mathcal{N}_2$ **do**

3: Compute $d = \mathcal{D}(\mathcal{N}_1, \mathcal{N}_2 p)$ using Algorithm 2

4: Append d to *graph_distances*

5: **end for**

6: Set

$$graph\_distance\_baseline = Mean(graph\_distances)$$

$$std = StandardDeviation(graph\_distances)$$

7: **return** *graph_distance_baseline*, *std*

4.6 Details of Experiments and Their Results

In Section 4.4 we described the procedure of creating the graph of full documents and in Section 4.4.4 we referenced the best model, although the details of the methods were presented in Section 4.5. This was not the only model we tried, and we will now discuss other experiments; they all support the conclusion of the non-accidental similarity of the graph of recommendations and the graphs of concepts. (As shown later in Section 4.7, this model also performs very well on other sets of guidelines).

4.6.1 Steps Used in All Our Experiments and Evaluation

In all our experiments we used the procedure in Algorithm 2 to compute the distance/distortion between the two labeled graphs, using the matrix of conceptual distances between full documents and the matrix in Table 4.3. As mentioned earlier, for

our best model the distortion was 0.31; therefore, the similarity was 0.69 (or 69%). We then asked the question: Could this distortion be accidental? In other words, could it be the case that we were lucky? If so, how lucky would we have to be? Since the distance between nodes of both graphs are fixed (in a given experiment), the only variable we can manipulate is the mapping from the nodes of one graph to another. In other words, if we did not have the labels, what are the chances of finding the right match from all possible labelings. We thus asked: Can other mappings produce similar results? To answer this question, we computed the average distortion and the standard deviation, based on all possible permutation of nodes ($5040 = 7!$ permutations). The pseudo-code for this computation is shown in Algorithm 3.

In all experiments, the difference between our results and average distortion was seven (or more) standard deviations. Therefore, we can conclude the that the matching of the two geometries is not accidental and is highly significant.

4.6.2 Results of the Experiments

In this section we first discuss the statistical properties of the experiments to show that our models capture statistically significant geometric correspondences between the graph of recommendation summaries and the graph of conceptual distances between the full document guidelines. Table 4.6 shows results of the main series of experiments we performed. Additional experiments are reported in section 4.7.

Table 4.6: This table shows the values obtained in multiple experiments. Column 2, `Distortion`, shows the distortions of graphs produced using corresponding models from Column 1. Average distortions per permutation are shown in Column 3. `STD` is the standard deviation of the distortion per permutation of vertices. Note that the distortion is somewhat depended on how we measure distances; however, the shapes of the distributions are very similar. (The cosine measures are capitalized for readability).

Model	Distortion	Distortion of Permutations	STD
BioASQ_WMD	0.31393366	0.38137817	0.00901798
Conceptualized_WMD	0.33504400	0.39118512	0.00929325
NoConcept_WMD	0.34457155	0.38822718	0.00909964
BioASQ_CosD	0.41787106	0.59569767	0.01572929
Conceptualized_CosD	0.53452523	0.61350075	0.01626678
NoConcept_CosD	0.51399564	0.59093162	0.01538653
BioASQ_CosD'	0.39343054	0.57170607	0.01494240
Conceptualized_CosD'	0.47697532	0.55849892	0.01458596
NoConcept_CosD'	0.47889093	0.55465835	0.01434584

Table 4.6 shows the results of the experiments with full text of the guidelines. For our best model, BioASQ_WMD, we found a 69% similarity (top line), or 0.31 distance (distortion). As can be seen, the average distortion of permutations (using the distances produced by BioASQ_WMD) is 38%; however, the standard deviation of the distortions is less than 1%. Thus, the distance between our model and the mean is about seven standard deviations. Therefore, we conclude that the similarity between the shapes of the two graphs is extremely unlikely to be coincidental. Hence the model represents a non-trivial similarity. Moreover, we performed the same kind of analysis using different models, i.e., different embeddings and different distance measures. Note that there is no natural transformation of WM distance applicable here. Additionally, while the distances and distortions change, the chances of similarities arising by accident are always smaller than 1/1000 (four standard deviations

from the mean of distortions). By this standard statistical criterion, no matter what measures of distance we use, the similarity between two graphs, one from human analysis [7] and the other from automated concept modeling, is non-trivial and not accidental. This observation is amplified by the additional experiments reported in section 4.7. We conclude that vector-based representation are capable of detecting conceptual differences, i.e., the types and densities of concepts brought to the writing of medical recommendations.

4.7 Additional Experiments

We performed several additional experiments, and we report on three of them in this section. The first experiment is a variant of the one described above using the CDC table in Figure 4.1. The second experiment is on lower back pain management guidelines, for which we could find an online summary table similar to the one in Figure 4.1. The third one is applying the graph comparison not to a table, but to the guideline abstracts. We could not find a tabular comparison of the hypertension management guidelines, so instead, we compared the concepts in full and abstracted texts. This shows the potential applicability of the proposed approach to other situations, where we might be interested in conceptual comparisons of related collections of documents.

4.7.1 Experimenting with the Full Texts of the Guidelines

We performed additional experiments with *modified* views of the full guideline documents, as enumerated below. This was driven by the fact that the levels of distances between full documents may change if we compute the similarities/distances between selected sentences, which are explicitly related to the statements from the CDC table in Figure 4.1. For these additional experiments, we split each full text guideline document into two different subsets:

1. `Related:` containing sentences that are related to the CDC table by having

common concepts, as represented by UMLS concepts. This was done in multiple ways, giving us six possible experiments:

(a) The CDC recommendations table was considered as a single bag of concepts. If a sentence in the full text had a *minimum* number of mutual concepts with this bag, that sentence was considered a related sentence.

(b) If a sentence in the full text had a *minimum* number of mutual concepts with at least one statement from the CDC table (again, viewed as a bag of concepts), that sentence was considered a related sentence.

Different minimum numbers of mutual concept(s) were examined in our experiment; that is the *minimum* was set to 1, 2, and 3.

2. `Unrelated:` the other sentences.

`Unrelated` sentences were not used for these additional experiments.

Concept extraction: For all experiments, we used MetaMap [5] to extract UMLS concepts [6] and semantic types [7] in sentences. We only considered concepts with informative (in our opinion) semantic types. This meant using concepts related to diagnosis and prevention, for example "findings", and not using ones related, e.g., to genomics. Our final list had the following: [[diap], [hlca], [dsyn], [neop], [qnco], [qlco], [tmco], [fndg], [geoa], [topp], [lbpr]].

[5] `https://metamap.nlm.nih.gov/`
[6] `https://www.nlm.nih.gov/research/umls/index.html`
[7] `https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html`

Table 4.7: **Using sentences in recommendations and minimum mutual concepts**. This table shows the values obtained in additional experiments, where full document guidelines were modified by attending to concepts in sentences (see above). Column 1 refers to the number of concepts overlapping with summaries. `Distortion` shows the distortions of graphs produced using corresponding models from Column 1. As before, in Table 4.6, the distortion depends on how we measure the distances; however, the shapes of the distributions are very similar.

Min. Mutual Concepts	**Model**	**Distortion**	**Distortion of Permutations**	**STD**
1	BioASQ CosD	0.526380991	0.602890558	0.011735664
	Conceptualized_CosD	0.635564038	0.646721788	0.011417208
	NoConcept_CosD	0.626087519	0.646906954	0.011131221
	NoConcept_WMD	0.352402031	0.383852647	0.006550777
	Conceptualized_WMD	0.359296888	0.390059373	0.006626223
	BioASQ_WMD	0.336903254	0.384735148	0.006498348
2	BioASQ_CosD	0.449264689	0.572620976	0.010916054
	Conceptualized_CosD	0.384945443	0.488740293	0.008608367
	NoConcept_CosD	0.433167046	0.501788823	0.008699466
	NoConcept_WMD	0.34284288	0.376371094	0.006467164
	Conceptualized_WMD	0.330059701	0.373155641	0.006466969
	BioASQ_WMD	0.32446554	0.38365857	0.006428759
3	BioASQ_CosD	0.468163076	0.537093759	0.010040669
	Conceptualized_CosD	0.564019791	0.57488789	0.010091071
	NoConcept_CosD	0.594326474	0.596293202	0.010300973
	NoConcept_WMD	0.360513492	0.375067469	0.006461442
	Conceptualized_WMD	0.37193217	0.383126986	0.006477258
	BioASQ_WMD	0.34276229	0.375886963	0.006455091

For full text guidelines (as per Table 4.4), the results of the experiments are shown in Table 4.6 and discussed in Sections 4.6 and 4.9. Tables 4.7 and 4.8 are based on the same type of comparisons as discussed in Section 4.6, except that we subtract the `Unrelated` sentences from the full guidelines. Again, we observed that the similarity is not accidental and that BioASQ embeddings with WM distance seem on average to give the best performance.

Note the potentially important observation about Tables 4.6, 4.7 and 4.8: They jointly show that the property we investigate, i.e., the conceptual distances between guidelines, is indeed geometric, and therefore, the word "distances" is not merely a metaphor. The correspondence between the two graphs is preserved no matter how we set up the experiments. That is, as with geometric properties such as being collinear or parallel, the structure remains the same when a transformation (such as a projection) is applied to the points, even though some of the measurements might change (e.g., measured distances, or the area of a parallelogram). The same happens when we transform the documents by removing `Unrelated` sentences: the values of distortions change, but the non-accidental correspondence with the summary graph (Figure 4.5) remains invariant.

Table 4.8: **Using the whole summary recommendations and minimum mutual concepts**. This table shows the values obtained in additional experiments, where the whole CDC summary was used to obtain sets of mutual concepts (see above). Column 1 refers to the number of concepts overlapping with the summary. `Distortion` shows the distortions of graphs produced using corresponding models from Column 1. As before, in Tables 4.6 and 4.7 the distortion is somewhat depended on how we measure distances; however, the shapes of the distributions are very similar.

Min. Mutual Concepts	**Model**	**Distortion**	**Distortion of Permutations**	**STD**
1	BioASQ_CosD	0.550516174	0.534742406	0.007178113
	Conceptualized_CosD	0.568149311	0.547872282	0.007613218
	NoConcept_CosD	0.559332088	0.54286484	0.007445151
	NoConcept_WMD	0.351230589	0.388633467	0.006465622
	Conceptualized_WMD	0.346202932	0.389016657	0.006561466
	BioASQ_WMD	0.320392721	0.38253681	0.006475659
2	BioASQ_CosD	0.553091569	0.536791251	0.00725238
	Conceptualized_CosD	0.558005588	0.543056307	0.00740679
	NoConcept_CosD	0.550200164	0.539443354	0.007298594
	NoConcept_WMD	0.341053017	0.380095604	0.006485268
	Conceptualized_WMD	0.328265638	0.378358521	0.006481775
	BioASQ_WMD	0.323598367	0.386020859	0.006486291
3	BioASQ_CosD 0.548898679	0.536773761	0.007261816	
	Conceptualized_CosD	0.555658633	0.544321589	0.007471369
	NoConcept_CosD	0.548497913	0.540891385	0.007362149
	NoConcept_WMD	0.351294868	0.377266094	0.006478541
	Conceptualized_WMD	0.352791102	0.37921564	0.006506027
	BioASQ_WMD	0.337147756	0.38514511	0.006439097

4.7.2 Lower Back Pain Management Guidelines

In this experiment, we used the summary tables on "clinical practice guidelines for the management of non-specific low back pain in primary care" from [1]. In the cited paper, several comparisons are made between 15 clinical practice guidelines from multiple continents and countries (Africa (multinational), Australia, Brazil, Belgium, Canada, Denmark, Finland, Germany, Malaysia, Mexico, the Netherlands, Philippine, Spain, the USA and the UK). In our experiments we used all of those for which an English text was available: (GER) [152], (MAL) [153], (SPA) [154], (UK) [155], (AUS) [156], (USA) [157], (CAN) [158], (DEN) [159], and (BEL) [160]. For this total of nine guideline texts, we experimented with Table 1 (describing methodologies for diagnosis) and Table 2 (treatment recommendations) from the article [1] containing, respectively, 12 and 60 features; in addition, we created a super-table combining the two tables and applied our method to it as well.

With the same process as described in Section 4.3 we converted the Tables 1 and 2 of [1] into Jaccard distances. Then, as before, we computed the distortion between the graphs of the full text and the graphs of distances between the extracted features; and, as before, we established that the probability of obtaining high similarity by chance is extremely small. For Table 1 of [1] our best model BioAsq_WMD produced about 28% distortion (or 72% similarity). Similar results hold for Table 2 and for the combined table, although the actual distortion numbers differ. In all cases, for the model BioAsq_WMD we found about 10-fold standard deviation, with distortion of about 16% for Table 2 and about 14% for the aggregated tables combining Tables 1 and 2 of [1].

All other models used in Table 4.6 performed in line with the previous results, with the only exception being the conceptualized models for Table 1 of [1], where for Conceptualized_CosD and Conceptualized_CosD' the distortion was slightly worse than random. We do not have an explanation for this subpar performance, but we

have seen a relatively weak performance of this model in Table 4.7. Table 4.9 shows the Jaccard distances and Table 4.10 shows the performance of all models on the combined table. Thus the performance of the model does not seem to degrade with a large number of comparisons.

Table 4.9: Jaccard distances based on the combined Tables 1 and 2 from [1]. The guidelines are about the management of non-specific lower back pain.

	US	**DEN**	**MAL**	**CAN**	**BEL**	**GER**	**UK**	**SPA**	**AUS**
US	0	46	43	35	39	32	43	41	46
DEN	46	0	50	45	47	44	47	49	39
MAL	43	50	0	33	36	39	40	36	42
CAN	35	45	33	0	33	23	33	34	32
BEL	39	47	36	33	0	26	15	36	38
GER	32	44	39	23	26	0	30	33	35
UK	43	47	40	33	15	30	0	41	36
SPA	41	49	36	34	36	33	41	0	44
AUS	46	39	42	32	38	35	36	44	0

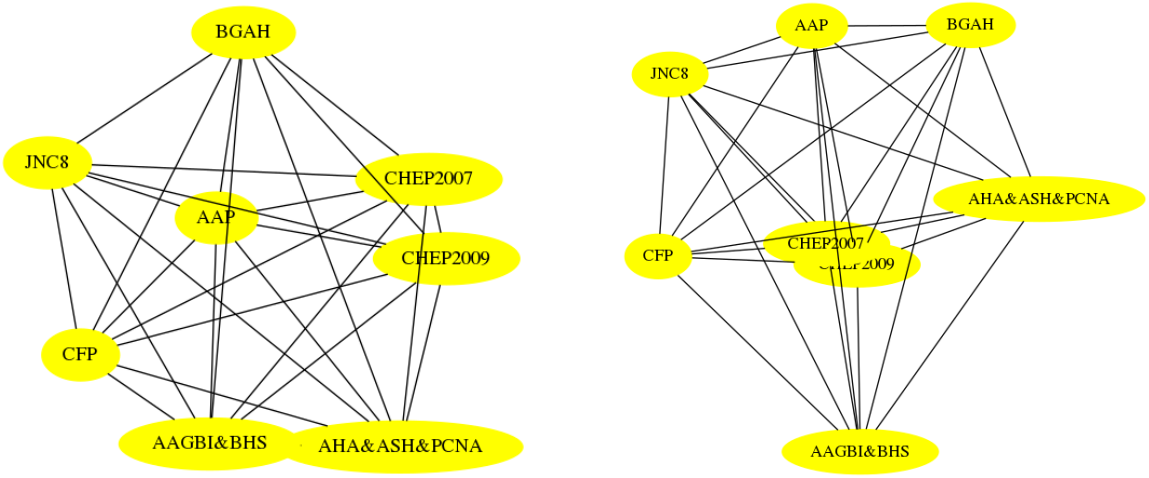
Table 4.10: The performance of the algorithms on the combined Tables 1 and 2 from [1] is in line with the results in Section 4.6.2, except for the weaker showing of the Conceptualized_WMD model.

Model	Distortion	Distortion of Permutations	STD
BioASQ_WMD	0.14157219	0.16717125	0.00186797
Conceptualized_WMD	0.15689518	0.16098291	0.00179856
NoConcept_WMD	0.13946498	0.16067458	0.00180108
BioASQ_CosD'	0.44899108	0.49891074	0.00595033
Conceptualized_CosD'	0.31577959	0.35477261	0.00389520
NoConcept_CosD'	0.27595783	0.33897971	0.00418504
BioASQ_CosD	0.40412785	0.45347124	0.00511851
Conceptualized_CosD	0.28283583	0.32039879	0.00342097
NoConcept_CosD	0.25530361	0.31717921	0.00378427

4.7.3 Comparing Hypertension Management Guidelines

In an additional experiment, we used a collection of hypertension management guidelines from different countries, including the USA, Canada, Brazil, the UK and Ireland [161, 162, 163, 164, 165, 166, 167, 168]. The corpus was created by searching PubMed for *'practice guideline'* as "publication type" and *'hypertension'* and as the "major MeSh" index. We selected eight of them from different medical bodies, where the guidelines' full texts were available. This corpus consists of the following eight documents: CHEP2007 [161], the 2007 Canadian Hypertension Education Program; AHA & ASH & PCNA [162], joint statement of the American Heart Association, American Society Of Hypertension, and Preventive Cardiovascular Nurses Association; BGAH [163], the Brazilian Guideline of Arterial Hypertension; CFP [164], the 2013 Canadian screening recommendations; AAGBI & BHS [165], the 2016 joint British and Irish guidelines; CHEP2009 [168], the 2009 Canadian Hypertension Education Program; AAP [166], 2017 guidelines focusing on children and adolescents; and JNC [167], the 2014 evidence-based guidelines focusing on adults.

Because we are not aware of any tabular summary of differences between hypertension guidelines, similar to the one shown earlier in Figure 4.1, we made the comparisons between full texts of the guidelines and their abstracts. That is, we created two graphs of embeddings, as shown in Figure 4.6, and measured their similarity, as well as the probability of the similarity arising by chance, as shown in Table 4.11. The experiment shows that the concepts appearing in the guidelines' abstracts strongly correlate with the concepts used in the guidelines' full texts. Moreover, the method, described earlier in Section 4.5, which we used to find this correspondence was very good at picking up this similarity; and, as before, a very good model was obtained by using BioASQ embeddings with the Word Mover Distance (WMD).



(a) The graph of the conceptual distances between the full texts of hypertension guidelines.

(b) The conceptual distances between the abstracts of the hypertension guidelines.

Figure 4.6: For the graphs of the eight hypertension guidelines and their abstracts a visual comparison is more difficult than it was earlier in Figure 4.5. Therefore, we need a quantitative comparison, which is given in Table 4.11.

Table 4.11: We see the robustness of the proposed method when comparing the abstracts' conceptual distances and the full documents of guidelines.

Model	Distortion	Distortion of Permutations	STD
BioASQ_WMD	0.10722113	0.15779628	0.00339429
Conceptualized_WMD	0.22486228	0.30471905	0.00416548
NoConcept_WMD	0.10659179	0.15903103	0.00339359
BioASQ_CosD'	0.63552553	0.67228101	0.00891151
Conceptualized_CosD'	0.4750315	0.62258280	0.00850719
NoConcept_CosD'	0.51297572	0.54567338	0.00727428
BioASQ_CosD	0.53894790	0.58653238	0.00754606
Conceptualized_CosD	0.34443154	0.47391939	0.00608174
NoConcept_CosD	0.42284040	0.45815692	0.00609919

4.8 Discussion

Our broad research objective is to create a computational model accurately representing medical guidelines' disagreements. Since the creation of such accurate models is beyond the current state of the art, in this study, we focused on an approximation, i.e., a model that is simple and general enough to be potentially applicable in other situations and which was useful for the question at hand, namely, whether conceptual vocabulary determines recommendations.

As mentioned earlier, this work was partly motivated by epistemology of disagreement and medical disagreement, viewed as "near-peer" disagreement. Our results show that it is possible to build computational models of "near-peer" disagreement. Additionally, they provide support for the empirical observations of disagreement adjudication among medical experts [134, 135], where the authors observe that the differences in experts' backgrounds increase the degree of disagreement.

A limitation of the study lies in testing the proposed method on a small number of case studies. In the main track, we focused on the CDC summaries of the breast cancer screening guidelines, and, in section 4.7, we discuss our experiments on the

lower back pain management and hypertension guidelines.

We showed that the method is robust in the case of these sample guidelines because even with the change of metrics, the similarities remain statistically significant. However, this study only describes a few case studies and leaves it as an open question of whether it will work equally well in other cases.

Unlike our earlier work [6], in this study, we have not performed any logical analysis of the guidelines. We also did not use text mining to extract relations from the guidelines' content. Although our focus was on concepts appearing in guidelines, we did not point to specific vocabulary differences. Instead, we measured semantic differences between guidelines using the distances between their vectorial representations. This has to do with the fact that, even though NLP methods have progressed enormously over the last decade [125], they are far from perfect. In our experiments, we used some of the simplest semantic types of words and simple collocations represented as vectors in high-dimensional spaces. This simplicity is helpful, as we can run several experiments and compare the effects of using different representations and metrics. This gives us the confidence that the similarities we are discovering tell us something interesting about guideline documents.

4.9 Conclusions

This work investigates the question of whether the disagreements in medical recommendations, for example, in breast cancer screening or back pain management guidelines, can be attributed to the differences in concepts brought to the problem by specific medical societies (and not,e.g., the style or formalization of recommendations). Our experiments answered this question in the affirmative and showed that a simple model using word embeddings to represent concepts could account for about 70% to 85% of disagreements in the recommendations. Another contribution is the articulation of a very natural graph clique-based algorithm/method for comparing the similarity of two collections of documents. Given two sets of documents, each

of the same cardinality, and a mapping between nodes, we computed the percent of distortion between the shapes of the two cliques and the chances that the mapping arose from a random process. We also documented all of the steps of the process and provided the data and the code (`https://github.com/hematialam/Conceptual_Distances_Medical_Recommendations`) to facilitate both extensions of this work and its replication.

Our work extends the state-of-the-art computational analysis of medical guidelines. Namely, instead of semi-automated conceptual analysis, we demonstrated the feasibility of automated conceptual analysis. That is, in our study, we used a representation derived from a (relatively shallow) neural network (BioASQ embeddings [151]), and knowledge-based annotations derived from MetaMap[8]. Our results, detailed in Section 4.6 and in section 4.7, show that both can be useful as representations of our set of guidelines. Overall, they show similar performance in modeling conceptual similarities. However, the BioAsq_WMD model, using the BioASQ embeddings and the Word Mover's Distance, seems to be most stable, as it performed very well in all our experiments.

Although this study is a collection of three case studies, bound by a common method, it could be a good starting point for an analysis of other medical guidelines and perhaps other areas of expert disagreement. The methods described in this chapter are easy to use and rely on well-known tools such as word embeddings and MetaMap. They can also be extended and improved to produce more accurate and deeper analyses due to the fast progress in text mining and deep learning techniques. From the point of view of methodology of analyzing medical guidelines, this work contains the first computational implementation of the "near-peer" model mentioned earlier. To our knowledge, ours is the first proposal to use automated methods of text analysis to investigate differences in recommendations.

[8] `https://metamap.nlm.nih.gov/`

CHAPTER 5: IDENTIFYING CONDITIONAL AND CONDITION-ACTION STATEMENTS IN MEDICAL GUIDELINES

This chapter proposes an automated extraction method using linguistic features to identify informative Statements, conditional and condition-action statements, and medical guidelines. The process includes three steps. In the first step, candidate sentences are selected based on a modifier word's appearance (e.g., "if") for conditions. In the second step, linguistic features are extracted from candidate sentences using Information Extraction methodologies. In the third step, we use supervised machine learning techniques to classify candidate sentences as to whether they express conditions and actions. With a domain expert's help, we annotated three sets of guidelines to create gold standards to measure our condition-action extracting models' performance. The sets of guidelines are: hypertension [167], chapter 4 of asthma [41], and rhinosinusitis [169]. chapter 4 of asthma guideline was selected for comparison with prior work of Wenzina and Kaiser [8]. The method was evaluated by extracting conditional and condition-action statements from these guidelines using several machine learning methods.

5.1 Introduction

Clinical decision support systems (CDSSs) typically address two major tasks: diagnosis — determining "what is true "about a patient; and recommendation — determining "what to do (or not)" for the patient. Medical guidelines provide the conceptual link between a diagnosis and a recommendation. For example, they may include sentences such as this:

"*In the population aged 18 years or older with CKD and hypertension*,

`initial (or add-on) antihypertensive treatment should include an ACEI or ARB to improve kidney outcomes"`

The italics show the diagnosis part, i.e., a *condition*, and the courier font a recommendation, i.e., an *action*. This chapter focuses on automated identification of *condition-action* sentences in medical guidelines. We present results of three studies, which use different text analytics techniques, and show that:

- Modern deep learning techniques using attention-based models give substantial improvements in accuracy (6-11%) and F1-score (17-25%) over earlier machine learning methods.

- Transfer learning can potentially be used on text of medical guidelines in new domains, even with small amounts of available training data. Namely, training on two guideline documents produces results better than hand-coded rules and comparable to standard machine learning methods, even though they only have 445 words in common, and their distributions are completely different.

The three studies use, respectively, syntactic, semantic, and deep learning methods, evaluated on a set of three annotated medical guidelines. Our main contribution is to show the applicability of the recently developed techniques, namely neural network transformers and transfer learning, to this particular problem, and in comparing them with alternatives based on older machine learning techniques.

In another contribution, we have released two annotated guidelines used in these experiments, adding to the one previously published data set of [8].

5.1.1 Motivation

There are over 35,000 clinical practice guidelines indexed by PubMed[1], with over 1500 appearing every year [107]. Such guidelines may disagree on their recommendations, as documented in prior work, including ours, [7, 170, 6]. Controversies over

[1](`https://pubmed.ncbi.nlm.nih.gov/`)

prostate screening (PSA), breast cancer screening, hypertension, and other treatment and prevention guidelines are well-known. In addition, clinical recommendations are often in conflict when managing comorbidities [171].

Notice that the disagreements focus on actions, i.e., what to do in particular situations (conditions of the patient). For example, for what ages and breast conditions should mammography be recommended.

We believe patient outcomes would be improved, overtreatment would be reduced, and possibly better processes for creating treatment guidelines could be established if only we could better reason about individual guidelines and guidelines corpora. In particular, it is natural to imagine decision support systems for healthcare professionals [23] accessing properly indexed and contextualized condition-action statements. Therefore, we should understand whether, and how, such condition-action statements can be automatically extracted from texts.

5.1.2 Organization of this Chapter and Brief Description of the Studies

After establishing some preliminaries and discussing related work in Section 5.2, we will methodically describe each of our set of experiments. In each study, multiple machine learning models are evaluated. Our studies progress through different methods of machine learning, starting with learning patterns based on part-of-speech (Study 1), adding syntactic and semantic information (Study 2), and several experiments in transfer learning using deep learning methods (Study 3).

In these studies, we focus on *condition-action* (CA), *condition-consequence* (CC), and *action* (A) sentences. For comparison with other works, and for easier summarization, we will refer to these classes as *conditional* sentences; and when finer distinctions are needed, we will use abbreviations. Thus, the class consisting of CC and CA classes will be abbreviated as CCA; and in a few instances, we will also discuss class CCA+A. With this naming convention, we now can summarize the topics of studies:

Study 1. (Described in Section 5.4.) Identifying conditional and condition-action statements using domain-independent *syntactic features*, part-of-speech (POS) tags.

Study 2. (Described in Section 5.4.3.) Identifying conditional and condition-action statements using both domain-independent features and UMLS *semantic types*.

Study 3. (Described in Section 5.5.3.) Experimenting with *deep learning, domain adaptation, and machine learning transfer learning*.

- Experiment 1. Identifying conditional and condition-action statements using pretrained transformer models.
- Experiment 2. Identifying conditional and condition-action statements using pretrained transformer models and features from Study 1 and Study 2.
- Experiment 3. Repeating Experiments 1 and 2 by training classifiers on two guidelines (rhinosinusitis+hypertension), and testing them on the third guideline (asthma).

All these experiments use three clinical guidelines: asthma, rhinosinusitis, and hypertension.

The data and its preparation are discussed in Section 5.3. Afterwards, we describe the experiments in Sections 5.4– 5.5. In particular, Section 5.5.3 describes several experiments in domain adaptation and transfer learning. Discussion and Conclusions follow in Sections 5.6 and 5.7.

5.2 Preliminaries and Related Work

This work focuses on a specific text analytics problem, namely, on finding sentences with condition-action recommendations in medical guidelines. For this purpose, we

use a variety of classification techniques ranging from traditional methods such as logistic regression and random forest to new deep learning and transfer learning methods introduced in the last few years.

Its motivation, as stated earlier, comes from two directions: clinical decision support and natural language processing. As such, it belongs to the broad category of applications of artificial intelligence in medicine, and in particular, in clinical decision support.

The specific techniques of supervised machine learning used in this chapter comprise both the classical approaches such as logistic regression and random forest, and more recently introduced deep learning methods involving domain adaptation and transfer learning. All of these are applied to the task of classifying sentences as to whether they express a condition and action or not.

Therefore, in the preliminaries, we need to cover both topics: natural language processing of medical text, as well as the newer machine learning techniques, including their applications to medical text, also in the context of clinical decision support.

5.2.1 Five Decades of Automated Analysis of Medical Texts

Text analysis of medical records is already mentioned in a 1975 article by N.Sager [172, 173], and included extracting information to populate relational databases. Over the five decades of research in this space, many new techniques have been developed and applied to medical texts.

Even though no system can claim to "really" understand natural language, the progress has been very fast in the last ten years [125], where in several domains, such as data extraction, classification and question answering, automated systems have dramatically improved their performance, in some cases performing better than humans. This progress is chiefly due to the unmatched pattern recognition and memorization capabilities of deep neural networks (see, e.g., [126] for an overview).

If we focus on medical texts, we see that modern NLP methods have been applied

to clinical decision support, e.g., [120], to clinical trials [121], to automatic extraction of adverse drug events and drug-related entities [122], and to other areas [174, 175].

5.2.2 Analysis of Medical Guidelines

A recent overview of research on clinical guidelines and its applications can be found in [176, 177]. However, over the past few decades, many problems have been encountered and approaches have been tried to represent and execute clinical guidelines over patient-specific clinical data. They include document-centric models, decision trees and probabilistic models, and "Task-Network Models" (TNMs) [178], which represent guideline knowledge in hierarchical structures containing networks of clinical actions and decisions that unfold over time. A general-purpose architecture for syntax-semantic translation of medical guidelines sentences, using classical NLP techniques, and based on GATE [179], has been recently proposed in [180]. A methodology for using linguistic patterns in guideline formalization to aid the human modelers and reduce human modeling effort has been proposed in [181]. A method to identify activities to be performed during a treatment which are described in a guideline document appears in [129], used relations of the UMLS Semantic Network [182].

Most related to our work has been a proposal [8] for a rule-based method to identify conditional activities in guideline documents. In their experiment, with document-specific rules, they achieved a recall of 75%, a precision of 88%, and 81% F-score on the same chapter of asthma guidelines which is used in our research.

Similarly, the use of specific heuristic patterns has been shown to lead to a relatively high 85.54% accuracy in identifying recommendation statements in the hypertension guideline [4]. Ensemble learning was applied [3] to the same set of three guidelines as used in this chapter, achieving 80-84% accuracy. Part of the ensemble was a deep learning module, but it was the weakest overall performer. These results were obtained on the same guidelines as in our experiments, but at different granularity, namely on classes of combined action and conditional sentences (CCA+A). As we

show later in Section 5.6, our current methods provide about 5% – 11% improvement over these results.

Our own earlier work [2] reported lower results than [8]. The difference was due to our using of completely automated feature selection when training on an annotated corpus and not relying on manually created extraction rules. In addition, the results in [8] demonstrate recalls on specific patterns. Thus, if applied to all activities in their annotated corpus, their recall was shown to be 56%, and on our annotated corpus, it was 39%. As we show later in this chapter, in Section 5.5, we can achieve the F1 scores of 81% and higher, using completely automated methods- even purely transfer-based methods can produce a 67% F1 score and 68% recall.

5.2.3 Deep Learning Methods, Domain Adaptation and Transfer Learning

There are plenty of overviews of deep learning methods, e.g., [183, 184]. In our experiments, we use pretrained transformer models such as BERT [185] and BioBERT [186], which are relatively well-known. However, we need to discuss the concepts of *transfer learning* and *domain adaptation*, which are the focus of some experiments reported in this chapter.

We will start by observing that the two concepts overlap and are often used interchangeably. In particular, in natural language processing, as observed by [187], transfer learning is sometimes referred to as domain adaptation. Wikipedia tries to make a distinction. It explains that the basic idea of *domain adaptation* is to learn a model on a dataset, in a way that would make it applicable in other, related situations. For example, adapting spam filtering models from one set of users to another.[2] On the other hand, transfer learning "focuses on storing knowledge gained while solving one problem and applying it to a different but related problem, "for example, "knowledge gained while learning to recognize cars could apply when trying to recognize trucks".[3]

[2] `https://en.wikipedia.org/wiki/Domain_adaptation`
[3] `https://en.wikipedia.org/wiki/Transfer_learning`

For the purpose of this study, we adopt the definition from a 2015 survey of the topic [188]: "*domain adaptation* is a subcategory of transfer learning. In domain adaptation, the source and target domains all have the same feature space (but different distributions); in contrast, *transfer learning* includes cases where the target domain's feature space is different from the source feature space or spaces" (our emphasis). We can contrast this with the traditional machine learning which generally assumes that the data is in the i.i.s. form (independent and identically distributed), and that from sampled, labeled data we can train a good model for test data.

Domain adaptation and transfer learning are very active areas of research [189]. We also observe their growing importance for natural language processing [190, 191] and in clinical NLP. For example, [192] argue that "researchers in clinical NLP should treat domain adaptation, transfer learning, etc. as a first-class problem rather than a niche area", and [193] as 'worth exploring'.

In our case, Experiment 1 of Study 3 (Section 5.5.1), where we used pretrained deep learning models to find conditional sentences, is an example of domain adaptation. We also perform two experiments in transfer learning in Study 3. In Experiments 2 (Section 5.5.2), we add features from Study 1 and 2, which are not used for training of the deep learning models. In Experiment 3 (Section 5.5.3), we train on a data set combining two guideline documents and test on a different document, showing earlier (Section 5.3) that the two frequency distributions of words are very different.

5.3 The Data

In this section, we discuss the data. First, we discuss the sources of our data and its basic statistical information. Next, to show that our experiments fall under the category of transfer learning, we discuss the differences between feature distributions used in the experiment of Study 3 (Section 5.5).

5.3.1 The Dataset of Three Annotated Guidelines

We used three medical guidelines documents to create gold standard datasets and (in 2017) made them publicly available[4]. We annotated three sets of guidelines to create gold standards to measure the performance of our condition-action extracting models. The sets of guidelines are: hypertension [194], Chapter 4 of asthma [195], and rhinosinusitis [169]. Chapter 4 of the asthma guidelines was selected for comparison with prior work of Wenzina and Kaiser [8]. Each sentence was annotated by one domain expert and us, with disagreements of less than 10 percent. We have annotated the guidelines for the conditions, consequences, modifiers of conditions, and type of consequences.

Our data preparation process proceeded as follows: we started by converting the guidelines from PDF or HTML to the text format, editing sentences only to manage conversion errors, most of which were bullet points. Tables and some figures posed a problem, which were simply treated as unstructured text.

The next step, the annotation of the guidelines text, focused on determining whether there were condition statements in the candidate sentences or not. The instruction to the annotators was to try to paraphrase candidate sentences as sentences with "if condition, then consequences." If the transformed/paraphrased sentence conveyed the same meaning as the original, we considered it to be a condition-consequence sentence, and we could annotate the condition and consequence parts. For example, the sentence *"Beta-blockers, including eye drops, are contraindicated in patients with asthma"* from [195] can be paraphrased to *"If patients have asthma, then beta-blockers, including eye drops, are contraindicated"*. The paraphrased sentence conveys the same meaning. So, it became a condition-consequence sentence in our dataset. On the other hand, for example, we cannot paraphrase *"Further, the*

[4]They are at `https://data.world/hematialam/condition-action-data`. To the best of our knowledge, these three annotated documents are the only such dataset, besides the original annotations from [8]

diagnostic criteria for CKD do not consider age-related decline in kidney function as reflected in estimated GFR" from [195] to an if-then sentence; it, therefore, belongs to the category no condition (nor action).

Table 5.1: Examples of classified sentences and their classes/types.

Type	Example
Condition-Action	*Timely referral is indicated if chronic or recurrent symptoms severely affect the patient's productivity or quality of life.*
Condition-Consequence	*Most patients with uncomplicated viral URIs do not have fever.*
Action	*Adjustment is necessary for fluticasone and mometasone and may also be necessary for alternative devices.*
No condition (nor action)	*"Further, the diagnostic criteria for CKD do not consider age-related decline in kidney function as reflected in estimated GFR"*

We annotated the type of sentences based on their semantics: We classified them into four classes: condition-action (CA), condition-consequence (CC) [5], action (A), and no condition nor action (NC)[6]. Examples of sentences we are trying to find are shown in Table 5.1.

Table 5.2 contains the basic statistical information about the three guidelines. The numbers do not add up, because certain types of sentences were omitted from the annotation process (see [2]). This is, among others, due to the fact that, to be interpreted, they require a model that crosses the sentence boundaries. For example, *"The most effective therapy is intranasal steroids."*

Table 5.2: Statistical information about annotated guidelines. `Words` – the total number of words in the document. `Avg Length` – average number of words per sentence (applies to all sentences). `CA` condition-action (recommendation); `CC` condition-consequence; `A` action; `NC` no condition (nor action).

Guidelines	`Words`	`Avg Length`	`Sentences`	`CA`	`CC`	`A`	`NC`
Asthma	3621	16	224	38	7	8	117
Rhinosinusitis	19870	27	726	97	39	15	610
Hypertension	8182	34	238	63	14	1	200

[5] which includes effect, intention, and event.
[6] which includes all other sentences

In all experiments, except for Experiment 3 in Study 3, described in Section 5.5.3, the data was split 75% for training and 25% for testing. In Study 3, Experiment 3 tests domain adaptation/transfer learning, and therefore, the rhinosinusitis and hypertension guidelines were used for training and the asthma guidelines for testing.

5.3.2 The Data from the Perspective of Domain Adaptation and Transfer Learning

We plan to argue that our experiments in Section 5.5 prove the applicability of transfer learning to the detection of sentences with conditions and actions (separately or jointly). For this argument, we need to establish that the feature distribution of training data is different than the test data and that their feature sets are different.

We will be using the following feature sets Study 3 (Section 5.5) where we are experimenting with deep learning, domain adaptation, and machine learning transfer.

— Experiment 1: Identifying conditional statements using pretrained transformers models. Here, the feature set consists of the vectors from BioBERT and other transformer models. Clearly, the distribution of words in the BioBERT training data (trained on a large set of diverse biomedical texts) is different than in the selected three guidelines.

— Experiment 2: Identifying conditional statements using pretrained transformers models and features from Study 1 and Study 2. Here, the features are the sum of BioBERT vectors and the features from Study1 and (separately) the features from Study 2. And, for the reason of vocabulary distribution alone, it can be viewed as a case of domain adaptation, as defined earlier in Section 5.2.3.

— Experiment 3: transfer learning. It consists of repeating Study 2, Experiments 1 and 2 by training classifiers on two guidelines (rhinosinusitis+hypertension) and testing them on the third guideline (asthma).

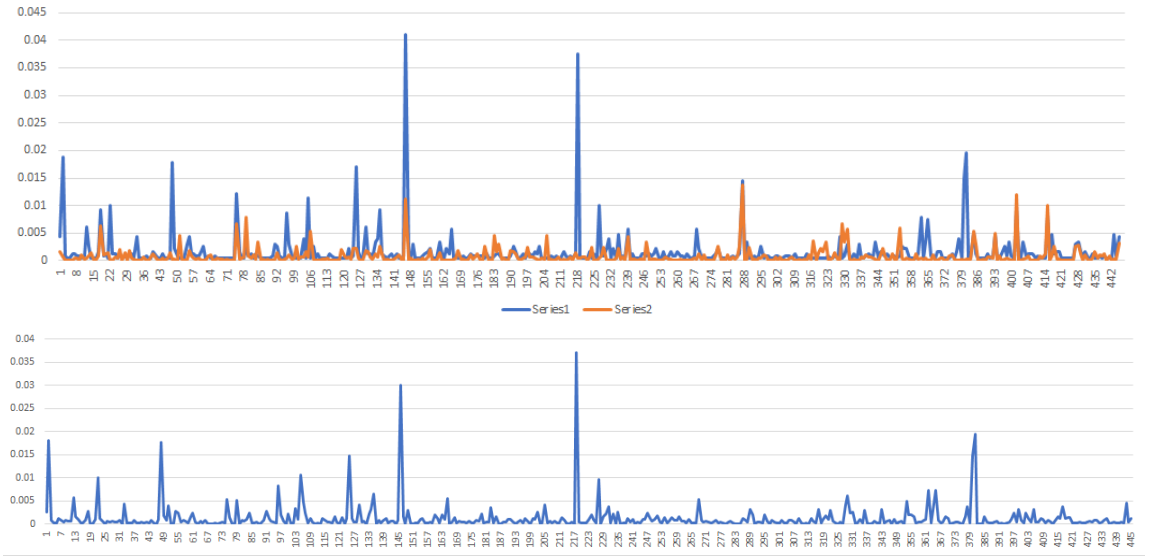


Figure 5.1: Top panel: the frequency distribution on 445 words-in-common in the training data (hypertension+rhinosinusitis; Series 1) and test data (asthma; Series 2). The bottom panel shows the large difference between the two distributions.

To establish the applicability of the concept of transfer learning to Experiment 3 (Study 3), we need to pay attention to feature distribution. In Experiment 3, the combined rhinosinusitis and hypertension guidelines has a vocabulary of 2,719 words (training set), while the asthma guidelines (test set) has a vocabulary of 661 words (test set). So, intuitively, these distributions should be different, and this difference is apparent in Figure 5.1.

This visual observation is indeed confirmed by the Kolmogorov-Smirnov test (K-S test), showing the difference with the significance level better than 0.001 on the total vocabulary. Even the distributions on the 445 words-in-common are very different according to the K-S test, with a significance level of about 0.025. The restricted distributions on the training and testing data set-based on the K-S test with a significance level of about 0.025. This is supported by another test on these two restricted distributions. The Kullback–Leibler relative entropy (K-L divergence) is high in both directions: training–testing has the value of 0.383, and for testing–training, we get 0.481. Thus, however we look at the distributions, they are very different.

5.4 Using Syntactic and Semantic Features (Studies 1 and 2)

Our baseline, Study 1, recognized condition-action statements by applying classical machine learning algorithms using a combination of domain-independent syntactic and semantic features and extending our earlier results [2]. It also extended (and improved) the results of [8] by using additional datasets, and it proved that finding sentences with conditions and actions does not have to be tailored to a specific document nor hand-coded in the form of regular expressions.

5.4.1 The Feature Set

We now briefly summarize these results, as they are given here for context and completeness. To provide the required preliminaries for Study 2 (Section 5.4.3) and Study 3 (Section 5.5), it is useful to discuss the features and methods used in [2].

The features, consisting of part-of-speech tags and syntactic patterns, were extracted from the sentences in the guidelines using the CoreNLP [196] shift-reduce constituency parser. More specifically, they consisted of POS tags and sequences POS tags (a modified algorithm, subsuming the one in [2], is shown in Section 5.4.3). For example, we have the following set of features, derived from the parse tree shown in Figure 5.2 for the sentence :*"If bp is greater than 50, follow the instruction."*

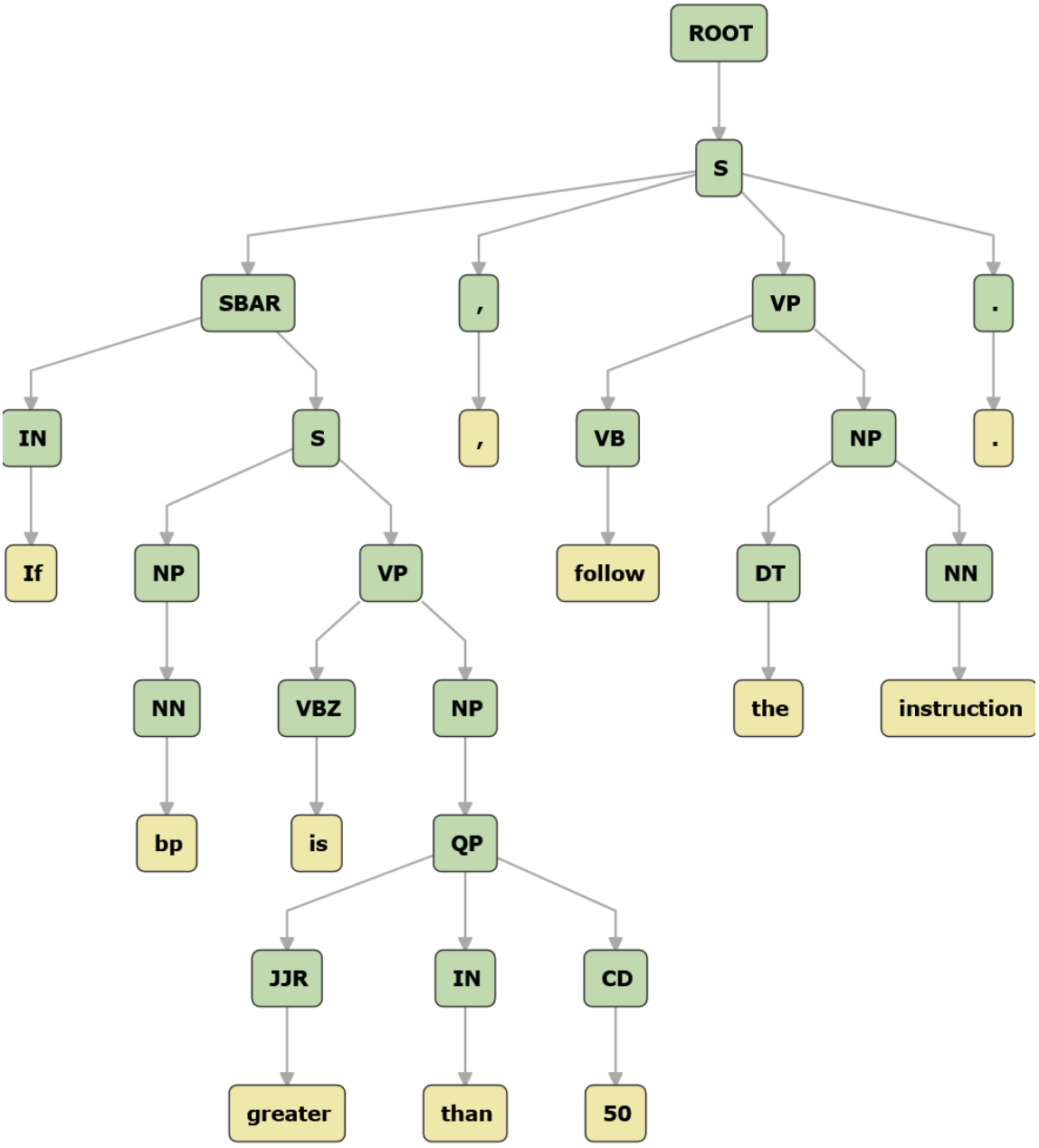


Figure 5.2: Example of the parse tree showing the part-of-speech (POS) features used in Study 1, and syntactic structures used in Study 2.

Example 5.4.1 Example list of features:

```
['IN', 'NN', 'VBZ', 'JJR', 'IN', 'CD', ',', 'VBP', 'DT', 'NN', '.',
'IN-NN', 'IN-NN-VBZ', 'NN-VBZ', 'NN-VBZ-JJR', 'VBZ-JJR', 'VBZ-JJR-IN',
'JJR-IN', 'JJR-IN-CD', 'IN-CD', 'IN-CD-,', 'CD-,', 'CD-,-VBP', ',-VBP',
',-VBP-DT', 'VBP-DT', 'VBP-DT-NN', 'DT-NN', 'DT-NN-.', 'NN-.']
```

5.4.2 Evaluation Measures and Baseline Results

We will be using the following *evaluation measures* to report results of our experiments: precision (P), recall (R), F1-measure, and accuracy (A).

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \qquad F1 = \frac{2 * P * R}{P + R} \qquad A = \frac{TP + TN}{N}$$

In the above definitions, TP is the number of items (e.g., condition expressions) that are correctly classified to a category, FP is the number of items that are misclassified, and FN is the number of items misclassified to other categories (e.g., condition as action or no-condition).

In our evaluation we used the `scikit-learn` implementation of the algorithms. The results for extracting condition-action (CA) statements are shown in Table 5.3. We achieved recall of 52%, precision of 81% for the class CA and F1-score of 63%.

Table 5.3: Classification results on annotated guidelines using only POS tags and their combinations, focusing on the detection of condition-action sentences. These baseline results are the core of Study 1.

Classifier	Class	Precision	Recall	F1-score	Accuracy
Random Forest	CA	0.95	0.42	0.58	0.86
Gradient Boosting	CA	0.81	0.52	0.63	0.84
Logistic Regression on RF&GB	CA	0.66	0.54	0.59	0.76

Table 5.4 shows combined results on the CC and CA classes, shown as CCA. Later, in Section 5.5, we show further improvements applying the three algorithms mentioned here to the vector representations produced by the deep learning models. In addition, we also show strong improvements on the combined classes CC, CA, and A, used in [4, 3].

Table 5.4: Summary of classification results on annotated guidelines, using domain independent syntactic features (Study 1), and based on [2]. The CA and CC classes are combined and shown as CCA.

Classifier	Class	Precision	Recall	F1-score	Accuracy
Random Forest	CCA	0.72	0.44	0.56	0.85
Gradient Boosting	CCA	0.62	0.44	0.52	0.83
Logistic Regression on RF&GB	CCA	0.63	0.52	0.57	0.84

5.4.3 Identifying Conditional Statements Using Semantic Types (Study 2)

Study 2 extends Study 1 by adding semantic features, that is, it is applying the domain knowledge to the process of finding condition-action sentences. For each candidate, we extract POS tags and syntactic tags at the phrase or clause level (see, Figure 5.2 and Example 5.4.2). We also use MetaMap[7]to extract UMLS semantic types of entities. We add these features to the features from Study 1.

Algorithm 4 shows the steps and the preconditions of the extraction process; the process requires the existence of specific syntactic modifiers. Here, `IN` denotes a preposition or a subordinating conjunction, `WRB` stands for a Wh-adverb, `WP` for a Wh-pronoun. and `TO` for the preposition 'to'.

[7] `https://metamap.nlm.nih.gov/`

Algorithm 4

Input: `Sentence`

Output: `Sentence type`

1: Parse the `Sentence`.
2: **if** There is a modifier tagged as `IN`, `WRB`, `WP`, or `TO` **then**
3: Extract linguistic (syntactic and semantic) features from the sentence
4: Using the extracted features, detect the sentence in one of the two categories: CA or CCA
5: **return** `Sentence type`
6: **end if**
7: **return** NC

In all experiments, after creating the lists of features for each sentence, we used the Random Forest classifier and Gradient Boosting classifier to classify sentences in our data sets. We also used the combined output probabilities from the first two classifiers to create features for a logistic regression classifier (see Tables 5.5 and 5.5).

We used the features created by the transformer models in three types of experiments:

1. We evaluate the performance of four classifiers, logistic regression, random forest, gradient boosting, and ensemble model using logistic regression on the output probabilities from random forest and gradient boosting. We only use the vectors created by the transformers as input to these four classical models.

2. We merged the features from Study 1 with the vectors from the transformer models. We evaluate the performance of the classifiers mentioned above in identifying conditional statements.

3. We used a combination of vectors and features from Study 2 to identify conditional statements using the classifiers mentioned earlier.

Example 5.4.2 shows the type of features that are added to the syntactic features shown earlier in Example 5.4.1. Note that the two sentences are different.

Example 5.4.2 Additional features for the sentence:

If bp is greater than 50, follow the instruction.

```
-- ['SBAR_IN-S-,', 'QP_JJR-IN-CD']
-- ['SBAR-VP-.', 'QP']
-- ['if-[gngm]', 'than-[fndg]']
-- ['if_[fndg]', 'if_[gngm]', 'than_[fndg]']
```

For the class CA (condition-action), we obtained a precision value of 88%, recall of 56%, and F1 score of 68% using gradient boosting and similar results using random forest and logistic regression, as shown in Table 5.5. Thus adding semantic features leads to at least 5% improvement (compared to Table 5.3).

We achieved a precision of 75%, a recall of 57 %, and an F1-score of 65% for the combined class CCA of conditional statements; see Table 5.6, showing improved results (over 10%), as compared to Table 5.4.

Table 5.5: Classification results on annotated guidelines, focusing on condition-action sentences and using all features (semantic and syntactic); Study 2.

Classifier	Class	Precision	Recall	F1-score	Accuracy
Random Forest	CA	1.0	0.52	0.68	0.86
Gradient Boosting	CA	0.88	0.56	0.68	0.85
Logistic Regression on RF&GB	CA	0.77	0.60	0.67	0.81

Table 5.6: This table shows the classification results on combined condition-consequence and condition-action classes using all features (semantic and syntactic); Study 2.

Classifier	Class	Precision	Recall	F1-score	Accuracy
Random Forest	CCA	0.81	0.48	0.60	0.87
Gradient Boosting	CCA	0.72	0.54	0.62	0.86
Logistic Regression on RF&GB	CCA	0.75	0.57	0.65	0.87

5.5 Deep Learning and Transfer Methods (Study 3)

This study uses pretrained deep neural language representation models. As before, we use these models to identify various types of conditional statements, including condition-action (CA) and condition-consequence (CC) statements.

Transfer learning through pretrained language models is a very common method in NLP. Typically, a deep learning model for target tasks are partially pretrained to create a language model and then fine-tuned on the supervised dataset. There are many well-known such language models (representations) that provide similar capabilities, but not necessarily a similar performance.

In our experiments, we used the following models: BERT [61], XLNet [197], DistilBERT [198], BioBERT `v. 1.1` [186], SciBERT `scivocab-uncased` [199], as well as BlueBERT `base-PubMed` and `base-PubMed+MIMIC-III` [200].

In contrast with the standard practice, because of the small size of our available data (discussed in Section 5.3), we could not retrain the deep learning models. Instead, we used the representations they produce to train a collection of *standard* classifiers, such as logistic regression and random forest. In other words, for any sentence of the guidelines, a deep learning model produced a vector. A learning algorithm views each dimension as a feature, and assigns importance to individual features or their collections. The result of this learning process is classifier.

In Experiment 1 of Study 3, we use pretrained deep learning models to find con-

ditional sentences, which is an example of domain adaptation. We also perform two experiments in transfer learning in Study 3. In Experiment 2, we add features from Study 1 and 2, which are not used for training of the deep learning models (although, in principle, they may be *latently* present in their neural representations [201, 202, 203, 204]).

In Experiment 3 of that study, we train on a data set combining two guideline documents and test on a different document. We showed earlier in Section 5.3 that the two frequency distributions of words are very different.

As recommended in [61], for the BERT-based models, the final hidden states corresponding to `[CLS]` token were used as the aggregate sequence representation for classification tasks. For the XLNet model, we used the final hidden states corresponding to the last token. This method provides us, for each transformer model, a sentence representation as a tensor of shape (1,768), i.e., a vector of 768 parameters/features.

5.5.1 Deep Learning. Study 3, Experiment 1

The results of two experiments using pretrained transformer models as a source of features for logistic regression are shown in Tables 5.7 and 5.8. We only show results using logistic regression, which is overall the best classifier in this context. The two experiments provide a baseline for the conditional classes CA and CCA. We can see that XLNet is the weakest overall performer and BioBERT the strongest.

Table 5.7: This table illustrates the classification results on identifying condition-action statements (CA) using transformer embeddings as features (Study 3, Experiment 1). In this table, we only report results from the logistic regression classifier, but use different vectorized representations of sentences coming from the models in the first column.

Model	Classifier	precision	recall	F1-score	Accuracy
BERT	Logistic Regression	0.78	070.	0.74	0.92
DistilBERT	Logistic Regression	0.80	0.80	0.80	0.93
XLNet	Logistic Regression	0.60	0.56	0.58	0.87
BioBERT	Logistic Regression	**0.89**	**0.82**	**0.85**	**0.95**
SciBERT	Logistic Regression	0.75	0.72	0.73	0.91
BlueBert	Logistic Regression	0.80	0.74	0.77	0.93
BlueBERTMIMIC	Logistic Regression	0.75	0.72	0.73	0.91

Table 5.8: This table illustrates the classification results on identifying conditional statements (CCA) using transformer embeddings as features (Study 3, Experiment 1). In this table, we only report results from Logistic Regression classifier.

Model	Classifier	precision	recall	F1-score	Accuracy
BERT	Logistic Regression	0.74	0.79	0.77	0.91
DistilBERT	Logistic Regression	0.80	0.78	0.79	0.92
XLNet	Logistic Regression	0.61	0.64	0.62	0.85
BioBERT	Logistic Regression	**0.85**	**0.79**	**0.82**	**0.93**
SciBERT	Logistic Regression	0.70	0.74	0.72	0.89
BlueBERT	Logistic Regression	0.81	0.72	0.76	0.91
BlueBERTMIMIC	Logistic Regression	0.78	0.72	0.75	0.91

We view this experiment as our first rudimentary experiment in domain adaptation. We note the large discrepancy in the size of the data (30K words vs. 20B words for BioBERT) and a potential for noise from the vocabulary mismatch. However, despite these potential problems, transformer-based models give a substantial boost in the performance compared with Tables 5.5 and 5.6 from Study 2 (and Study 1).

5.5.2 Deep learning. Study 3, Experiment 2

In Experiment 2, we add to transformer vectors the syntactic and semantic features from Study 1 and Study 2. As before, the data was split 75% for training and 25% for testing. We use the transformer models as the source of vectors, and each vector consists of 768 numerical features. To these vectors, we add the features from Study 1 and Study 2 and then apply logistic regression as the learning mechanism. We can view this experiment as another case of domain adaptation, as the new feature set adds UMLS concepts, and the distribution of common features is different as well. Also, BioBERT was not trained on identifying conditional sentences, so the task is new as well.

The results are better than in the earlier experiments in Study 2, reported in Tables 5.5 and 5.6. However, we can see from Tables 5.9 and 5.10 that these additional features *decrease* the performance of BioBERT (perhaps because some of them are implicitly encoded in BioBERT vectors). Interestingly enough, they improve the performance of other models, even though they are known to also encode syntactic and semantic information (as shown in previously cited [201, 202, 203, 204]).

Table 5.9: This table illustrates the classification results on identifyng condition-action statements (type CA) using different features (Study 3, Experiment 2). In this table, we only report results from the Logistic Regression classifier.

Model	Transformer vectors		Adding features from Study 1		Adding features from Study 2	
	F1	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**
BERT	0.74	0.92	0.81	0.94	0.82	0.94
DistilBERT	0.80	0.93	0.74	0.92	0.80	0.94
XLNet	0.58	0.87	0.67	0.90	0.67	0.90
BioBERT	**0.85**	**0.95**	0.79	0.93	0.81	0.94
SciBERT	0.73	0.91	0.78	0.93	0.76	0.93
BlueBERT	0.77	0.93	0.76	0.93	0.79	0.94
BlueBERTMIMIC	0.73	0.91	0.76	0.93	0.80	0.94

Table 5.10: This table illustrates the classification results on identifying conditional statements (type CCA) using different features (Study 3, Experiment 2). In this table, we only report results from the Logistic Regression classifier.

Model	Transformer vectors		Adding features from Study 1		Adding features from Study 2	
	F1	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**
BERT	0.77	0.91	0.77	0.91	0.81	0.93
DistilBERT	0.79	0.92	0.77	0.92	0.78	0.92
XLNet	0.62	0.85	0.63	0.85	0.66	0.90
BioBERT	**0.82**	**0.93**	0.80	0.92	0.80	0.93
SciBERT	0.72	0.89	0.77	0.91	0.79	0.92
BlueBERT	0.76	0.91	0.79	0.92	0.81	0.93
BlueBERTMIMIC	0.75	0.91	0.71	0.90	0.73	0.90

5.5.3 Extracting Conditional Statements using Transfer Learning

Based on the 2015 survey of the topic [188], we defined *transfer learning* as focused on cases where the target domain's feature space is different from the source feature space or spaces.

In Study 3, Experiment 3 tests the applicability of transfer learning by using the rhinosinusitis and hypertension guidelines for training and the asthma guidelines for testing. As observed earlier in Section 5.3.2, the training and testing data sets have different vocabularies, different distributions (established by Kolmogorov-Smirnov test, and by K-L divergence), and even different distributions on the common vocabulary as shown in Figure 5.1, and confirmed by the K-S test.

Table 5.11: Study 3. Experiment 3. On the class of conditional sentence (CCA), 72% F1 and 87% accuracy (A) shows applicability of machine learning transfer; it beats results of Study 2 Table 5.6 of 65%. Syntactic and semantic features from Study 2 were used in the first and third experiments.

Model	**Classifier**	**P**	**R**	**F1**	**A**
All Study 2 features	Random Forest	1.0	0.11	0.20	0.72
	Gradient Boosting	0.82	0.34	0.48	0.77
	Logistic Regression on RF&GB	0.85	0.32	0.47	0.77
BioBERT (only)	Logistic Regression	0.85	0.62	**0.72**	**0.87**
	Random Forest	0.56	0.11	0.19	0.74
	Gradient Boosting	0.58	0.47	0.52	0.77
	Logistic Regression on RF&GB	0.56	0.49	0.52	0.76
BioBERT + all Features	Logistic Regression	0.91	0.47	0.62	0.85
	Random Forest	0.75	0.07	0.12	0.75
	Gradient Boosting	0.62	0.44	0.52	0.78
	Logistic Regression on RF&GB	0.64	0.47	0.54	0.79

As we can see in Tables 5.11 and 5.12, we get results comparable to Study 2, which shows that out of the box transfer learning on unseen documents, and with completely

different distribution of features, can perform on the level of classical algorithms trained under the i.i.s. (independent and identically distributed) assumption with 75%-25% train-test split.

Table 5.12: Study 3. Experiment 3. On the class of condition-action (CA) sentences the 67% F1 score shows the applicability of transfer learning to this class, closely matching the 68% F1 score of Table 5.5. Syntactic and semantic features from Study 2 were used in first and third experiments.

Model	**Classifier**	**P**	**R**	**F1**	**A**
All Study 2 features	Random Forest	1.0	0.03	0.05	0.78
	Gradient Boosting	0.89	0.21	0.34	0.82
	Logistic Regression on RF&GB	0.89	0.21	0.34	0.82
BioBERT (only)	Logistic Regression	0.65	0.68	**0.67**	**0.85**
	Random Forest	0.50	0.29	0.37	0.78
	Gradient Boosting	0.50	0.50	0.50	0.78
	Logistic Regression on RF&GB	0.51	0.50	0.51	0.78
BioBERT+all Features	Logistic Regression	0.71	0.53	0.61	0.85
	Random Forest	0.53	0.21	0.30	0.78
	Gradient Boosting	0.61	0.53	0.56	0.82
	Logistic Regression on RF&GB	0.57	0.42	0.48	0.80

5.6 Discussion

In this section, we summarize our work from the point of view of comparison with prior art. Our results (Table 5.13 and Table 5.14) show that significant improvements of the prior art are possible using domain adaptation and transfer methods. We start with the pioneering work of Wenzina and Kaiser [8], who proposed a heuristic-based information extraction method for identifying condition-action statements.

Table 5.13: This table illustrates the improvements in classification results on identifying condition-action statements.

Experiment	Classifier	Features	F1	F1-gain	A	A-gain
Study 1	GB	POS tags	0.63	0	0.84	0
Study 2	R	Semantic + Syntactic	0.68	+5%	0.86	+2%
Study 3 Ex. 1	LR	BioBERT vectors	0.85	+17%	0.95	+9%

Table 5.14: This table illustrates the classification results on identifying conditional statements.

Experiment	Classifier	Features	F1	F1-gain	A	A-gain
Study 1	LR on R & GB	POS tags	0.57	0	0.84	0
Study 2	LR on R & GB	Semantic + Syntactic	0.65	8%	0.87	3%
Study 3 Ex. 1	LR	BioBERT vectors	0.82	17%	0.93	6%

The authors calculate a score for statements based on the appearances of trigger words ("if" and "should") and sequences of semantic types from UMLS. They achieved a recall of 75%, a precision of 88%, and 81% F1 score on the same chapter of asthma guidelines as the one used in our research. Their results only demonstrate recall on activities with specific patterns — the appearance of the trigger words "if" or/and "should."

However, if we consider all activities in their annotated corpus, the recall drops to 56%. Furthermore, we disagree with some of their annotations. We believe there are more condition-action statements in the chapter of asthma guidelines. If we apply their approach to our annotated corpus, which we used in our experiments, their recall will be 39%. In the experiments reported in this chapter, we achieved precision of 89%, recall of 82%, and an 85% F1 score on identifying condition-action statements.

Hussain et al. [4] used only the hypertension guideline annotations from our gold standard dataset to develop a heuristic model for identifying medical recommendations. In their study, they considered all condition-action, condition-consequence, and

action statements as recommendations. They achieved 85.54% accuracy in detecting recommendations using ten heuristic patterns identified manually by authors.

Hussain and Lee [3] proposed two methods to detect *recommendations* from clinical practice guidelines. They were defined in [4, 3] as the combined classes CCA+A, i.e., condition-consequence, condition-action and action.

In their experiment, first they used the TF-IDF vectors of preprocessed sentences as features for machine learning models. Second, they added aspects (UMLS concepts) of the tokens to the sentences and used the TF-IDF vectors of the modified sentences. They trained and evaluated their models on hypertension and rhinosinusitis guideline annotations from our gold standard dataset. They achieved approximately 80% accuracy for the first experiment and 84% accuracy for the second one. Although deep learning was used as a part of an ensemble learning model [3], it was the weakest overall performer.

In contrast, using the transfer method described earlier, with BioBERT vectors as features for a logistic regression classifier, we achieved, as shown in Table 5.15, a 91 % accuracy in detecting *recommendations*. They are defined in [4, 3] as consisting of the combined classes CCA+A, i.e., condition-consequence, condition-action and action. However, we should note that accuracy is perhaps a less informative measure than F1, in cases of imbalanced classes. For example, in the top row of Table 5.7, we see the 78% accuracy of random forest with an abysmal 5% F1 score.

Table 5.15: This table illustrates the classification results on identifying recommendations, defined in [3] and [4], as CCA+A . This experiment uses as features the embeddings from the transformer models, as previously shown in Study 3, Example 1.

Model & Classifier	**data**	**features**	**precision**	**recall**	**F1-score**	**Accuracy**
Logistic Regression	Hypertension	BioBERT vectors	0.94	0.75	0.83	0.91
Logistic Regression	Hypertension & rhinosinusitus	BioBERT vectors	0.77	0.65	0.71	0.87
Logistic Regression	Hypertension & rhinosinusitus	BlueBERT vectors	0.88	0.64	0.74	0.89
Logistic Regression	Hypertension & rhinosinusitus & Asthma	BioBERT vectors	0.79	0.73	0.76	0.90
Logistic Regression	Hypertension & rhinosinusitus & Asthma	SciBERT vectors	0.81	0.75	0.78	0.91
Heuristic model [4]	Hypertension	heuristic patterns	-	-	-	0.86
Ensamble learner[3]	Hypertension	TF-IDF from sentences	-	-	-	0.80
Ensamble learner [3]	Hypertension & rhinosinusitis	TF-IDF from sentences and concepts	-	-	-	0.84

5.7 Conclusions

In this chapter, we showed that modern deep learning methods, when applied to the text of clinical guidelines, yield substantial improvements in our ability to find sentences expressing the relations of condition-consequence, condition-action, and action.

As shown in a series of experiments, a combination of machine learning domain adaptation and transfer can improve the ability to automatically find conditional sentences in clinical guidelines. We showed substantial improvements over the prior art (+5% minimum, +25% maximum), and discussed several directions of extending this work, including addressing the problem of paucity of annotated data.

In summary, we presented three studies using syntactic, semantic, and deep learning methods, and performed an in-depth evaluation on a set of three annotated medical guidelines. Despite the limitation of having only a small set of annotated data, we showed the applicability of the recently developed techniques, namely neural network transformers and transfer learning to the problem of detection of conditional sentences.

CHAPTER 6: FROM KNOWLEDGE EXTRACTION TO INFORMATION RETRIEVAL AND INSIGHTS

In this chapter, we discuss the problem of finding relevant recommendations. As discussed earlier, the volume of medical texts, including clinical guidelines and case studies, is rapidly increasing. While current information retrieval tools provide good search capabilities on the level of documents, fine-grained access to specific recommendations, e.g., condition-action sentences, is difficult without additional instrumentation.

Therefore, it should be possible to use semantic indexing tools to provide such access. In the remainder of the chapter, we discuss a prototype of a semantic search engine, which is capable of:

1. Retrieving statements based on keywords, semantic concepts, and semantic types.
2. Providing different relevance metrics.
3. Providing metadata about the document, e. g. indexed MeSH terms.
4. Indexing and retrieving table data.

From the public health perspective, policy makers in health care face some challenges when dealing with various medical guidelines for the same condition. They need to find answers to these questions: Are guidelines for the same condition truly comparable? If there is a disagreement, can we identify potential reasons for it? Can we track the progression of recommendations? What is the average time of universal

adoption of innovations in medicine? With the help of text analytics capabilities, a repository of medical guidelines could help researchers investigate these questions.

Thus, we start with an overview of PubMed, discuss prior art, and proceed to the description of our fine-grained semantic search engine. We continue to discuss possible applications, including answering certain questions from the public health perspective.

6.1 PubMed – a Dominant Paradigm in Medical Search

PubMed is a search engine that is developed and maintained by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM). PubMed provides access to the MEDLINE database, which includes bibliographic information for journals covering life sciences with a concentration on biomedicine. Each article in MEDLINE is indexed with Medical Subject Headings (MeSH). MEDLINE introduced "publication type" as an indexing term to facilitate queries on trials [205]. PubMed also provides links to full-text articles in PubMed Central or other resources. PubMed Central (PMC) is a free full-text archive of articles that concentrate on biomedical and life science research. MEDLINE/PubMed data and PMC articles are available for download via the NLM website.[1] As of April 2021, there are more than 32 million citations available on PubMed (April 2021); PMC provides access to more than 6 million full-text records.

Millions of users use PubMed and PMC to retrieve medical literature each day [206].

On the one hand, retrieving relevant literature has become more challenging due to the growth rate of biomedical literature [207].

On the other hand, although PubMed and PMC are comprehensive repositories of documents, they provide the relevant documents, *not the relevant statements*. Users interested in finding specific types, e.g., recommendations or evidence, must go through additional steps to retrieve those statements.

[1] `https://www.nlm.nih.gov/`

6.2 Related Works

In this section we review some of the works related to enhancing the process of retrieving documents from PubMed. Earlier works focused on improving the relevance of the retrieved document using additional information extracted from the literature. In recent years, we have seen more interest in sentence-level information retrieval systems.

Ohta et al. [208] developed MEDIE, a search engine for MEDLINE. MEDIE was introduced with three types of search: keyword search, semantic search, and GCL search. Semantic search provides results based on the appearances of gene and disease entities. GCL search utilized parsed trees to provide results representing the relation between entities as requested in the query. Semantic MEDLINE [209] is a search engine that provides semantic relation of medical concepts as a connected graph of concepts. It uses SemRep [210] system to extract the semantic relationships between concepts that appeared in MEDLINE abstracts. A review of more traditional search engines is available in [207].

Muller et al. [211] introduced Textpresso, an ontology-based information retrieval and extraction system for biological literature. It provides a sentence-level semantic search engine. The authors labeled and indexed keywords with 33 categories from their ontology. They populated an ontology on biological concepts and relations. Their work has some limitations. First, retrieval is done at the document-level. After that, keywords, not semantic types, are highlighted in the sentences of the document. Second, semantic search can be done only on semantic type, not medical concepts.

Siadaty et al. [212] implemented a sentence-level search engine for MEDLINE. They designed a relevance metric based on the words' co-occurrence in the title of the article, a sentence of the abstract, the abstract of the article, and the indexed MeSH terms. The authors used a SQL database and SQL queries to store the data and search the repository. They evaluated their search engine by conducting two

case studies. Their result shows a significant improvement in precision for retrieving relevant documents compared to PubMed results.

LitSense [206] is another example of sentence-level retrieval systems. It uses a combination of the traditional term-weighting approach and a neural embedding model to retrieve and rank sentences from a unified repository of PubMed abstracts and PMC full-text articles. It retrieves sentences using Solr and reranks them based on their similarities with the query using a Sent2Vec model. It does not provide the capability of searching for medical types or relations.

EVIDENCEMINER [213] was developed to retrieve textual evidence from PubMed abstracts and PMC articles. It performs a keyword search and ranks the retrieved statements based on the textual evidence patterns. Since EVIDENCEMINER focuses on retrieving evidence, it has some limitations in retrieving other types of statements, i.e., recommendations. EVIDENCEMINER was used by Wang et al. [214] to provide a textual evidence retrieval system for Covid-19 literature.

None of the reviewed work addressed the information residing in tables. We believe tables are a crucial part of literature knowledge when users search for a specific statement type, i.e. evidence or recommendation.

6.3 Semantic Medical Guideline Information Retrieval System

In this section, we report our process of creating a repository of medical guidelines and the methods we use to retrieve statements from the guidelines. First, we show the data gathering process and how we processed the data. After that, we introduce a model to enable formulating data from tables. We continue by describing the Word2Vec models we created to provide alternatives to, rather than TF-IDF, relevance measures. In the end, we show how we created our search engine.

6.3.1 Data Acquisition and Preparation

To proceed with the data gathering process, we downloaded the PubMed annual baseline dataset (Januray 2020). The dataset included more than 28 million citations. Each citation in PubMed has a unique ID which is called PMID. For each entry in the PubMed dataset, we extracted the PMID and all available information from MeSH indexes, publication types, abstracts, and PMCIDs. PMCID is the unique ID for each article in the PMC.

We were interested in collecting full-text clinical practice guidelines. CPGs are indexed as "Guideline" or "Practice Guideline" in PubMed. There were 29,434 articles with "Guideline" or "Practice Guideline" as the publication type in the extracted entries. Out of these 29,434 entries, 1,901 articles had PMCID and a link to the article's PMC web page.

At first, we tried to extract the text of each article from PMC bulk articles packages, which are available from PMC FTP web service. Unlike the PubMed dataset, the PMC dataset does not use the same structure for every entry. Some article text file names were the PMCID of the article, and the journal name and date of publication were used for some other file names. This inconsistency led us to try an alternative approach. We used PMCIDs to download HTML pages from the PMC website.

We collected 1,901 HTML web pages. For extracting the text of the article, we used HTML tags to find the body text. Each section of the article is embedded inside a tag with a class attribute which starts with "tsec," and the type of that section is available inside a <h2> tag. Since we already had extracted the abstract of each article, we skipped the abstraction section on the web page. We also extracted the References section separately. All other sections were considered as the body text of the articles. Regular expressions were used for cleaning HTML tags.

After the previous steps, we noticed two issues. First, the size of the text file was notably small for some articles. We found out that some old articles are not converted

to HTML format in the PMC. Those articles' PMC web pages include only links to images or PDF files of the scanned version. We determined that we have enough materials to work on without considering these guidelines. We skipped those articles, but it is possible to use OCR tools to extract those articles' text if needed. Second, there are many tables in the guidelines we have gathered, and the extracted text does not represent table information accurately. So, we needed a model to extract and represent tables.

6.3.2 Processing Tables

We need a model which enables us to keep relations inside tables. Figure 6.1 shows an example table in a guideline. Although each cell in a table represents a data point, it needs to be combined with other elements of the table to convey the information correctly. Generally, each cell relates to the first-column cell on the same row, the header cell on the same column, and the caption. We defined a model to formalize data in a table in a textual format.

Table 4

Changes in 5-year relative survival (%) for the most common cancers, all stages, all ages, SEER 9*, 1975–2012

Cancer site	5-y relative survival (95% CI)		Change over time (95% CI)	
	1975–1977	2006–2012	Absolute, %	Proportional, %
All sites (case-mix adjusted)	50.3 (50.1 to 50.6)	66.4 (66.2 to 66.5)	16.0 (15.7 to 16.3)	31.9 (31.1 to 32.6)
Lung and bronchus	12.2 (11.8 to 12.6)	18.7 (18.4 to 19.1)	6.5 (6.0 to 7.1)	53.6 (47.5 to 59.7)
Colon and rectum	49.8 (49.1 to 50.6)	66.2 (65.7 to 66.7)	16.4 (15.5 to 17.3)	32.9 (30.7 to 35.1)
Breast (female)	74.8 (74.2 to 75.5)	90.8 (90.5 to 91.1)	16.0 (15.3 to 16.7)	21.4 (20.3 to 22.5)
Prostate	67.8 (66.7 to 68.9)	99.3 (99.1 to 99.5)	31.5 (30.4 to 32.6)	46.5 (44.2 to 48.9)
Oral cavity and pharynx	52.5 (51.1 to 54.0)	67.0 (66.1 to 67.9)	14.4 (12.7 to 16.1)	27.4 (23.5 to 31.4)
Esophagus	5.0 (4.0 to 6.2)	20.5 (19.4 to 21.7)	15.5 (13.9 to 17.1)	308.1 (217.6 to 398.6)
Stomach	15.2 (14.1 to 16.3)	31.1 (30.1 to 32.2)	15.9 (14.4 to 17.4)	104.7 (88.2 to 121.1)
Pancreas	2.5 (2.0 to 3.0)	8.5 (8.0 to 9.0)	6.0 (5.3 to 6.7)	244.7 (175.9 to

Figure 6.1: An example table from (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5409140/)

For each column, the header cell is the aggregation of cells from the top-down with a delimiter. For example, for the table in Figure 6.1, "cancer site", "5-y relative survival (95% CI) –> 1975-1977","5-y relative survival (95% CI) –> 2006-2012", "Change over time (95% CI) –> Absolute, %", and "Change over time (95% CI) –> Proportional, %" are considered as the header cells.

For each cell on the first column, the header cell is going to be added to the cell. For example, for the table in Figure 6.1, "Cancer site –> Lung and bronchus" is the first-column cell for the second row.

Finally, each cell, except the ones in the first column and header, is considered as

a data point. Each data point is stored as a line in the below format:

"Caption:() ##:## table_cell:() ##:## first_column:() ##:## first_row:() "

We applied our model to the articles we had selected. Out of 1,461 guidelines, 1,024 one had at least 1 table. We extracted 4,981 tables with 116,010 data points from the downloaded HTML webpages. For example, below is an instance data point for the table in Figure 6.1:

"Caption: Changes in 5-year relative survival (%) for the most common cancers, all stages, all ages, SEER 9*, 1975-2012 ##:## table_cell: 18.7 (18.4 to 19.1) ##:## first_column: Cancer site –> Lung and bronchus ##:## first_row: 5-y relative survival (95% CI) –> 2006-2012 "

6.3.3 Training Word Embedding Models

We believe that we can benefit from word embedding models when we want to find statements with mutual context. We used the Python implementation of Word2Vec model library (Gensim) [150] to train word embedding models. Since we focus on the biomedical domain in our research, we chose the PubMed dataset as our corpus resources. As it was mentioned before, PubMed provides access to abstracts of articles which are indexed in MEDLINE. We extracted abstracts of more than 17 million articles which were available in the PubMed dataset with their PIDs.

We considered some general preprocessing steps before we start training our models:

1. Each abstract were split into sentences.
2. Punctuations were removed.
3. Stop words were removed.
4. Characters were converted to lower cases.

We trained four different Word2Vec models. Each one is designed to be used for various studies. The first model was the baseline model. A corpus was created using

sentences from the abstracts. The corpus was used by Gensim [150] library to train the Word2Vec model. The output model had 468,869 distinctive words.

In the second model, each PMID was considered as a concept. Each PMID was paired with each word in its abstract and the pair was added to the corpus as a training sentence. The output model had 17,422,292 terms and concepts. With this approach, we created Doc2Vec alongside the Word2Vec that we had in the same vector space.

We tried to conceptualize the corpus we had in order to capture the semantic aspect of the text corpus more thoroughly. We used NLM Medical Subject Headings (MeSH) terms as our vocabulary for biomedical concepts. The MeSH vocabulary included 29,351 concepts (December 2018). Each concept has various types of representation. For example, *cancer*, *cancers*, *neoplasms*, *tumor*, and 12 other entry terms represent the MeSH term *neoplasms*. We extracted 219,499 entry terms for all MeSH terms and created a dictionary from entry terms to related MeSH term. We also considered MeSH terms with more than one word as n-grams. By replacing entry terms with MeSH terms, we conceptualized our corpus. We trained two models similar to the previous models with the new corpus. The model with PMIDs had 17,425,309 terms and concepts. The other model's vocabulary length was 471,886.

6.3.4 Implementing the Search Engine

We used a Solr instance for indexing our corpus. Each split sentence and each table data point were considered as a document. For each document, UMLS terms and semantic types were extracted using MetaMap. Each document was indexed with its UMLS terms and semantic types. A total number of 655,351 documents were indexed in this process.

We scored the similarity between sentences of each pair with the Word2Vec models we have trained. We are able to use two models (conceptualized and not conceptualized) and four methods for calculating similarities:

1. n_similarity function from Gensim at the sentence level.

2. n_similarity function from Gensim at the clause or phrase level.

3. averaging the cosine similarity between a term in the query and the most similar word in the result at the sentence level.

4. averaging the cosine similarity between a term in the query and the most similar word in the result at the clause or phrase level.

6.4 Applications

This section provides some examples of how we can benefit from the medical guideline repository and semantic search capabilities. First, we formulate some semantic queries to illustrate the advantages of semantic search. Second, we investigate the adoption/discussion of asthma medicine during a timeline.

6.4.1 Semantic Search

We created a simple UI to illustrate some of the advantages of performing the semantic search. We provide a query expansion suggestion option for users who are not familiar with UMLS concepts and MeSH terms of their query. It returns the UMLS concepts, UMLS semantic types, and MeSH terms in the user's query.

Breast Cancer Screening:

Let's assume we are interested in finding information (i.e., recommendations or evidence) for breast cancer screening from evidence-based guidelines. We need to formulate a query to retrieve the most relevant statements. Figure 6.2 shows our formulated query. We need to search for all representations of *breast cancer* in the guidelines. We can achieve this by searching for *malignant neoplasm of the breast* as a UMLS concept. After that, we are interested in evidence based guidelines. We add the *evidence based medicine* to the indexed_mesh box to filter out the guidelines not tagged as EBM in PubMed. Finally, since we are interested in screening methods,

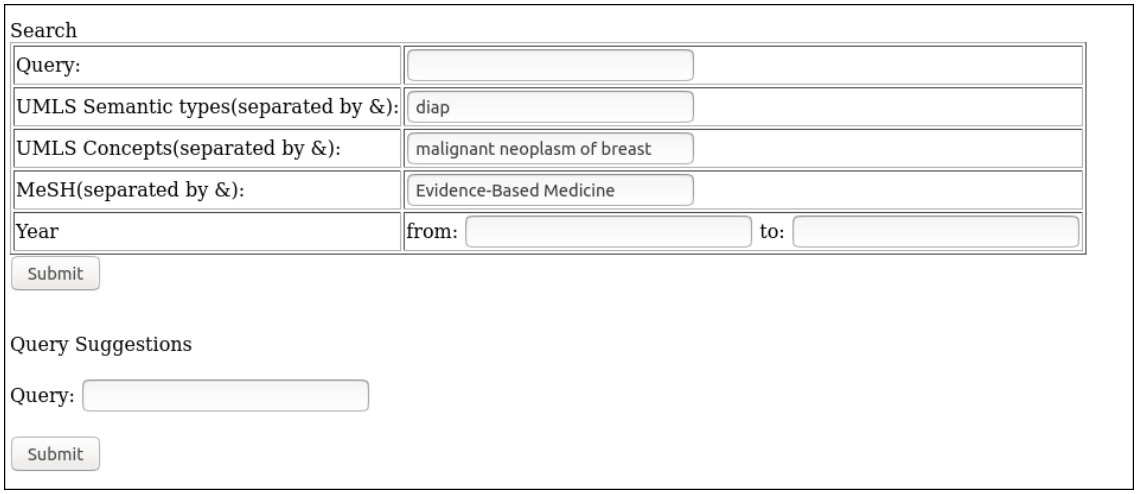


Figure 6.2: A simple search engine UI. *Query* can be used for keyword searches. UMLS semantic types and concepts are designed for semantic search. *MeSH* can filter guidelines based on their indexed MeSH type in PubMed. *Query Suggestion* can be used to translate users' queries into UMLS concepts and MeSH terms.

we use UMLS semantic type *diap* (diagnostic procedure) to find concepts related to screening.

By using the described query, we retrieved 65 statements from our corpus. Figure 6.3 shows some of the retrieved statements. Note that the first result is coming from a table.

Rank	Title	ID	Abstract	Full Text	Sentence	UMLS Concepts	MeSh Terms	Year
1	Recommendations for breast cancer surveillance for female survivors of childhood, adolescent, and young adult cancer given chest radiation: a report from the International Late Effects of Childhood Cancer Guideline Harmonization Group.	24275135_t_29	Abstract	Full Text	['table caption: Concordances and discordances among breast cancer surveillance recommendations #:#.# #:#:# first_row: Concordant/ discordant #:#:# table_cell: Discordant ']'	['Recommendation [idcn]', 'Malignant neoplasm of breast [neop]', 'Medical Surveillance [hlca]', 'Discordant [cnce]', 'Mammography [diap]', 'agreement [socb]']	['Adolescent', 'Age Factors', 'Breast Neoplasms/diagnosis', 'Breast Neoplasms/etiology', 'Breast Neoplasms/prevention & control', 'Breast Neoplasms', 'Child', 'Early Detection of Cancer/adverse effects', 'Early Detection of Cancer/methods', 'Early Detection of Cancer/standards', 'Early Detection of Cancer', 'Evidence-Based Medicine', 'Female', 'Humans', 'Interdisciplinary Communication', 'International Cooperation', 'Magnetic Resonance Imaging', 'Mammography', 'Mass Screening/adverse effects', 'Mass Screening/methods', 'Mass Screening/radiotherapy', 'Neoplasms', 'Population Surveillance/methods', 'Population Surveillance', 'Radiotherapy/adverse effects', 'Radiotherapy', 'Risk Assessment', 'Survivors', 'Time Factors', 'Young Adult']	[2013]
2	Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society.	26501536_s_244	Abstract	Full Text	['At this time, both early detection and modern therapy have important roles in the control of breast cancer.']	['Malignant neoplasm of breast [neop]', 'Social Role [socb]', 'Early Diagnosis [diap]', 'Time [tmco]', 'Therapeutic procedure [topp]', 'control substance [sbst]', 'Important [qlco]']	['Adult', 'Age Factors', 'Breast Neoplasms/diagnostic imaging', 'Breast Neoplasms/mortality', 'Breast Neoplasms', 'Early Detection of Cancer', 'Evidence-Based Medicine', 'Female', 'Health Status', 'Humans', 'Life Expectancy', 'Mammography/standards', 'Mammography', 'Middle Aged', 'Review Literature as Topic', 'Risk', 'Ultrasonography']	[2015]
3	Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society.	26501536_s_310	Abstract	Full Text	['If the woman has an average risk of developing breast cancer, the ACS encourages a discussion of screening around the age of 40 years.']	['Average [qnco]', 'around [idcn]', 'Discussion (procedure) [topp]', 'Malignant neoplasm of breast [neop]', '40% [qnco]', 'Andorra [geoa]', 'Woman [popg]', 'Disease Screening [diap]', 'Risk [idcn]', 'Age-Years [tmco]']	['Adult', 'Age Factors', 'Breast Neoplasms/diagnostic imaging', 'Breast Neoplasms/mortality', 'Breast Neoplasms', 'Early Detection of Cancer', 'Evidence-Based Medicine', 'Female', 'Health Status', 'Humans', 'Life Expectancy', 'Mammography/standards', 'Mammography', 'Middle Aged', 'Review Literature as Topic', 'Risk', 'Ultrasonography']	[2015]
4	Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society.	26501536_s_1	Abstract	Full Text	['Early detection has been shown to be associated with reduced breast cancer morbidity and mortality.']	['Associated with [qlco]', 'Togo [geoa]', 'Malignant neoplasm of breast [neop]', 'Reduced [qlco]', 'Show [anim]', 'Early Diagnosis [diap]', 'Mortality Vital Statistics [qnco]', 'Morbidity - disease rate [qnco]']	['Adult', 'Age Factors', 'Breast Neoplasms/diagnostic imaging', 'Breast Neoplasms/mortality', 'Breast Neoplasms', 'Early Detection of Cancer', 'Evidence-Based Medicine', 'Female', 'Health Status', 'Humans', 'Life Expectancy', 'Mammography/standards', 'Mammography', 'Middle Aged', 'Review Literature as Topic', 'Risk', 'Ultrasonography']	[2015]
5	Breast Cancer Screening for Women at Average Risk: 2015	[illegible]	[illegible]	Full	['Even though a substantial proportion of breast cancers are self-detected, the relative [illegible]	['Proportion [qnco]', 'Malignant neoplasm of breast [neop]', 'Unknown [qlco]', 'subscriber - self [impr]', [illegible]	['Adult', 'Age Factors', 'Breast Neoplasms/diagnostic imaging', 'Breast Neoplasms/mortality', 'Breast Neoplasms', 'Early Detection of Cancer', 'Evidence-Based Medicine', [illegible]	[illegible]

Figure 6.3: First five statements retrieved for breast cancer screening from evidence-based guidelines. The first retrieved statement is a formulated table information from table caption, first row, first column, and the cell.

Keyword Search vs Semantic Search:

We used the conditional statements from Chapter 5 as our query statements. We believe that conditional statements are the most informative pieces of each guideline.

We used the indexed corpus and performed the keyword search and semantic search on our sentences. The whole sentence was the query for keyword search, and UMLS terms with semantic types were used as queries for semantic search. For each search, we paired the top 100 search results with the query statements as the candidate pairs. We had some overlap in the search results for keyword search and semantic search.

We scored the similarity between sentences of each pair with the Word2Vec model ((not conceptualized) we have trained. We used "n_similarity function from Gensim at the sentence level" for calculating similarities.

We provide one example here to illustrate some differences between keyword search and semantic search. We used the below statement as our query: "The panel also recognizes that an SBP goal of lower than 130 mm Hg is commonly recommended for adults with diabetes and hypertension."

Our best result in the keyword search did not appear in the top 100 results in the semantic search: "For patients with diabetes mellitus who are at least 18 years of age, the panel originally appointed by the National Heart, Lung, and Blood Institute to review the evidence on treatment of hypertension recommends initiating pharmacologic treatment to lower BP at SBP of >=140 mm Hg or DBP of >=90 mm Hg and to treat to a goal SBP of <140 mm Hg and a goal DBP <90 mm Hg."

Our best result in the semantic search was the 4th retrieved document in keyword search results: "(C) Goals A goal SBP <130 mmHg is appropriate for most patients with diabetes."

Our 4th retrieved document in semantic search did not appear in keyword search results: "(C) Goals A goal systolic blood pressure <130 mmHg is appropriate for most patients with diabetes."

Drug-Drug Interaction:

In order to retrieve information about drug-drug interactions, we searched our repository for statements that mentioned *interaction* and have at least a concept tagged as *phsu* (Pharmacologic Substance). We retrieved 115 statements. After that, we excluded negative sentences, e.g., 'Riociguat has no pharmacodynamic interaction with warfarin.' We collected 95 sentences with positive mentions of *interaction* and pharmacologic substances, e.g., 'Due to effects on protein binding, there is a potential interaction with warfarin requiring careful monitoring.'

6.4.2 Case Study: Asthma Medicine

Since we have access to each guidelines' published year, we can analyze temporal changes between guidelines during the time. We selected the changes of recommended medications in asthma guidelines to investigate our system's capabilities.

We retrieved 1,400 statements by searching for concepts tagged as "phsu" (Pharmacologic Substance) in guidelines indexed as *Asthma* guidelines. Figure. 6.4 shows the word cloud of these concepts in our retrieved statements.

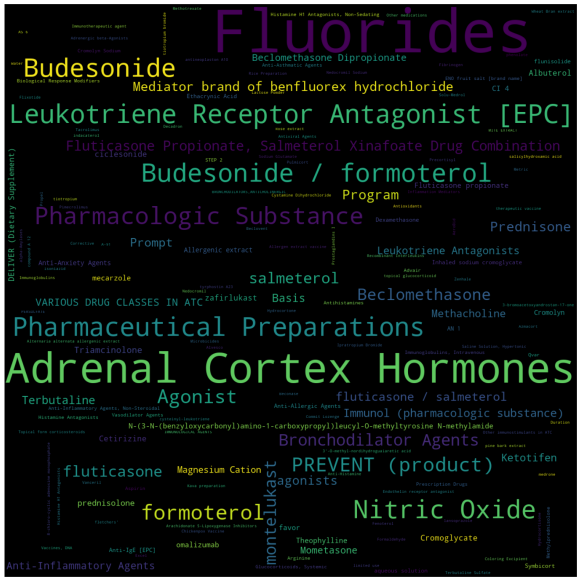


Figure 6.4: Word cloud of *Pharmacologic Substances* based on their frequencies in guidelines tagges as *Asthma* guidelines.

Table 6.1 shows some of the drugs approved by FDA after 2004 and the time they were introduced. Our extracted list of drugs from guidelines has some common items with the drugs as shown in Table 6.1,e.g., Budesonide/formoterol and Fluticasone.

Table 6.1: Asthma Medicine Products Approved by the U.S. FDA [5]

Drug type	**Year**	**product name**	**Drug(s)**
Hydrofluoroalkane Metered Dose Inhaler	2005	Xopenex HFA	Levalbuterol tartrate
	2006	Aerospan	Flunisolide
	2006	Advair HFA	Fluticasone propionate/salmeterol
	2006	Flovent HFA	Fluticasone propionate
	2006	Symbicort	Budesonide/formoterol
	2008	Alvesco	Ciclesonide
	2010	Dulera	Mometasone/formoterol fumarate
	2014	Asmanex HFA	Mometasone furoate
Dry Powder Inhalers	2005	Asmanex Twisthaler	Mometasone furoate
	2006	Exubera	Recombinant human insulin
	2006	Pulmicort Flexhaler	Budesonide
	2006	Foradil Certihaler	Formoterol fumarate
	2010	Aridol	Mannitol
	2011	Arcapta Neohaler	Indacaterol maleate
	2013	Tobi Podhaler	Tobramycin inhalation powder
	2013	Breo Ellipta	Fluticasone furoate/vilanterol
	2014	Incruise Ellipta	Umeclidinium
	2014	Afrezza	Human recombinant insulin

In order to visualize the temporal changes of appearances of extracted drugs in asthma guidelines, we created a heat map of frequencies of each concept in each guideline. Figure 6.5 shows the created heat map.

Figure 6.5: We show a heat map of the frequencies of the drugs in each guideline. Each guideline is represented by "Year::PMID". For all frequencies higher than 10, we used dark green in order to emphasize the appearances of concepts better.

From the public health perspective, as users with no medical background, by looking at the Figure 6.5 we can extract some knowledge about the discussion of drugs in the guidelines such as:

1. Substances, such as Agonist and Budesonide/formoterol, were not mentioned in the guidelines until 2009.

2. Discussion over formoterol was increased from 2005 to 2012.

3. Interest in montelukast was decreased since 2005.

4. We can observe that the focus of the substances in the item labeled as "2013 :: 23457669" is clearly different than the other guidelines. When we look at the index MeSH terms for the guidelines, we can see "2013 :: 23457669" is the only guideline focused on diagnosis exclusively.

We can expand this experiment by searching for co-occurrences of a drug and a symptom/disease in order to track the discussion of treatment for symptom/disease by a specific drug. We retrieved 574 statements with co-occurrences of a concept tagged as "phsu" and a concept tagged as "dsyn"/"sosy". Figure. 6.6 illustrate the created heat map from frequencies of pairs of drug-disease or drug-symptom in our guidelines.

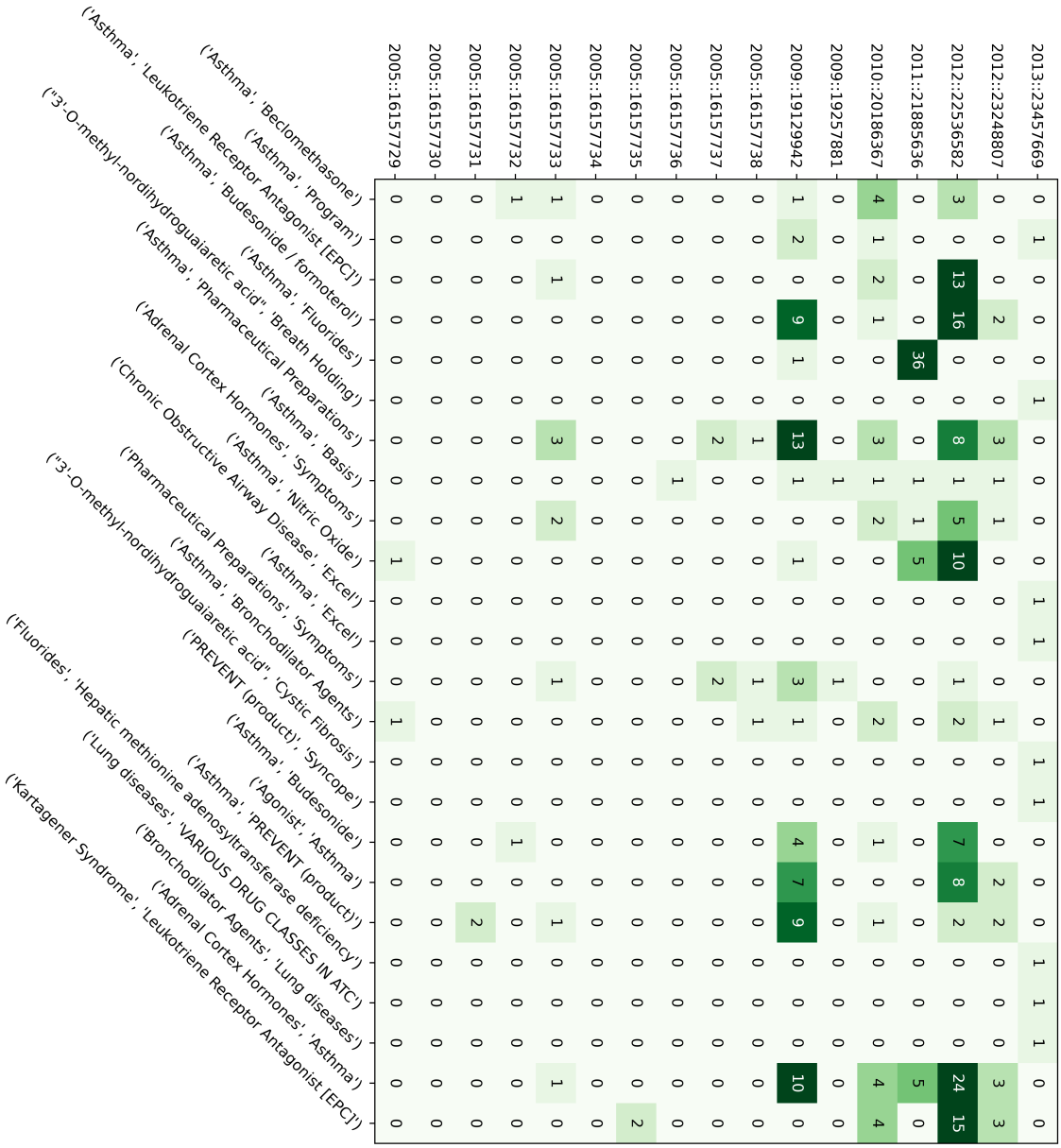


Figure 6.6: A heatmap of frequencies of a *Pharmacologic Substances* and a disease or a symptom in *asthma* guidelines in our repository. Values more than 10 are being shown in dark green. Each guideline is represented by "Year::PMID".

Similar to Figure 6.5, Figure 6.6 can help us understand the change of focus in asthma medicine better. For example, although *fluorides* were mentioned in guidelines from 2005, but it was not discussed with *asthma* until 2009.

6.5 Discussion

Since we do not have a gold standard at this point, we cannot evaluate the performance of the semantic search and the relevance measures numerically at this point. But we can make the below observations by reviewing the results of a sample query list:

- Both keyword and semantic search retrieve short sentences that have a common word or concept in the query with a high relevance metric. This issue came from the fact that Solr is using TF-IDF scores for ranking documents.

- When two statements have a similar length, cosine similarity at the sentence level works fine. But if the length of the query and result are not similar, sentence level similarity calculation does not perform well.

- Semantic search is able to retrieve results with different representation terms of the same concept.

- Generally, each metric performs well on some examples.

From the public health perspective, we showed that we can provide raw materials to answer some of the policy makers' questions about the adaptation of the drugs. But we should mention some challenges we faced in that experiment:

- MetMap is a useful tool in extracting concepts, but it is not perfect. For example, in Figure 6.5, we can see that *EXCEL* is labeled as a drug, but it is the name of a trial in the document.

- Physicians are interested in knowing the background information, such as the reliability of the guideline or the authors' background. At this point, our system doesn't provide these types of information.

- To perform a temporal analysis on medicine, we should add more medical guidelines to our repository. For example, as shown in Figure 6.5, we miss guidelines from 2005 to 2009.

CHAPTER 7: IMPROVING BLOOD TRANSFUSION MEDICAL RECORDS USING TEXT ANALYSIS

7.1 Introduction

The electronic health record (EHR) is the primary source of patient medical information. While efficiencies of data control have been created with templates and structured data, much of the EHR still consists of free-text information. In the clinical laboratory, it is required to have a reason why a test or treatment is being provided for billing and quality assurance reasons. Most lab orders are structured with standardized diagnosis coding applied for order justification, such as CPT or ICD-10 codes.

There are four common blood components that are used in patient care. They are derived from separating whole blood from blood donors to maximize the storage potential for each component. Red blood cells (RBCs) are the most commonly transfused blood component, used to replace patient RBCs that may be lost due to bleeding or not manufactured by the patient (often due to chemotherapy treatments). Platelets (PLTs) are small cell fragments important in blood clotting that are necessary to support cancer treatment and cardiac surgery. Plasma is the liquid portion of blood that contains all of the coagulation factors needed to make a strong blood clot, and cryoprecipitate (CRYO) is a derivative of plasma that is concentrated in key coagulation factors, both used to treat bleeding patients. One laboratory order that still involves significant clinical variability in decision-making is blood transfusion. Blood transfusion has been noted as one of the most overused treatments in healthcare [215].

To address the overuse of transfusion, much work has been done providing more

specific guidance on when a blood transfusion is appropriate [216, 217, 218]. This has resulted in many institutions adopting a clinical indication field in the computerized provider order entry system (CPOE) when ordering a blood transfusion. By including standardized indications as choices for this field, it applies a clinical decision support (CDS) tool. The standardized options (i.e. *predefined reasons*) inform the ordering provider "here are the reasons we approve for blood transfusion." Table 7.1 shows the the products and their standardized options.

However, the evidence-based guidelines do not cover all clinical situations, and there are always new or uncommon reasons that would be burdensome to include as standardized options. The choice of "other" with a free-text field for explanation has been included for these situations. The explanation could then be reviewed by a physician for clinical judgment.

Table 7.1: Blood Transfusion Indications from the CPOE. The "Other – ..." fields are the source of free-text data used in our analysis.

Product	Standardized Reason
CRYO	Bleeding with Fibrinogen < 200 mg/dL
	Uncontrolled hemorrhage/massive transfusion
	Bleeding with Uremia
	Factor deficiency, approved by coagulation service attending (970-2414)
	Other - enter reason and attending MD in comments
RBCs	Hgb < 7.0 g/dL
	Hgb < 8.0 g/dL and Coronary Artery Disease
	Hgb < 8.0 g/dL in outpatient/oncology patient
	Hgb < 10.0 g/dL and Symptomatic Anemia – please describe (including Attending MD name)
	Uncontrolled hemorrhage/massive transfusion
	Erythrocytapheresis
	ECMO
	Intra-uterine Transfusion (IUT)
	Perioperative/procedural bleeding
	Perioperative/procedural bleeding
	Other - enter reason and attending MD in comments
Plasma	PT INR > 2.0 on Warfarin, with bleeding (give 10-20 mL/kg)
	PT INR > 1.5 x normal AND uncontrolled bleeding (describe) (give 10-20 mL/kg)
	PT INR > 1.5 x normal AND prior to non-elective invasive procedure-describe (give 10-20 mL/kg)
	PT INR > 1.3 x normal with CNS/ocular trauma, bleed or surgery
	Uncontrolled hemorrhage/massive transfusion
	Plasma for therapeutic plasma exchange (call Blood Bank to confirm)
	Other - enter reason and attending MD in comments
PLT	PLT < = 10K
	PLT < = 50K and bleeding – please describe in comments
	PLT < 50K and pending invasive procedure or surgery
	PLT < 100K and neurosurgical procedure
	Uncontrolled hemorrhage/massive transfusion
	Other - enter reason and attending MD in comments

Problem statement:

On review over 8,000 orders in a 12 month period (20% of blood orders) had an indication of "Other" with free-text attached. The manual review suggested that

there were repeating themes in the free-text. The question arose as to whether these "Other" reasons could be automatically analyzed, grouped or mapped to preexisting EHR options. If so, the new frequently used reasons would be added to the standard drop down options. In addition, the results of this analysis would be used to educate providers to choose the standardized answers.

We seek to analyze the text to quantify commonly appearing free-text answers. Our idea is to extract conditions from each reason and compare bags of conditions to identify similarities between reasons. For example, we have 'PLT $<$ 100K and neurosurgical procedure' as a predefined reason in our dataset. It includes two conditions: 'PLT $<$100' and 'neurosurgical procedure'. In our dataset, we also have 'Neurosurgery 8/2 plt goal $>$ 100 for 2 weeks' as a free-text reason. These two reasons provide identical conditions with different language.

We need to extract conditions from reasons and normalize them to enable comparisons and mappings. For example, 'plt goal $>$100' should be normalized into 'PLT $<$ 100' to be matched with the predefined reason's condition.

In the first example, 'PLT $<$100K' is a *numerical condition* and 'neurosurgical procedure' is a *conceptual condition*.

We define the former, *a numerical condition*, as a chunk of text which includes 3 components: A numerical value, a comparative sign, and a medical concept. The latter, *a conceptual condition*, is any combination of medical concepts, such as diseases or symptoms, which describe the patient's condition.

7.2 Overview of Prior Art

Natural processing methods have been applied to medical texts for about half a century. Text analysis of medical records is already mentioned in a 1975 article by N.Sager [172, 173], and included extracting information to populate relational

databases. Since then, many new techniques have been developed, and applied to medical texts.

The state-of-the-art of text analysis in the context of electronic medical records is reviewed in [219] and contains an observation that "statistical analyses or machine learning, followed by NLP techniques, are gaining popularity over the years in comparison with rule-based systems." Since the early 2010s, which this observation correctly characterizes, this trend accelerated and has been amplified by the use of deep learning [125].

For improved accuracy of classification and data extraction, the statistical techniques often rely on preprocessing aimed at extraction of important entities. For example, MetaMap[1], or GATE-based TextHunter [220] can be used to extract concepts to be used in a classification task, as in [221] for pneumonia identification from narrative reports, or in [222] in the context extracting symptoms of mental illness from clinical text.

Rotmensch et al. [223] proposed an automated process to create a knowledge graph linking diseases and symptoms using probabilistic models on electronic medical records (EMRs). They used string-matching to extract concepts (in different forms such as acronyms and ICD-9 codes). In the next step, they used statistical models to relate diseases and symptoms. Finally, they translated the statistical models into knowledge graphs. The authors evaluated their knowledge graph by comparing it against a subset of Google health knowledge graph and a clinical evaluation from domain experts. They achieved a precision of 0.23 for a recall of 0.5 against GHKG. In the clinical evaluation, they achieved a precision of 0.87 for a recall of 0.5.

Chen et al. [224] proposed a methodology on analyzing a health knowledge graph, proposed by Rotmensch et al. [223], relating diseases and symptoms extracted from EHRs. They evaluated the knowledge graph by computing the F1 measure for each

[1] `https://metamap.nlm.nih.gov/`

disease. They found out that like sample size and the number of co-occurring diseases in patients.

Ma et al. [225] introduced a framework consisting of rule-based and machine learning models to capture disease as a causal chain of abnormal states from electronic health records (EHRs). They utilized text mining abilities, i.e., Word2Vec and regular expression, to detect and expand abnormal states. The authors reported a significant positive impact of each text mining method on retrieving abnormal states.

Bjarnadottir et al. [226] used text mining to find related content related to fall risk and prevention in medical notes. They extracted unigrams, bigrams, and trigrams from a pre-processed dataset on nursing notes. They used NOTEEVENT nursing notes from Medical Information Mart for Intensive Care (MIMIC) III open-source dataset [227] to evaluate their work. They report the frequencies of n-grams related to risk factors, events, and prevention detected in a notes lexicon of words and terms that are clinically or theoretically related to patient falls.

Cobb et al. [228] used a bag of words and a bag of concepts (UMLS concepts) as features to predict the outcomes during patient care. They achieved an F1 score of 0.5 on predicting the outcomes utilizing SVM and Nearest Centroid classifiers.

Transfusion studies that have looked at manual text entry for order indications have used manual review and classification of the text [229]. A Canadian group developed a computerized audit tool that still required manual extraction of key elements from the order indication then applied an algorithm to judge appropriateness [230]. No applications of computerized text analysis can be found in this setting, and this article is the first study of this kind.

7.3 Classification objectives and Process

The first objective of our text analysis process is to map the free-text reasons to *predefined reasons* (i.e. reasons which show up as options in EHR). The second objective is to detect *repetitive reasons*, that is frequently appearing free-text reasons,

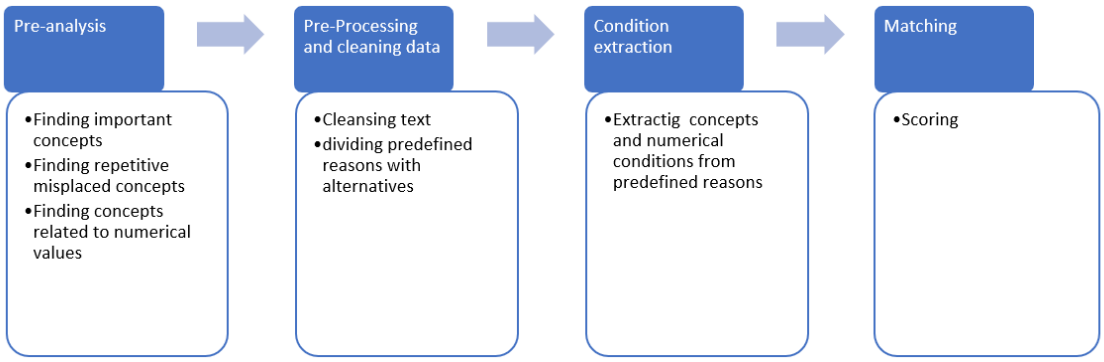


Figure 7.1: A proposed architecture for mapping reasons to standardized options (predefined reasons).

to be possibly added to the list of standardized options.

Some of the free-text order reasons echo predefined transfusion reasons with a slight change of language or format. For example, 'plt goal > 100' is a representation of 'PLT < 100'. Such reasons include conditional statements of conceptual entities (e.g., quantitative concepts), procedures, diseases, and symptoms. Therefore, by normalizing the language used in the free-text reasons, we should be able to match those reasons with their counterparts in the predefined categories.

For finding repetitive reasons, we performed statistical analysis of the appearance of concepts as bags of concepts and bags of numerical conditions.

7.3.1 Mapping to Predefined Reasons

Our idea is to transform reasons into normalized (conceptualized) bags of conditions and compare them to bags of conditions from predefined reasons. We divided conditions into two forms: conceptual conditions (e.g., diseases) and numerical conditions.

The process of mapping free-text reasons to predefined reasons includes several steps shown in Figure 7.1. We used MetaMap to extract UMLS concepts from each statement. Using UMLS concepts enabled us to match acronyms (e.g., BMT and 'bone marrow transplant') and synonyms (e.g., 'hemorrhage' and 'bleeding'). We

faced two problems using MetaMap for extracting concepts. First, not all concepts mapped by MetaMap are important for us. Second, MetaMap has some limitations in resolving word-sense disambiguation.

To resolve these limitations, we start with the pre-analysis step in our experiment. In this step, we selected a list of UMLS informative, in our opinion, concepts to be used in the next steps: 'Deficiency Aspects', 'Plasma Exchange', 'Coronary Arteriosclerosis', 'Open Approach', 'Extracorporeal Membrane Oxygenation', 'On Warfarin', 'Neurosurgical Procedures', 'Elective Procedure', 'Elective Surgical Procedures', 'Perioperative Procedures', 'International Normalized Ratio', 'Operative Surgical Procedures', 'Fibrinogen Concentrate (human)', 'Eye Injuries', 'Central Nervous System', 'Anemia', 'Blood Transfusion, Intrauterine', 'Uremia', 'Massive Blood Transfusion', 'Non-invasive', 'Injury of Central Nervous System', 'Factor', 'Hemorrhage', 'Uncontrolled', 'Symptomatic', 'Invasive', 'Ocular (qualifier)', and 'Erythrocytapheresis'.

To resolve the word-sense disambiguation problem, we manually checked concepts from predefined reasons and repetitive concepts extracted from free-text reasons. For example, we found that both 'PT' (Prothrombin time) and 'Pt' (Patient) were mapped to 'Physical Therapy' by MetaMap. We made a list of misplaced concepts to be used in the next steps.

In the next part of the pre-analysis step, we address the limitations of numerical conditions. First, we need to find the concepts that appear in numerical conditions. In our list of predefined reasons, we have four concepts that appeared in the numerical conditions: 'Fibrinogen', 'Hgb', 'PT INR', and 'PLT'. After we have our target list of concepts for numerical conditions, we should address the below limitations for extracting numerical conditions:

- Missing comparative signs for numbers. For example, 'PLT 12' was typed by physicians instead of 'PLT = 12'.

- For 'PLT' product, since the product and concept are the same, some orders

are missing 'PLT' as a concept in the reason and they start with comparative signs or a number.

- Different representations of concepts. For example, *'PLTs* and 'Platelet' are other terms for 'PLT' in our dataset.

- Orders which set a threshold of concepts.

To resolve the first limitation, we considered cardinal numbers and the nouns next to them in the reasons with no comparative signs as numerical condition candidates. If a reason starts with a comparative sign or a number under the 'PLATELET' product, we added 'PLT' to the start of the reason to address the second limitation.

To find the terms which represent our list of concepts for numerical conditions, we extracted all numerical condition candidates and manually check the list of concepts to find the other terms representing our list of numerical concepts.

We noticed some reasons asked for some concepts to be set at some threshold (e.g., 'BMT pt requires plt $>$ 50'). To handle the extraction of numerical conditions for these types of reasons, we reversed the comparative sign of the numerical condition if a trigger term, such as keep, maintain, or requires, appeared before the condition.

We pre-processed statements by removing stop words, replacing comparative signs with their meaning, removing punctuation, replacing concepts with 'preferred_name' (provided by MetaMap), and replacing capital letters with lower cases. We also divided predefined reasons, which include alternatives. For example, 'Uncontrolled hemorrhage/massive transfusion' was divided into 'Uncontrolled hemorrhage' and 'massive transfusion.'

In the next step, we extracted conditions as numerical conditions and conceptual conditions from each statement. For extracting numerical conditions, we defined 3 parts for each condition: concept, comparative operator, and value. We used CoreNLP to parse each statement and find these parts. Comparative operators were

found under ‘JJD’ or ‘RBR’ tags. Values were tagged as ‘CD’. Concepts were extracted as noun phrases placed on the left side of the comparative operator. Each numerical condition was stored as a tuple for each statement. For each reason, UMLS concepts with semantic types included in the list created in the pre-analysis step were extracted and stored as a bag of concepts for that statement. For example, (’PLT’, ’Less than’, ’100’) was stores as the numerical condition for ‘plt goal >100’.

In the last step, we scored the similarities between free-text reasons and each predefined reason. For the bag of concepts, we defined the similarity score between a free-text reason and a predefined reason as the number of mutual conditions over the number of concepts from that reason. For example, ‘Hgb 7.0 and actively bleeding’ and ‘Hgb $<$ 7.0 g/dL’ have one mutual concept (‘Hgb’). The *similarity score* for this pair will be 0.33 since the reason has 3 concepts (‘Hgb’, ‘actively’, and ‘bleeding’).

For the numerical conditions, we used the numerical subsumption to test if the numerical condition from a free-text reason fits the numerical condition from predefined reasons. For example, both ‘ ‘PLT = 30’ ’and ‘ ‘PLT $\leq$ 40’ ’ fit “PLT $\leq$ 50”, i.e. are subsumed by the last condition.

7.3.2 Detecting Repetitive Reasons

To provide candidate repetitive reasons, we followed similar steps, with some changes, to the matching process. Figure 7.2 shows these steps.

To find our relevant list of concepts, we start with the pre-analysis step in this experiment. In this step, instead of using a list of concepts, we selected a list of UMLS informative, in our opinion, semantic types to be used in the next steps. The list includes below semantic groups and semantic types:

- Conceptual Entity : [qnco], [qlco], [tmco], [ftcn], [bdsy]
- Finding: [fndg], [sosy], [lbtr]
- Health Care Activity : [hlca], [lbpr], [topp], [diap]

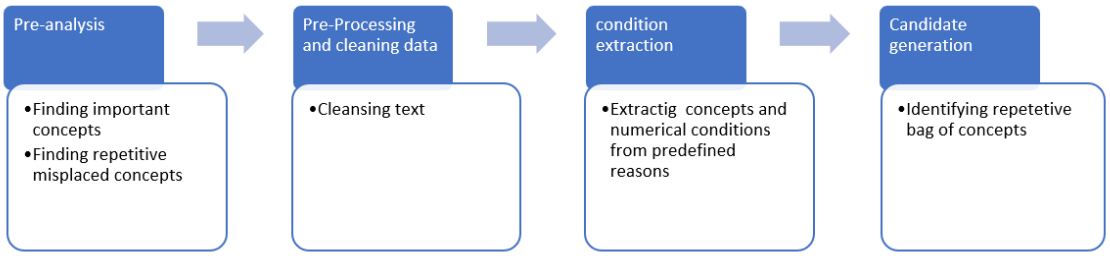


Figure 7.2: A proposed architecture to generate candidates for new standardized options.

- Pathologic Function : [dsyn], [neop], [patf]
- Injury or Poisoning : [inpo]
- Pharmacologic Substance and body Substance: [phsu], [bdsu]

After cleaning reasons by removing stop words, punctuations, and symbols, we extract concepts using MetaMap. We also extract numerical conditions by the method we used in 7.3.1.

At the last step, instead of mapping reasons, we provide statistical information for the bag of concepts to be considered for candidate generation.

7.4 Results

In this section, we review the results of the two experiments explained in the previous sections.

7.4.1 Mapping to a Predefined Reason

Our data set includes 3,908 reasons tagged as *free-text* or *other* reasons. We were able to extract at least one numerical condition from 1,476 reasons. We mapped 1,105 of those conditions to their counterparts for predefined reasons. After adding the conceptual score to the numerical score, we had 79 perfect matches (e. g. "Neurosurgery 8/2 plt goal $>$ 100 for 2 weeks" and "PLT $<$ 100K and neurosurgical procedure") and

1,206 partial matches(e. g. "fibrinogen $<$ 100" and "Bleeding with Fibrinogen $<$ 200 mg/dL"). We could not find a counterpart for 371 numerical conditions because either the numerical concept was not in our list or the condition did not fit any numerical condition.

We detected 2,432 reasons without any numerical condition. We divided the results for these reasons into four categories:

- perfect match: we were able to match 202 reasons.
- conceptual match with a predefined reason with a numerical condition: we got conceptual score = 1 for 36 reasons. Some of these reasons include a numerical condition, but our algorithm could not formulate it (e.g., "Hgb-7.1 Dr. Phys33")
- Partial match (conceptual score $<$ 1): 634 reasons were partially matched with a predefined reason based on some mutual concepts. For example, 'symptomatic anemia; transfusion dependent' fits some part of "Hgb $<$ 10.0 g/dL and Symptomatic Anemia – please describe including Attending MD name."
- No match: For 1,560 reasons, we could not find any mutual concept with the predefined reasons.

We selected a random subset of 400 reasons(about 10% of the dataset) to evaluate numerical condition extraction and our mapping process.

We extracted numerical conditions from 149 reasons. two extracted conditions were selected incorrectly. Out of 251 reasons without numerical conditions, 231 reasons had no numerical conditions. 20 reasons had numerical conditions, and we could not extract them because they were mentioned in an unstructured format. For example, we could not detect the numerical condition in "Hgb steadily downtrending now 7.3 pt very tachycardic strong suspicion patient is bleeding". We achieved 95% accuracy in detecting numerical conditions.

Out of the 149 reasons that we extracted numerical conditions from, we could match 144 of them with their counterparts in the predefined reasons correctly. This means that we achieved 97% accuracy in comparing numerical conditions.

All 400 reasons were examined by a domain expert to evaluate if the mapped reason is the preferable choice for a reason or not. Table 7.2 shows the evaluation results of reasons in different categories. We achieved 83% accuracy in mapping free-text reasons to a predefined reason.

Table 7.2: Classification results on free-text reasons. *Conceptual Match*: How the concepts in a reason match with the concepts from the assigned predifned reason. *Numerical Match*: whether a reason and its assigned predefined reason match or not. *Has a Numerical Condition*: Whether a reason has a numerical condition or not. *Frequency*: reports of the reasons classified based on *Conceptual Match* and *Has a Numerical Condition*. *Mapped Correctly*: The number of reasons from that class which our domain expert agrees that the mapped reason is the preferable choice for a reason.

Conceptual Match	**Numerically Matched**	**Has a Numerical Condition**	**Frequency**	**Mapped Correctly**
Yes	Yes	Yes/No	39	39
Yes	No	Yes/No	5	4
Partially	Yes	Yes/No	129	119
Partially	No	Yes/No	49	27
No	No	Yes/No	178	143

7.4.2 Repetitive Reasons

We provide the statistics of repetitive concepts in four categories: Numerical concepts, pairs of a numerical concept and one medical concept, individual concepts, and pairs of concepts. Table 7.3 shows the frequencies of detected repetitive concepts in different product orders. For *concepts*, we present any concept with frequency greater than 20. We set the threshold 10 for *Pairs of concepts* and *Pairs of a numerical concept and one medical concept*. We listed any *numerical condition* with frequency more than 5.

Table 7.3: Frequencies of the repetitive concepts for different products in the blood management system.

Repetitive Concept Category	Product	Repetitive Concept	Frequency
Numerical concepts	PACKED RBC	Age	18
	PACKED RBC	Neonatal transfusion age	6
Pairs of a numerical concept and one medical concept	PACKED RBC	Hgb & ABMT	72
	PLATELETS	Platelets & BMT	41
Concepts	PLATELETS	Aspirin	86
	PLASMA (FFP)	ECMO	80
	PLATELETS	Bone marrow transplantation	61
	PACKED RBC	Gastrointestinal	41
	PACKED RBC	Hypotension	40
	PACKED RBC	Autologous bone marrow transplant	38
	PLATELETS	ECMO	34
	PLATELETS	Thrombocytopenia	26
	PLATELETS	Plavix	24
	PLATELETS	Defibrotide	21
Pairs of concepts	PLATELETS	aspirin & Neurosurgical procedures	36
	PLASMA (FFP)	ECMO & Vascular cannula removal	14

7.5 Discussion and Conclusion

Manual analysis of free-text data is labor-intensive and inefficient. While individual review aids in specific order inquiry and allows for direct education, it will not identify patterns of orders and shifts in practice. The ability to automate this analysis and review a large set of data allows for ordering patterns that may require addressing. Once trained, a number of concepts appeared that were reasonable indications for transfusion that are not included in formal guidelines, which also meant they were not in the formal list of transfusion reasons. The analysis also identified concepts that may have been included in the indication for one blood product type but not in another; ECMO (extracorporeal membranous oxygenation) was codified for RBC orders, but not platelets – both products are often indicated to support the ECMO procedure.

By quickly classifying order indications into concept groups, reasons that are out-

side the norm can be more easily identified. Addressing and eliminating these indications quickly reduces unnecessary transfusions, which in turn, reduces transfusion risk to patients and eliminates the added cost of transfusion.

In this experiment, we used MetaMap to detect medical concepts. As mentioned earlier, MetaMap is not perfect in resolving word-sense disambiguation. We chose to manually resolve misannotations in the post-processing step since we needed to cover a limited number of concepts, and we knew all forms of representation of them.

We achieved 95% accuracy in extracting numerical conditions using the syntactic structure of the reasons. Using the proposed method, we were able to map 83% of the reasons in our test dataset to the preferable standardized option.

This chapter showed that using text analysis techniques can help patient management systems detect records that do not match their health system goals. This will help physicians prevent overtreatment. Also, the textual analysis we presented in this chapter helped to modify standardized options to include repetitive justifiable reasons.

CHAPTER 8: SUMMARY OF DISSERTATION, OPEN PROBLEMS, AND FUTURE DIRECTIONS

In this dissertation, we addressed some problems resulting from the enormous number of guidelines available in the medical domain. We utilized NLP capabilities to analyze medical guidelines and differences between guidelines addressing the same topic, namely, disagreements between guidelines, conceptual distances between guidelines, and identifying informative segments of the guidelines. In addition, we introduced a sentence-level information retrieval system for a corpus of medical guidelines. Finally, we showed that text analysis capabilities can be used in other medical texts to improve patient care and decision-making.

Some of the contributions and possible future directions of this dissertation are as follows:

- A novel formal analysis of types of contradictions in texts. Namely, we introduce and formally characterize the distinction between *contradictions* and *disagreements* alongside a proposed architecture to identify contradictions and disagreement in medical guidelines.

  We showed the feasibility of using the proposed architecture with a simple approach, but a more thorough evaluation should be done on the task in the future. Perhaps, creating gold standard datasets on disagreements and contradictions between guidelines is the first step.

- An automated method of text analytics for computing conceptual differences between documents addressing the same topic. We showed that the differences in recommendation in guidelines can be computed to a large scale (69% to 86%)

from the concepts used in the text. Additionally, we introduced a novel graph clique-based algorithm/method for comparing the similarity of two collections of documents.

An obvious extension of this work would be to compare other groups of guidelines, e.g., European medical societies vs. US medical societies. We know that for years their recommendations, e.g., on managing blood cholesterol, differed. Another potential extension would be to experiment with other representations, such as more complex word and document embeddings, or with more subtle semantic representations based on entity and relationship extraction or formal models, cf. [231], and on formal modeling of contradictions, like the ones discussed in Chapter 3.

- Showing the applicability of the new neural network transformers and transfer learning in identifying informative statements. We improve state of the art on identifying conditional or condition-action statements by 5% to 25%.

  The open issues and possible extensions of this work can go in several directions. The most obvious next step, after identifying conditional sentences, is to extract the specifics: conditions, actions, and consequences.

  A discourse-oriented direction of analysis would allow us to find conditions, actions, and consequences spread over paragraphs or sections of texts. Combining discourse analysis with the extraction of specific entities and events should result in improved accuracy of both classification and extraction, and would open the possibility of applications, e.g., analysis of electronic health records (EHR). Finally, creating more annotated guidelines would lessen the problem of the lack of data mentioned in Section 5.3.

- A sentence-level semantic search engine and a corpus of medical guidelines with these capabilities: keyword and semantic search, alternative relevance scores,

and handling table data.

Perhaps, evaluating the relevance measures and using the semantic search introduced in 6.3.4 is the first possible future work. As it was discussed in 6.5, we need to add more reliable guidelines to the repository.

- We developed a data extraction and classification model on mapping medical records to a standardized list of blood transfusion orders. We achieved 83% accuracy in finding the most preferable matched for free-text reasons. We also proposed a method for extracting numerical conditions from the reasons. Our evaluation shows a 95% accuracy of extracting numerical conditions.

  Next steps for this study would include additional training on acronyms and concepts. Our results show that the situations that can represent some concepts, such as bleeding, are varied, and additional mapping can catch more of them.

REFERENCES

[1] C. B. Oliveira, C. G. Maher, R. Z. Pinto, A. C. Traeger, C.-W. C. Lin, J.-F. Chenot, M. van Tulder, and B. W. Koes, "Clinical practice guidelines for the management of non-specific low back pain in primary care: an updated overview," *European Spine Journal*, vol. 27, no. 11, pp. 2791–2803, 2018.

[2] H. Hematialam and W. Zadrozny, "Identifying condition-action statements in medical guidelines using domain-independent features," *arXiv preprint arXiv:1706.04206*, 2017.

[3] M. Hussain and S. Lee, "Information extraction from clinical practice guidelines: A step towards guidelines adherence," in *International Conference on Ubiquitous Information Management and Communication*, pp. 1029–1036, Springer, 2019.

[4] M. Hussain, J. Hussain, M. Sadiq, A. U. Hassan, and S. Lee, "Recommendation statements identification in clinical practice guidelines using heuristic patterns," in *2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pp. 152–156, IEEE, 2018.

[5] S. W. Stein and C. G. Thiel, "The history of therapeutic aerosols: a chronological review," *Journal of aerosol medicine and pulmonary drug delivery*, vol. 30, no. 1, pp. 20–41, 2017.

[6] W. Zadrozny, H. Hematialam, and L. Garbayo, "Towards semantic modeling of contradictions and disagreements: A case study of medical guidelines," *Proc. 12th International Conference on Computational Semantics (IWCS); arXiv preprint arXiv:1708.00850*, 2017.

[7] CDC, *Breast Cancer Screening Guidelines for Women.* Centers for Disease Control and Prevention, Mar 2017.

[8] R. Wenzina and K. Kaiser, "Identifying condition-action sentences using a heuristic-based information extraction method," in *Process Support and Knowledge Representation in Health Care*, pp. 26–38, Springer, 2013.

[9] M. Field and K. Lohr, *Clinical Practice Guidelines: Directions for a New Program.* National Academies Press, 1990.

[10] E. Steinberg, S. Greenfield, D. M. Wolman, M. Mancher, R. Graham, *et al.*, *Clinical practice guidelines we can trust.* National Academies Press, 2011.

[11] K. N. Lohr, M. J. Field, *et al.*, *Guidelines for clinical practice: from development to use*. National Academies Press, 1992.

[12] D. Stewart, "Throat infections in children," *Canadian Medical Association Journal*, vol. 105, no. 6, p. 559, 1971.

[13] D. Saslow, C. Boetes, W. Burke, S. Harms, M. O. Leach, C. D. Lehman, E. Morris, E. Pisano, M. Schnall, S. Sener, *et al.*, "American Cancer Society guidelines for breast screening with MRI as an adjunct to mammography," *CA: a cancer journal for clinicians*, vol. 57, no. 2, pp. 75–89, 2007.

[14] J. L. Paradise, "Tonsillectomy and adenoidectomy," *Pediatric otolaryngology*, vol. 2, pp. 1054–1065, 1996.

[15] U. P. S. T. Force *et al.*, "Screening for breast cancer: US Preventive Services Task Force recommendation statement.," *Annals of internal medicine*, vol. 151, no. 10, p. 716, 2009.

[16] W. H. Organization, *International travel and health: situation as on 1 January 2010*. World Health Organization, 2010.

[17] T. C. Hilton, R. C. Thompson, H. J. Williams, R. Saylors, H. Fulmer, and S. A. Stowers, "Technetium-99m sestamibi myocardial perfusion imaging in the emergency room evaluation of chest pain," *Journal of the American College of Cardiology*, vol. 23, no. 5, pp. 1016–1022, 1994.

[18] W. I. McDonald, A. Compston, G. Edan, D. Goodkin, H.-P. Hartung, F. D. Lublin, H. F. McFarland, D. W. Paty, C. H. Polman, S. C. Reingold, *et al.*, "Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis," *Annals of neurology*, vol. 50, no. 1, pp. 121–127, 2001.

[19] M. B. Landon, S. Leindecker, C. Y. Spong, J. C. Hauth, S. Bloom, M. W. Varner, A. H. Moawad, S. N. Caritis, M. Harper, R. J. Wapner, *et al.*, "The MFMU Cesarean Registry: factors affecting the success of trial of labor after previous cesarean delivery," *American journal of obstetrics and gynecology*, vol. 193, no. 3, pp. 1016–1023, 2005.

[20] M. C. Fiore, W. C. Bailey, S. J. Cohen, S. F. Dorfman, M. G. Goldstein, E. R. Gritz, R. B. Heyman, C. R. Jaen, T. E. Kottke, H. A. Lando, *et al.*, "Treating tobacco use and dependence: clinical practice guideline," *Rockville, MD: US Department of Health and Human Services*, pp. 00–0032, 2000.

[21] I. Sim and A. Berlin, "A framework for classifying decision support systems.," *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, vol. 2003, no. Figure 1, pp. 599–603, 2003.

[22] I. Sim, P. Gorman, R. A. Greenes, R. B. Haynes, B. Kaplan, H. Lehmann, and P. C. Tang, "Clinical decision support systems for the practice of evidence-based medicine," *Journal of the American Medical Informatics Association*, vol. 8, no. 6, pp. 527–534, 2001.

[23] M. A. Musen, B. Middleton, and R. A. Greenes, "Clinical decision-support systems," in *Biomedical informatics*, pp. 643–674, Springer, 2014.

[24] R. Moskovitch and Y. Shahar, "Vaidurya: A multiple-ontology, concept-based, context-sensitive clinical-guideline search engine," *Journal of Biomedical Informatics*, vol. 42, no. 1, pp. 11–21, 2009.

[25] J. Cowie and W. Lehnert, "Information extraction," *Communications of the ACM*, vol. 39, no. 1, pp. 80–91, 1996.

[26] C.-H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan, "A survey of web information extraction systems," *IEEE transactions on knowledge and data engineering*, vol. 18, no. 10, pp. 1411–1428, 2006.

[27] L. F. Rau, "Extracting company names from text," in *Artificial Intelligence Applications, 1991. Proceedings., Seventh IEEE Conference on*, vol. 1, pp. 29–32, IEEE, 1991.

[28] C. Thielen, "An approach to proper name tagging for german," *arXiv preprint cmp-lg/9506024*, 1995.

[29] S. Coates-Stephens, "The analysis and acquisition of proper names for the understanding of free text," *Computers and the Humanities*, vol. 26, no. 5-6, pp. 441–456, 1992.

[30] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pp. 188–191, Association for Computational Linguistics, 2003.

[31] M. Asahara and Y. Matsumoto, "Japanese named entity extraction with redundant morphological analysis," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pp. 8–15, Association for Computational Linguistics, 2003.

[32] R. Bunescu and M. Paşca, "Using encyclopedic knowledge for named entity disambiguation," in *11th conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[33] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[34] V. Yadav and S. Bethard, "A survey on recent advances in named entity recognition from deep learning models," *arXiv preprint arXiv:1910.11470*, 2019.

[35] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[36] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.," in *Proceedings of the AMIA Symposium*, p. 17, American Medical Informatics Association, 2001.

[37] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[38] O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[39] D. Jurafsky, *Speech & language processing*. Pearson Education India, 2000.

[40] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of English: The Penn Treebank," *Computational linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[41] I. G. N. Scottish, "British guideline on the management of asthma.," *Thorax*, vol. 58, p. i1, 2003.

[42] E. Loper and S. Bird, "NLTK: The natural language toolkit.," *URL http://arxiv. org/abs/cs/0205028*, 2002.

[43] D. M. Cer, M.-C. De Marneffe, D. Jurafsky, and C. D. Manning, "Parsing to Stanford Dependencies: Trade-offs between Speed and Accuracy.," in *LREC*, Floriana, Malta, 2010.

[44] J. R. Hobbs, "Resolving pronoun references," *Lingua*, vol. 44, no. 4, pp. 311–338, 1978.

[45] C. L. Sidner, "Focusing for interpretation of pronouns," *Computational Linguistics*, vol. 7, no. 4, pp. 217–231, 1981.

[46] N. Ge, J. Hale, and E. Charniak, "A statistical approach to anaphora resolution," in *Sixth Workshop on Very Large Corpora*, 1998.

[47] K. L. Wagstaff and C. Cardie, *Intelligent clustering with instance-level constraints*. Cornell University USA, 2002.

[48] V. Ng and C. Cardie, "Improving machine learning approaches to coreference resolution," in *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 104–111, Association for Computational Linguistics, 2002.

[49] F. Y. Choi, "Advances in domain independent linear text segmentation," in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pp. 26–33, Association for Computational Linguistics, 2000.

[50] J. C. Reynar, "Topic segmentation: Algorithms and applications," *University of Pennsylvania, PA*, 1998.

[51] M. Tsytsarau and T. Palpanas, "Survey on mining subjective data on the web," *Data Mining and Knowledge Discovery*, vol. 24, no. 3, pp. 478–514, 2012.

[52] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[53] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[54] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79–86, Association for Computational Linguistics, 2002.

[55] IBM, "IBM watson products and services," *IBM Cognitive advantage reports*, Oct 2017.

[56] M. T. Pazienza, M. Pennacchiotti, and F. M. Zanzotto, "Terminology extraction: an analysis of linguistic and statistical approaches," in *Knowledge mining*, pp. 255–279, Springer, 2005.

[57] K. Heylen and D. De Hertog, "Automatic term extraction," *Handbook of Terminology*, vol. 1, no. 01, 2015.

[58] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[59] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[60] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[61] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[62] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, pp. 19–27, 2015.

[63] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding with unsupervised learning," tech. rep., Technical report, OpenAI, 2018.

[64] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[65] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," *arXiv preprint arXiv:1804.07461*, 2018.

[66] J. Baldridge, "The OpenNLP project," *URL: http://opennlp. apache. org/index. html,(accessed 2 February 2012)*, p. 1, 2005.

[67] H. Cunningham, "GATE, a general architecture for text engineering," *Computers and the Humanities*, vol. 36, no. 2, pp. 223–254, 2002.

[68] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, and L. S. Zettlemoyer, "AllenNLP: A Deep Semantic Natural Language Processing Platform," *arXiv preprint arXiv:1803.07640*, 2017.

[69] L. Doyle, *Information Retrieval and Processing*. A Wiley-Becker & Hayes series book, Melville Publishing Company, 1975.

[70] V. Bush *et al.*, "As we may think," *The atlantic monthly*, vol. 176, no. 1, pp. 101–108, 1945.

[71] C. N. Mooers, "Information retrieval viewed as temporal signaling," in *Proceedings of the International Congress of Mathematicians*, vol. 1, pp. 572–573, 1950.

[72] G. Salton, "A Blueprint for Automatic Indexing," *SIGIR Forum*, vol. 16, pp. 22–38, Sept. 1981.

[73] F. Song and W. B. Croft, "A General Language Model for Information Retrieval," in *Proceedings of the eighth international conference on Information and knowledge management. ACM*, (New York, New York, USA), p. 316321, ACM Press, 1999.

[74] H. Turtle and W. B. Croft, "Inference Networks for Document Retrieval," in *Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '90, (New York, NY, USA), pp. 1–24, ACM, 1990.

[75] W. S. Cooper, "Getting beyond boole," *Information Processing & Management*, vol. 24, no. 3, pp. 243–248, 1988.

[76] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, pp. 613–620, nov 1975.

[77] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[78] M. E. Maron and J. L. Kuhns, "On Relevance, Probabilistic Indexing and Information Retrieval," *Journal of the ACM*, vol. 7, pp. 216–244, jul 1960.

[79] N. Fuhr, "Probabilistic models in information retrieval," *The computer journal*, vol. 35, no. 3, pp. 243–255, 1992.

[80] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986.

[81] C. J. V. Rijsbergen, *Information Retrieval*. Newton, MA, USA: Butterworth-Heinemann, 2nd ed., 1979.

[82] N. J. Belkin and W. B. Croft, "Retrieval techniques," in *Annual Review of Information Science and Technology, Vol. 22* (M. E. Williams, ed.), pp. 109–145, New York, NY, USA: Elsevier Science Inc., 1987.

[83] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and D. Johnson, "Terrier information retrieval platform," in *Proceedings of the 27th European Conference on Advances in Information Retrieval Research*, ECIR'05, (Berlin, Heidelberg), pp. 517–519, Springer-Verlag, 2005.

[84] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The Vocabulary Problem in Human-system Communication," *Commun. ACM*, vol. 30, pp. 964–971, Nov. 1987.

[85] C. Carpineto and G. Romano, "A Survey of Automatic Query Expansion in Information Retrieval," *ACM Computing Surveys*, vol. 44, pp. 1–50, jan 2012.

[86] X. Liu and W. B. Croft, "Cluster-based retrieval using language models," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 186–193, ACM, 2004.

[87] E. M. Voorhees and D. Harman, "The text retrieval conferences (trecs)," in *Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998*, pp. 241–273, Association for Computational Linguistics, 1998.

[88] D. Harman, "Towards interactive query expansion," in *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 321–331, ACM, 1988.

[89] D. Harman, "Relevance feedback revisited," in *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 1–10, ACM, 1992.

[90] I. Ruthven and M. Lalmas, "A survey on the use of relevance feedback for information access systems," *The Knowledge Engineering Review*, vol. 18, no. 2, pp. 95–145, 2003.

[91] R. Navigli, "Word Sense Disambiguation: A Survey," *ACM Comput. Surv.*, vol. 41, pp. 10:1–10:69, Feb. 2009.

[92] S. F. Weiss, "Learning to disambiguate," *Information Storage and Retrieval*, vol. 9, no. 1, pp. 33–41, 1973.

[93] A. Tombros, R. Villa, and C. J. Van Rijsbergen, "The effectiveness of query-specific hierarchic clustering in information retrieval," *Information processing & management*, vol. 38, no. 4, pp. 559–582, 2002.

[94] M. A. Hearst and J. O. Pedersen, "Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results," in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, (New York, NY, USA), pp. 76–84, ACM, 1996.

[95] S. Osinski and D. Weiss, "A concept-driven algorithm for clustering search results," *IEEE Intelligent Systems*, vol. 20, no. 3, pp. 48–54, 2005.

[96] C. Carpineto, S. Osiński, G. Romano, and D. Weiss, "A Survey of Web Clustering Engines," *ACM Comput. Surv.*, vol. 41, pp. 17:1–17:38, July 2009.

[97] A. V. Leouski and W. B. Croft, "An evaluation of techniques for clustering search results," tech. rep., MASSACHUSETTS UNIV AMHERST DEPT OF COMPUTER SCIENCE, 2005.

[98] Y. S. Maarek, R. Fagin, I. Z. Ben-Shaul, and D. Pelleg, "Ephemeral document clustering for web applications," in *IBM RESEARCH REPORT RJ 10186*, Citeseer, 2000.

[99] P. Ferragina and A. Gulli, "A personalized search engine based on web-snippet hierarchical clustering," *Software: Practice and Experience*, vol. 38, no. 2, pp. 189–225, 2008.

[100] K. Järvelin and J. Kekäläinen, "IR Evaluation Methods for Retrieving Highly Relevant Documents," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, (New York, NY, USA), pp. 41–48, ACM, 2000.

[101] M.-C. De Marneffe, A. N. Rafferty, and C. D. Manning, "Finding contradictions in text.," in *ACL*, vol. 8, pp. 1039–1047, 2008.

[102] J. Kloetzer et al, "Two-stage method for large-scale acquisition of contradiction pattern pairs using entailment," *EMNLP*, pp. 693–703, 2013.

[103] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st ed., 2009.

[104] M. C. McCord, J. W. Murdock, and B. K. Boguraev, "Deep parsing in Watson," *IBM Journal of Research and Development*, vol. 56, no. 3.4, pp. 3–1, 2012.

[105] A. H. McClintock, A. L. Golob, and M. B. Laya, "Breast Cancer Risk Assessment: A step-wise approach for primary care providers on the front lines of shared decision making," in *Mayo Clinic Proceedings*, vol. 95, pp. 1268–1275, Elsevier, 2020.

[106] L. E. Pace and N. L. Keating, "A systematic assessment of benefits and risks to guide breast cancer screening decisions," *Jama*, vol. 311, no. 13, pp. 1327–1335, 2014.

[107] M. Catillon, "Medical knowledge synthesis: A brief overview," 2017.

[108] W. Zadrozny and L. Garbayo, "A sheaf model of contradictions and disagreements. preliminary report and discussion," *arXiv preprint arXiv:1801.09036*, 2018.

[109] D. Christensen, J. Lackey, and T. Kelly, *The Epistemology of Disagreement: New Essays*. Oxford University Press, 2013.

[110] J. Lackey, *Taking Religious Disagreement Seriously*, pp. 299–316. Oxford University Press, 2014.

[111] P. Grim, A. Modell, N. Breslin, J. Mcnenny, I. Mondescu, K. Finnegan, R. Olsen, C. An, and A. Fedder, "Coherence and correspondence in the network dynamics of belief suites," *Episteme*, vol. 14, no. 2, pp. 233–253, 2017.

[112] L. Garbayo, *Epistemic Considerations on Expert Disagreement, Normative Justification, and Inconsistency Regarding Multi-criteria Decision Making*, pp. 35–45. Cham: Springer International Publishing, 2014.

[113] L. Garbayo, M. Ceberio, S. Bistarelli, and J. Henderson, "On modeling multi-experts multi-criteria decision-making argumentation and disagreement: Philosophical and computational approaches reconsidered," in *Constraint Programming and Decision Making: Theory and Applications* (M. Ceberio and V. Kreinovich, eds.), pp. 67–75, Springer International Publishing, 2018.

[114] L. Garbayo, "Dependence logic & medical guidelines disagreement: an informational (in) dependence analysis," in *Logic Colloquium 2019*, p. 112, 2019.

[115] N. Peek, C. Combi, R. Marin, and R. Bellazzi, "Thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes," *Artificial intelligence in medicine*, vol. 65, no. 1, pp. 61–73, 2015.

[116] J. Bowles, M. Caminati, S. Cha, and J. Mendoza, "A framework for automated conflict detection and resolution in medical guidelines," *Science of Computer Programming*, vol. 182, pp. 42–63, 2019.

[117] R. Tsopra, J.-B. Lamy, and K. Sedki, "Using preference learning for detecting inconsistencies in clinical practice guidelines: Methods and application to antibiotherapy," *Artificial intelligence in medicine*, vol. 89, pp. 24–33, 2018.

[118] K. C. Lee, B. V. Udelsman, J. Streid, D. C. Chang, A. Salim, D. H. Livingston, C. Lindvall, and Z. Cooper, "Natural Language Processing Accurately Measures Adherence to Best Practice Guidelines for Palliative Care in Trauma," *Journal of Pain and Symptom Management*, vol. 59, no. 2, pp. 225–232, 2020.

[119] S. A. Waheeb, N. Ahmed Khan, B. Chen, and X. Shang, "Machine learning based sentiment text classification for evaluating treatment quality of discharge summary," *Information*, vol. 11, no. 5, 2020.

[120] O. Seneviratne, A. K. Das, S. Chari, N. N. Agu, S. M. Rashid, C.-H. Chen, J. P. McCusker, J. A. Hendler, and D. L. McGuinness, "Enabling trust in clinical decision support recommendations through semantics," in *SeWeBMeDa@ ISWC*, 2019.

[121] X. Chen, H. Xie, G. Cheng, L. K. Poon, M. Leng, and F. L. Wang, "Trends and features of the applications of natural language processing techniques for clinical trials text analysis," *Applied Sciences*, vol. 10, no. 6, p. 2157, 2020.

[122] M. Ju, N. T. Nguyen, M. Miwa, and S. Ananiadou, "An ensemble of neural models for nested adverse drug events and medication extraction with subwords," *Journal of the American Medical Informatics Association*, vol. 27, no. 1, pp. 22–30, 2020.

[123] F. Benedetti, D. Beneventano, S. Bergamaschi, and G. Simonini, "Computing inter-document similarity with context semantic analysis," *Information Systems*, vol. 80, pp. 136–147, 2019.

[124] M. Rospocher, F. Corcoglioniti, and M. Dragoni, "Boosting document retrieval with knowledge extraction and linked data," *Semantic Web*, vol. 10, no. 4, pp. 753–778, 2019.

[125] M. Zhou, N. Duan, S. Liu, and H.-Y. Shum, "Progress in Neural NLP: Modeling, Learning, and Reasoning," *Engineering*, 2020, in press.

[126] N. A. Smith, "Contextual word representations: putting words into computers," *Communications of the ACM*, vol. 63, no. 6, pp. 66–74, 2020.

[127] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[128] W. Shalaby, W. Zadrozny, and H. Jin, "Beyond word embeddings: learning entity and concept representations from large scale knowledge bases," *Information Retrieval Journal*, vol. 22, no. 6, pp. 525–542, 2019.

[129] K. S. Kalyan and S. Sangeetha, "SECNLP: A survey of embeddings in clinical natural language processing," *Journal of biomedical informatics*, vol. 101, p. 103323, 2020.

[130] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, and F. Rudzicz, "A survey of word embeddings for clinical text," *Journal of Biomedical Informatics: X*, vol. 4, p. 100057, 2019.

[131] H. T. Nguyen, P. H. Duong, and E. Cambria, "Learning short-text semantic similarity with word embeddings and external knowledge sources," *Knowledge-Based Systems*, vol. 182, p. 104842, 2019.

[132] N. H. Tien, N. M. Le, Y. Tomohiro, and I. Tatsuya, "Sentence modeling via multiple word embeddings and multi-level comparison for semantic textual similarity," *Information Processing & Management*, vol. 56, no. 6, p. 102090, 2019.

[133] R. K. Lie, F. K. Chan, C. Grady, V. H. Ng, and D. Wendler, "Comparative effectiveness research: what to do when experts disagree about risks," *BMC medical ethics*, vol. 18, no. 1, p. 42, 2017.

[134] M. Schaekermann, G. Beaton, M. Habib, A. Lim, K. Larson, and E. Law, "Capturing Expert Arguments from Medical Adjudication Discussions in a Machine-readable Format," in *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 1131–1137, 2019.

[135] M. Schaekermann, G. Beaton, M. Habib, A. Lim, K. Larson, and E. Law, "Understanding expert disagreement in medical data analysis through structured adjudication," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–23, 2019.

[136] J. Grant and A. Hunter, "Analysing inconsistent first-order knowledgebases," *Artificial Intelligence*, vol. 172, no. 8-9, pp. 1064–1093, 2008.

[137] V. S. Subrahmanian and L. Amgoud, "A general framework for reasoning about inconsistency.," in *IJCAI*, pp. 599–504, 2007.

[138] J. Grant and A. Hunter, "Analysing inconsistent information using distance-based measures," *International Journal of Approximate Reasoning*, 2016.

[139] T. H. Tran *et al.*, "Inconsistency measures for probabilistic knowledge bases," in *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*, pp. 148–153, IEEE, 2017.

[140] L. Garbayo, W. Zadrozny, and H. Hematialam, "Converging in breast cancer diagnostic screening: A computational model proposal," *Diagnosis*, vol. 6, no. 4, p. eA60, 2019a.

[141] American College of Obstetricians-Gynecologists, "Practice bulletin no. 122: Breast cancer screening," *Obstetrics & Gynecology*, vol. 118, no. 2, 2011.

[142] A. P. Action, "Summary of Recommendations for Clinical Preventive Services," *American Academy of Family Physicians*, 2017.

[143] T. J. Wilt, R. P. Harris, and A. Qaseem, "Screening for cancer: advice for high-value care from the American College of Physicians," *Annals of internal medicine*, vol. 162, no. 10, pp. 718–725, 2015.

[144] C. H. Lee, D. D. Dershaw, D. Kopans, P. Evans, B. Monsees, D. Monticciolo, R. J. Brenner, L. Bassett, W. Berg, S. Feig, *et al.*, "Breast cancer screening with imaging: recommendations from the society of breast imaging and the acr on the use of mammography, breast mri, breast ultrasound, and other technologies for the detection of clinically occult breast cancer," *Journal of the American college of radiology*, vol. 7, no. 1, pp. 18–27, 2010.

[145] K. C. Oeffinger, E. T. Fontham, R. Etzioni, A. Herzig, J. S. Michaelson, Y.-C. T. Shih, L. C. Walter, T. R. Church, C. R. Flowers, S. J. LaMonte, *et al.*, "Breast cancer screening for women at average risk: 2015 guideline update from the American Cancer Society," *Jama*, vol. 314, no. 15, pp. 1599–1614, 2015.

[146] K. J. Jørgensen and S. Bewley, "Breast-cancer screening—viewpoint of the iarc working group," *N Engl J Med*, vol. 373, p. 1478, 2015.

[147] A. L. Siu, "Screening for breast cancer: US Preventive Services Task Force recommendation statement," *Annals of internal medicine*, vol. 164, no. 4, pp. 279–296, 2016.

[148] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *International conference on machine learning*, pp. 957–966, 2015.

[149] G. Monge, "Mémoire sur la théorie des déblais et des remblais," *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[150] R. Rehurek and P. Sojka, "Gensim—statistical semantics in Python," *Retrieved from gensim.org*, 2011.

[151] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, *et al.*, "An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition," *BMC bioinformatics*, vol. 16, no. 1, p. 138, 2015.

[152] J.-F. Chenot, B. Greitemann, B. Kladny, F. Petzke, M. Pfingsten, S. G. Schorr, *et al.*, "Non-specific low back pain," *Deutsches Ärzteblatt International*, vol. 114, no. 51-52, p. 883, 2017.

[153] M. Mansor, "The Malaysian low back pain management guidelines.first edition.," 2009.

[154] E. L. Marques, "The treatment of low back pain and scientific evidence," in *Low Back Pain*, IntechOpen, 2012.

[155] T. F. de Campos, "Low back pain and sciatica in over 16s: assessment and management NICE Guideline [NG59]," *J Physiother*, vol. 63, no. 2, p. 120, 2017.

[156] N. A. for Clinical Innovation, "Management of people with acute low back pain: model of care," *Chatswood*, 2016.

[157] A. Qaseem, T. J. Wilt, R. M. McLean, and M. A. Forciea, "Noninvasive treatments for acute, subacute, and chronic low back pain: a clinical practice guideline from the American College of Physicians," *Annals of internal medicine*, vol. 166, no. 7, pp. 514–530, 2017.

[158] T. O. P. L. B. P. W. Group *et al.*, "Evidence-informed primary care management of low back pain," *Edmonton, Canada: Toward Optimized Practice*, 2015.

[159] M. J. Stochkendahl, P. Kjaer, J. Hartvigsen, A. Kongsted, J. Aaboe, M. Andersen, M. Ø. Andersen, G. Fournier, B. Højgaard, M. B. Jensen, *et al.*, "National clinical guidelines for non-surgical treatment of patients with recent onset low back pain or lumbar radiculopathy," *European Spine Journal*, vol. 27, no. 1, pp. 60–75, 2018.

[160] P. Van Wambeke, A. Desomer, L. Ailiet, A. Berquin, C. Dumoulin, B. Depreitere, J. Dewachter, M. Dolphens, P. Forget, V. Fraselle, *et al.*, "Low back pain and radicular pain: assessment and management. Good Clinical Practice (GCP)Brussels: Belgian Health Care Knowledge Centre (KCE).," *KCE Report*, vol. 287, 2017.

[161] R. S. Padwal, B. R. Hemmelgarn, F. A. McAlister, D. W. McKay, S. Grover, T. Wilson, B. Penner, E. Burgess, P. Bolli, M. Hill, *et al.*, "The 2007 canadian Hypertension Education Program recommendations for the management of hypertension: Part 1–blood pressure measurement, diagnosis and assessment of risk," *Canadian Journal of Cardiology*, vol. 23, no. 7, pp. 529–538, 2007.

[162] T. G. Pickering, N. H. Miller, G. Ogedegbe, L. R. Krakoff, N. T. Artinian, and D. Goff, "Call to action on use and reimbursement for home blood pressure monitoring: a joint scientific statement from the American Heart Association, American Society of Hypertension, and Preventive Cardiovascular Nurses Association," *Hypertension*, vol. 52, no. 1, pp. 10–29, 2008.

[163] M. Malachias, M. Gomes, F. Nobre, A. Alessi, A. Feitosa, and E. Coelho, "7th brazilian guideline of arterial hypertension: chapter 2-diagnosis and classification," *Arquivos brasileiros de cardiologia*, vol. 107, no. 3, pp. 7–13, 2016.

[164] P. Lindsay, S. C. Gorber, M. Joffres, R. Birtwhistle, D. McKay, and L. Cloutier, "Recommendations on screening for high blood pressure in canadian adults," *Canadian Family Physician*, vol. 59, no. 9, pp. 927–933, 2013.

[165] A. Hartle, T. McCormack, J. Carlisle, S. Anderson, A. Pichel, N. Beckett, T. Woodcock, and A. Heagerty, "The measurement of adult blood pressure and management of hypertension before elective surgery: Joint Guidelines from the Association of Anaesthetists of Great Britain and Ireland and the British Hypertension Society," *Anaesthesia*, vol. 71, no. 3, pp. 326–337, 2016.

[166] J. T. Flynn, D. C. Kaelber, C. M. Baker-Smith, D. Blowey, A. E. Carroll, S. R. Daniels, S. D. de Ferranti, J. M. Dionne, B. Falkner, S. K. Flinn, *et al.*, "Clinical practice guideline for screening and management of high blood pressure in children and adolescents," *Pediatrics*, vol. 140, no. 3, 2017.

[167] P. A. James, S. Oparil, B. L. Carter, W. C. Cushman, C. Dennison-Himmelfarb, J. Handler, D. T. Lackland, M. L. LeFevre, T. D. MacKenzie, O. Ogedegbe, *et al.*, "2014 evidence-based guideline for the management of high blood pressure in adults: report from the panel members appointed to the eighth joint national committee (jnc 8)," *Jama*, vol. 311, no. 5, pp. 507–520, 2014.

[168] R. S. Padwal, B. R. Hemmelgarn, N. A. Khan, S. Grover, D. W. McKay, T. Wilson, B. Penner, E. Burgess, F. A. McAlister, P. Bolli, *et al.*, "The 2009 Canadian Hypertension Education Program recommendations for the management of hypertension: Part 1–blood pressure measurement, diagnosis and assessment of risk," *Canadian Journal of Cardiology*, vol. 25, no. 5, pp. 279–286, 2009.

[169] A. W. Chow, M. S. Benninger, I. Brook, J. L. Brozek, E. J. Goldstein, L. A. Hicks, G. A. Pankey, M. Seleznick, G. Volturo, E. R. Wald, *et al.*, "IDSA clinical practice guideline for acute bacterial rhinosinusitis in children and adults," *Clinical Infectious Diseases*, vol. 54, no. 8, pp. e72–e112, 2012.

[170] H. Hematialam, L. Garbayo, S. Gopalakrishnan, and W. W. Zadrozny, "A method for computing conceptual distances between medical recommendations: Experiments in modeling medical disagreement," *Applied Sciences*, vol. 11, no. 5, p. 2045, 2021.

[171] E. Bilici, G. Despotou, and T. N. Arvanitis, "The use of computer-interpretable clinical guidelines to manage care complexities of patients with multimorbid conditions: A review," *Digital health*, vol. 4, p. 2055207618804927, 2018.

[172] N. Sager, *Computerized discovery of semantic word classes in scientific fields*. New York University, 1977.

[173] R. Grishman, *Directions in artificial intelligence: Natural language processing*. Courant Institute of Mathematical Sciences, New York University, 1975.

[174] A. Nogales, Á. García-Tejedor, D. Monge, J. S. Vara, and C. Antón, "A survey of deep learning models in medical therapeutic areas," *Artificial Intelligence in Medicine*, p. 102020, 2021.

[175] M. Nadif and F. Role, "Unsupervised and self-supervised deep learning approaches for biomedical text mining," *Briefings in Bioinformatics*, vol. 22, no. 2, pp. 1592–1603, 2021.

[176] R. Gatta, M. Vallati, C. Fernandez-Llatas, A. Martinez-Millana, S. Orini, L. Sacchi, J. Lenkowicz, M. Marcos, J. Munoz-Gama, M. Cuendet, *et al.*, "Clinical guidelines: A crossroad of many research areas. challenges and opportunities in process mining for healthcare," in *International Conference on Business Process Management*, pp. 545–556, Springer, Cham, 2019.

[177] R. Gatta, M. Vallati, C. Fernandez-Llatas, A. Martinez-Millana, S. Orini, L. Sacchi, J. Lenkowicz, M. Marcos, J. Munoz-Gama, M. A. Cuendet, *et al.*, "What role can process mining play in recurrent clinical guidelines issues? a position paper," *International Journal of Environmental Research and Public Health*, vol. 17, no. 18, p. 6616, 2020.

[178] M. Peleg, S. Tu, J. Bury, P. Ciccarese, J. Fox, R. A. Greenes, R. Hall, P. D. Johnson, N. Jones, A. Kumar, *et al.*, "Comparing computer-interpretable guideline models: a case-study approach," *Journal of the American Medical Informatics Association*, vol. 10, no. 1, pp. 52–68, 2003.

[179] H. Cunningham, "Gate: A framework and graphical development environment for robust nlp tools and applications," in *Proc. 40th annual meeting of the association for computational linguistics (ACL 2002)*, pp. 168–175, 2002.

[180] D. R. Schlegel, K. Gordon, C. Gaudioso, and M. Peleg, "Clinical tractor: A framework for automatic natural language understanding of clinical practice guidelines," in *AMIA Annual Symposium Proceedings*, vol. 2019, p. 784, American Medical Informatics Association, 2019.

[181] R. Serban, A. ten Teije, F. van Harmelen, M. Marcos, and C. Polo-Conde, "Extraction and use of linguistic patterns for modelling medical guidelines," *Artificial intelligence in medicine*, vol. 39, no. 2, pp. 137–149, 2007.

[182] A. T. McCray, "The UMLS semantic network.," in *Proceedings. Symposium on Computer Applications in Medical Care*, pp. 503–507, American Medical Informatics Association, 1989.

[183] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, "Dive into deep learning. 2020," *URL https://d2l. ai*, 2020.

[184] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pp. 2672–2680, 2014.

[185] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[186] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[187] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[188] S. Sun, H. Shi, and Y. Wu, "A survey of multi-source domain adaptation," *Information Fusion*, vol. 24, pp. 84–92, 2015.

[189] A. Ramponi and B. Plank, "Neural unsupervised domain adaptation in nlp—a survey," *arXiv preprint arXiv:2006.00632*, 2020.

[190] S. Ruder, *Neural transfer learning for natural language processing*. PhD thesis, NUI Galway, 2019.

[191] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pp. 15–18, 2019.

[192] E. Laparra, S. Bethard, and T. A. Miller, "Rethinking domain adaptation for machine learning over clinical language," *JAMIA open*, vol. 3, no. 2, pp. 146–150, 2020.

[193] S. Rosenthal, S. Das, P.-Y. S. Hsueh, K. Barker, and C.-H. Chen, "Efficient goal attainment and engagement in a care manager system using unstructured notes," *JAMIA Open*, vol. 3, no. 1, pp. 62–69, 2020.

[194] J. PA, O. S, C. BL, and et al, "2014 evidence-based guideline for the management of high blood pressure in adults: Report from the panel members appointed to the eighth joint national committee (jnc 8)," *JAMA*, vol. 311, no. 5, pp. 507–520, 2014.

[195] B. T. S. S. I. G. Network *et al.*, "British guideline on the management of asthma.," *Thorax*, vol. 63, p. iv1, 2008.

[196] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60, 2014.

[197] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *arXiv preprint arXiv:1906.08237*, 2019.

[198] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[199] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676*, 2019.

[200] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets," *arXiv preprint arXiv:1906.05474*, 2019.

[201] E. A. Chi, J. Hewitt, and C. D. Manning, "Finding universal grammatical relations in multilingual BERT," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5564–5577, 2020.

[202] G. Jawahar, B. Sagot, and D. Seddah, "What does BERT learn about the structure of language?," in *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[203] R. Rosa and D. Mareček, "Inducing syntactic trees from BERT representations," *arXiv preprint arXiv:1906.11511*, 2019.

[204] Z. Luo, "Have Attention heads in BERT learned constituency grammar?," *arXiv preprint arXiv:2102.07926*, 2021.

[205] I. of Medicine (US) Committee on the Use of Complementary andAlternative Medicine by the American Public, "State of emerging evidence on cam," Jan 1970.

[206] A. Allot, Q. Chen, S. Kim, R. Vera Alvarez, D. C. Comeau, W. J. Wilbur, and Z. Lu, "LitSense: making sense of biomedical literature at sentence level," *Nucleic acids research*, vol. 47, no. W1, pp. W594–W599, 2019.

[207] Z. Lu, "PubMed and beyond: a survey of web tools for searching biomedical literature," *Database*, vol. 2011, 2011.

[208] T. Ohta, Y. Miyao, T. Ninomiya, Y. Tsuruoka, A. Yakushiji, K. Masuda, J. Takeuchi, K. Yoshida, T. Hara, J.-D. Kim, *et al.*, "An intelligent search engine and GUI-based efficient MEDLINE search tool based on deep syntactic parsing," in *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pp. 17–20, 2006.

[209] T. C. Rindflesch, H. Kilicoglu, M. Fiszman, G. Rosemblat, and D. Shin, "Semantic MEDLINE: An advanced information management application for biomedicine," *Information Services & Use*, vol. 31, no. 1-2, pp. 15–21, 2011.

[210] T. C. Rindflesch and M. Fiszman, "The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text," *Journal of biomedical informatics*, vol. 36, no. 6, pp. 462–477, 2003.

[211] H.-M. Müller, E. E. Kenny, and P. W. Sternberg, "Textpresso: an ontology-based information retrieval and extraction system for biological literature," *PLoS biol*, vol. 2, no. 11, p. e309, 2004.

[212] M. S. Siadaty, J. Shu, and W. A. Knaus, "Relemed: sentence-level search engine with relevance score for the MEDLINE database of biomedical articles," *BMC medical informatics and decision making*, vol. 7, no. 1, pp. 1–11, 2007.

[213] X. Wang, Y. Guan, W. Liu, A. Chauhan, E. Jiang, Q. Li, D. Liem, D. Sigdel, J. Caufield, P. Ping, *et al.*, "Evidenceminer: Textual evidence discovery for life sciences," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 56–62, 2020.

[214] X. Wang, W. Liu, A. Chauhan, Y. Guan, and J. Han, "Automatic textual evidence mining in covid-19 literature," *arXiv preprint arXiv:2004.12563*, 2020.

[215] J. Commission *et al.*, "Proceedings from the national summit on overuse," *National Summit on Overuse, Oakbrook Terrace, IL*, 2012.

[216] J. D. Roback, S. Caldwell, J. Carson, R. Davenport, M. J. Drew, A. Eder, M. Fung, M. Hamilton, J. R. Hess, N. Luban, *et al.*, "Evidence-based practice guidelines for plasma transfusion," *Transfusion*, vol. 50, no. 6, pp. 1227–1239, 2010.

[217] J. Carson, B. Grossman, S. Kleinman, A. Tinmouth, M. Marques, M. Fung, and J. Holcomb, "Clinical transfusion medicine committee of the AABB. red blood cell transfusion: a clinical practice guideline from the aabb*," *Ann Intern Med*, 2012.

[218] R. M. Kaufman, B. Djulbegovic, T. Gernsheimer, S. Kleinman, A. T. Tinmouth, K. E. Capocelli, M. D. Cipolle, C. S. Cohn, M. K. Fung, B. J. Grossman, *et al.*, "Platelet transfusion: a clinical practice guideline from the AABB," *Annals of internal medicine*, vol. 162, no. 3, pp. 205–213, 2015.

[219] C. Shivade, P. Raghavan, E. Fosler-Lussier, P. J. Embi, N. Elhadad, S. B. Johnson, and A. M. Lai, "A review of approaches to identifying patient phenotype cohorts using electronic health records," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 221–230, 2014.

[220] M. Ball, R. Patel, R. D. Hayes, R. J. Dobson, and R. Stewart, "TextHunter–a user friendly tool for extracting generic concepts from free text in clinical research," in *AMIA Annual Symposium Proceedings*, vol. 2014, p. 729, American Medical Informatics Association, 2014.

[221] C. A. Bejan, F. Xia, L. Vanderwende, M. M. Wurfel, and M. Yetisgen-Yildiz, "Pneumonia identification using statistical feature selection," *Journal of the American Medical Informatics Association*, vol. 19, no. 5, pp. 817–823, 2012.

[222] R. G. Jackson, R. Patel, N. Jayatilleke, A. Kolliakou, M. Ball, G. Gorrell, A. Roberts, R. J. Dobson, and R. Stewart, "Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project," *BMJ Open*, vol. 7, no. 1, 2017.

[223] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, and D. Sontag, "Learning a health knowledge graph from electronic medical records," *Scientific reports*, vol. 7, no. 1, pp. 1–11, 2017.

[224] I. Y. Chen, M. Agrawal, S. Horng, and D. Sontag, "Robustly extracting medical knowledge from EHRs: A case study of learning a health knowledge graph," in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020*, pp. 19–30, World Scientific, 2019.

[225] X. Ma, T. Imai, E. Shinohara, S. Kasai, K. Kato, R. Kagawa, and K. Ohe, "EHR2CCAS: A framework for mapping EHR to disease knowledge presenting causal chain of disorders–chronic kidney disease example," *Journal of Biomedical Informatics*, vol. 115, p. 103692, 2021.

[226] R. I. Bjarnadottir and R. J. Lucero, "What can we learn about fall risk factors from EHR nursing notes? a text mining study," *eGEMs*, vol. 6, no. 1, 2018.

[227] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[228] R. Cobb, S. Puri, D. Z. Wang, T. Baslanti, and A. Bihorac, "Knowledge extraction and outcome prediction using medical notes," in *ICML workshop on Role of Machine Learning in Transforming Healthcare*, 2013.

[229] K. L. O'Brien, Y. Chen, and L. Uhl, "Assessing inpatient platelet ordering practice: evaluation of computer provider order entry overrides," *Vox Sanguinis*, 2020.

[230] M. Hill-Strathy, P. H. Pinkerton, T. A. Thompson, A. Wendt, A. Collins, R. Cohen, W. O. BComm, T. Cameron, Y. Lin, W. Lau, *et al.*, "Evaluating the appropriateness of platelet transfusions compared with evidence-based platelet guidelines: An audit of platelet transfusions at 57 hospitals," *Transfusion*, vol. 61, no. 1, pp. 57–71, 2021.

[231] Q. Zhu, X. Li, A. Conesa, and C. Pereira, "GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text," *Bioinformatics*, vol. 34, pp. 1547–1554, 12 2017.