COMPUTER-AIDED APPROACHES TO ENHANCE SYSTEMATIC REVIEW DEVELOPMENT

by

Duy Duc An Bui

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biomedical Informatics

The University of Utah

May 2016

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Duy Duc An Bui**

has been approved by the following supervisory committee members:

Guilherme Del Fiol, Chair **3/11/2016**
Date Approved

John F. Hurdle, Member **3/25/2016**
Date Approved

Siddhartha Jonnalagadda, Member **3/11/2016**
Date Approved

Gang Luo, Member **3/11/2016**
Date Approved

Bruce E. Bray, Member **3/11/2016**
Date Approved

and by Wendy W. Chapman, Chair/Dean of

the Department/College/School of **Biomedical Informatics**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Medical knowledge learned in medical school can become quickly outdated given the tremendous growth of the biomedical literature. It is the responsibility of medical practitioners to continuously update their knowledge with recent, best available clinical evidence to make informed decisions about patient care. However, clinicians often have little time to spend on reading the primary literature even within their narrow specialty. As a result, they often rely on systematic evidence reviews developed by medical experts to fulfill their information needs. At the present, systematic reviews of clinical research are manually created and updated, which is expensive, slow, and unable to keep up with the rapidly growing pace of medical literature. This dissertation research aims to enhance the traditional systematic review development process using computer-aided solutions.

The first study investigates query expansion and scientific quality ranking approaches to enhance literature search on clinical guideline topics. The study showed that unsupervised methods can improve retrieval performance of a popular biomedical search engine (PubMed). The proposed methods improve the comprehensiveness of literature search and increase the ratio of finding relevant studies with reduced screening effort.

The second and third studies aim to enhance the traditional manual data extraction process. The second study developed a framework to extract and classify texts from PDF reports. This study demonstrated that a rule-based multipass sieve approach is more effective than a machine-learning approach in categorizing document-level structures and that classifying and filtering publication metadata and semistructured texts enhances the performance of an information extraction system. The proposed method could serve as a document processing step in any text mining research on PDF documents. The third study proposed a solution for the computer-aided data extraction by recommending relevant sentences and key phrases extracted from publication reports. This study demonstrated that using a machine-learning classifier to prioritize sentences for specific data elements performs equally or better than an abstract screening approach, and might save time and reduce errors in the full-text screening process.

In summary, this dissertation showed that there are promising opportunities for technology enhancement to assist in the development of systematic reviews. In this modern age when computing resources are getting cheaper and more powerful, the failure to apply computer technologies to assist and optimize the manual processes is a lost opportunity to improve the timeliness of systematic reviews. This research provides methodologies and tests hypotheses, which can serve as the basis for further large-scale software engineering projects aimed at fully realizing the prospect of computer-aided systematic reviews.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to express my special thanks to Dr. Qing Zeng-Treitler and Dr. Guilherme Del Fiol, two of my advisors who significantly contributed to my doctorate training and dissertation research. Dr. Qing provided guidance and financial support in my early stage while Dr. Guilherme had a critical role in the success of my dissertation research. I've learned a lot from their supervisions in both research and life experiences.

I also want to thank my committee members, Dr. Siddhartha Jonnalagadda, Dr. John F. Hurdle, Dr. Bruce E. Bray, and Dr. Gang Luo for their dedication to the guidance and review of this dissertation. Despite having a diverse set of backgrounds, we were able to harmonize and take advantage of the differences to move the research forward.

I appreciate the financial support at different times of Qing, John, and Siddhartha's research grants and the student award sponsored by the John Morgan family. Those funding supports were great helps to keep me fully concentrating on studying and research.

I want to thank my family and my friends for their social support. I have had a wonderful time in Utah partially because people here are really nice. My friends from the BMI department, Northwestern University, and the Vietnamese community are wonderful to be with. Lastly, I'm always grateful to have a family although they are almost half-earth distance away. That's the only place on earth that I can share the emotional happiness and sadness. Without them, I might not have the confidence and motivation to march to the end of this journey.

CHAPTER 1

INTRODUCTION

1.1 Objectives and Hypotheses

The systematic review process attempts to comprehensively identify, appraise, and synthesize the best quality research to find reliable answers to research questions based on the best available evidence [1]. Healthcare practitioners, in the process of seeking information for patient care, become the information consumers for some kinds of systematic review products such as Clinical Practice Guidelines or Cochrane reviews. Systematic reviews have been criticized as slow and effortful to develop, and potentially as biased [2]. This is partially because the systematic review process involves many labor-intensive manual steps that face human limitations such as limited time and resources, in addition to human inconsistency and errors. The main objective of this research is to explore computer-aided techniques that can help humans in performing systematic reviews. Specifically, we investigated algorithms to improve the citation retrieval and the data extraction processes for systematic reviews.

In the following three studies presented in three distinct chapters, the following hypotheses will be explored.

- H1.1 Query expansion techniques improve the performance of a widely used biomedical search engine (i.e., PubMed) (Chapter 3).

- H1.2 An unsupervised citation ranking approach performs better than a general purpose machine-learning classifier in ranking scientifically sound studies (Chapter 3).
- H2.1 In the classification of PDF texts, the rule-based multipass sieve approach is more accurate than the machine-learning approach. (Chapter 4).
- H2.2 PDF text classification improves performance of information extraction from full-text publications (Chapter 4).
- H3.1 Machine-learning classification to prioritize sentences in full-text performs equally or better than abstract screening (Chapter 5).

1.2 Rationale for Analysis

Evidence-based resources (EBR) such as practice guidelines (CPG) and systematic reviews (SR) are important expert-synthesized information sources to enable evidence-based medicine practice [3]. Experts perform systematic reviews of best available evidence to develop EBR. At present state, the development of systematic reviews still relies on an extensive amount of manual effort. Therefore, the production and updating of EBR is often costly, slow, and unable to keep up with the rapid growth of the biomedical literature [4, 5]. Citations indexed in PubMed have grown from 4 million (pre 1975) to 22 million today [5]. It takes 2.5 to 6.5 years for a primary study publication to be included and published in a new SR [6] and takes 1 to 1.5 years to finish peer-review for CPG development [7]. As a result, about 23% of SRs have not updated new evidence in 2 years after first publication, [6] and many clinical questions were not found in existing SRs [8]. Those issues of limited time and resources can make EBRs easily become outdated and suboptimal for patient care.

That highlights the need of investigating computer technologies to aid humans with this labor-intensive task.

Researchers have investigated computerized techniques to aid EBR developers with the systematic review development. However, there are many research gaps that have not been filled in this area. This dissertation research attempts to address some of those gaps in the course of three projects.

First, researchers have investigated automated and semiautomated approaches to aid with citation screening. The prominent approaches were based on machine-learning, active-learning, and rule-based methods [9, 10]. Those approaches always need some sort of labeled data from a process of manual review to train a classification model or to derive a rule set. However, in the early stages of systematic review development, training data are rare and insufficient to train a competent classifier. As a result, searchers often rely on typical functionalities that search engines provide to fulfill the citation retrieval and screening task. Chapter 3 investigates alternative approaches for query expansion and citations ranking that outperform the standard functionality of a popular biomedical search engine (PubMed).

Second, data extraction to generate evidence summaries is a standard process in systematic review development. Studies on information extraction (IE) have investigated techniques to automatically extract key data elements from texts, which have the potential to aid the manual extraction task. The sources of extraction are often Medline abstracts, PubMed Central (PMC) archives, and journal websites [11-13]. However, the data extraction process requires extraction of information from the full-text study reports rather than the study abstract, and full-text reports are not always available in PMC and journal websites. Chapters 4 and 5 investigate IE techniques to extracting key information from clinical trials published in PDF format. Chapter 4 focuses on solving problems associated with the PDF file format. In a PDF document, contents are often mixed with publication metadata (e.g., header, footer, author information, journal information) and semistructure (e.g., tables, figures) texts. Publication metadata are often not relevant to systematic review development, and add noise to IE systems. Semistructured texts can contain relevant information, but they often do not consist of prose narrative form and require different extraction strategies than narrative texts. In Chapter 4, text classification techniques to categorize PDF texts are investigated, and the impact on an information extraction system is measured.

Last, Chapter 5 investigates a computer system to aid in the extraction of clinical trial characteristics from full-text PDF publications. The system has the ability to find relevant sentences specific to target data elements and to recommend key phrases to help the work of SR developers.

To increase the practicality of the results, this research developed gold standards that are very close to real-world systematic reviews. Clinical practice guidelines developed by The American College of Cardiology (ACC) / American Heart Association (AHA), and systematic reviews developed by the Cochrane Collaboration were used as sources for gold standards. These are popular evidence-based resources that are commonly used in patient care [14, 15].

1.3 Significant Contributions

This research proposes and evaluates promising computerized techniques that can aid human reviewers with systematic review development. First, our experiments demonstrated that the proposed approaches for query expansion and citation ranking improve retrieval performances of a widely used biomedical search engine (PubMed). The approaches were validated in retrieving relevant high-quality citations for cardiovascular guideline development.

Information extraction from full-text publications is sparse; extraction from PDF documents is even rarer and more challenging. The second and third studies provided evidence that it is possible to apply advanced natural language processing techniques to extract key clinical trial information from a PDF document. The second study solves the heterogeneity of PDF texts by proposing a text classification algorithm that can automatically categorize PDF texts into three dimensions: grammatical texts, semistructured texts, and publication metadata. The rule-based multipass sieve approach demonstrated a superior performance over a machine-learning classifier. The last study demonstrated the feasibility of text mining solutions that might assist humans in extracting clinical data elements from primary study reports. Future enhancements based on this approach are promising to change the traditional data extraction process, which currently relies on expensive manual review approaches.

1.4 References

[1] D.J. Cook, C.D. Mulrow, R.B. Haynes, Systematic reviews: synthesis of best evidence for clinical decisions, Ann. Intern. Med. 126(5) (1997) 376-380.

[2] I. Roberts, K. Ker, P. Edwards, D. Beecher, D. Manno, E. Sydenham, The knowledge system underpinning healthcare is not fit for purpose and must change, BMJ 350 (2015) h2463.

[3] D.L. Sackett, W.M. Rosenberg, J.A. Gray, R.B. Haynes, W.S. Richardson, Evidence based medicine: what it is and what it isn’t, BMJ (Clinical research ed). 312 (7023) (1996) 71–72.

[4] M. Ware, M. Mabe, An Overview of Scientific and Scholarly Journal Publishing, The STM Report, 2009.

[5] Statistical Reports on MEDLINE®/PubMed® Baseline Data, <https://www.nlm.nih.gov/bsd/licensee/baselinestats.html>.

[6] K.G. Shojania, M. Sampson, M.T. Ansari, J. Ji, S. Doucette, D. Moher, How quickly do systematic reviews go out of date? A survival analysis, Ann. Intern. Med. 147 (4) (2007) 224–233.

[7] ACCF/AHA TaskForce on Practice Guidelines, Methodology Manual and Policies From the ACCF/AHA Task Force on Practice Guidelines 2010, <http://assets.cardiosource.com/Methodology_Manual_for_ACC_AHA_Writing_Committees.pdf>.

[8] P. Bragge, O. Clavisi, T. Turner, E. Tavender, A. Collie, R.L. Gruen, The global evidence mapping initiative: scoping research in broad topic areas, BMC Med. Res. Methodol. 11 (1) (2011) 92.

[9] A.M. Cohen, W.R. Hersh, K. Peterson, P.Y. Yen, Reducing workload in systematic review preparation using automated citation classification, J. Am. Med. Inform. Assoc. 13 (2) (2006) 206–219. March–April. PubMed PMID:16357352. PubMed Central PMCID: 1447545.

[10] M. Fiszman, E. Ortiz, B.E. Bray, T.C. Rindflesch, Semantic processing to support clinical guideline development, in: AMIA Annual Symposium proceedings/AMIA Symposium AMIA Symposium, 2008, pp. 187–191. PubMed PMID:18999127. Pubmed Central PMCID: 2656081.

[11] F. Boudin, J.-Y. Nie, J.C. Bartlett, R. Grad, P. Pluye, M. Dawes, Combining classifiers for robust PICO element detection, BMC Med. Inform. Decis. Mak. 10 (2010) 29.

[12] S. Kiritchenko, B. de Bruijn, S. Carini, J. Martin, I. Sim, ExaCT: automatic extraction of clinical trial characteristics from journal publications, BMC Med. Inform. Decis. Mak. 10 (2010) 56.

[13] W. Hsu, W. Speier, R.K. Taira, Automated extraction of reported statistical analyses: towards a logical representation of clinical trial literature, in: AMIA Annual Symposium Proceedings/AMIA Symposium AMIA Symposium, 2012, pp. 350–359.

[14] M. Farahzadi, A. Shafiee, A. Bozorgi, M. Mahmoudian, S. Sadeghian, Assessment of adherence to ACC/AHA guidelines in primary management of patients with NSTEMI in a referral cardiology hospital, Crit. Pathw. Cardiol. 14(1) (2015) 36-38.

[15] R. Bortolus, F. Blom, F. Filippini, M.N. van Poppel, E. Leoncini, D.J. de Smit, P.P. Benetollo, M.C. Cornel, H.E. de Walle, P. Mastroiacovo, Prevention of congenital malformations and other adverse pregnancy outcomes with 4.0 mg of folic acid: community-based randomized clinical trial in Italy and the Netherlands, BMC Pregnancy Childbirth 14(1) (2014) 166.

CHAPTER 2

BACKGROUND

2.1 Challenges to Evidence-Based Medicine

Practice of evidence-based medicine (EBM) requires integrating individual clinical expertise and the best external evidence in making decisions about patient care. However, health care practitioners have little time to keep up with the tremendous growth of biomedical literature. In 2009, there were about 25400 peer-reviewed journals, and the number increases 3.5% a year [1]. PubMed citations have grown from 4 million in 1975 to 22 million today at a rate of about half million a year [2]. Each year, about 3000 clinical trial studies have posted results in ClinicalTrial.gov [3]. Fraser and Dunstan showed that it is almost impossible to keep up with medical literature even within a narrow specialty [4]. In a review of information-seeking behavior, Karen Davies showed that clinicians' lack of time, issues with information technology, and limited search skills are top barriers for information searching [5]. Clinicians are overwhelmed by new information, and most clinical questions still remain unanswered. In 1985, Covell et al. showed that clinicians raised two questions every three patients, and 70% of questions remained unanswered [6]. In a recent systematic review, Del Fiol et al. showed that clinicians raised roughly one question out of every two patients seen, and more than 60% of these questions were not answered [7].

As a result, clinicians search information from evidence-based resources (EBR) compiled by clinical experts or professional organizations. Examples include the ACCF/AHA guidelines, Cochrane systematic reviews, Drug effectiveness reviews, and evidence summaries such as those provided by UpToDate. However, the development and update of those knowledge sources is costly, slow and unable to keep up with the rate of new evidence in the medical literature. In a 2003 survey of guideline developers, the average cost for CPG development was $200,000 per guideline in the US [8]. High quality guidelines that meet strict quality criteria [9, 10] require more time and resources. The total expenditure of the Cochrane Collaboration in 2011 fiscal year was $2.4 million [11], and this number increased to 3.9 million in 2013 [12]. It takes 2.5 to 6.5 years for a primary study publication to be included and published in a new systematic review [6] and takes 1 to 1.5 years to finish peer-review for guideline development [7]. About 23% of reviews have not been updated with new evidences within 2 years after their first publication [13] and many clinical questions are not covered in existing reviews [14]. Thus, development and updating evidence-based knowledge is unable to keep up with the rate of new evidence in the biomedical literature. As a result, evidence-based resources become quickly outdated and suboptimal for patient care.

2.2 The Systematic Review Development Process

Experts performed a standardized systematic review development process which involves a series of scientifically rigorous steps. The Cochrane Handbook listed eight general steps to prepare a systematic review [15]:

- Defining the review question and criteria for including studies
- Searching the literature

- Selecting studies and extracting data from study reports
- Assessing risk of bias of included studies
- Performing data analysis and meta-analyses
- Addressing publication biases
- Summarizing the results in tables/figures
- Interpreting the results and drawing conclusions

This dissertation research focuses on the literature search and data extraction tasks, since they are conducted early in the SR development process and are critical for the quality of the review. Those two steps also require a significant amount of manual effort, for which there is substantial opportunity for optimization through the adoption of advanced technology solutions.

Literature search is the early step in the systematic review process, conducted by individuals with database searching skills (librarians or information specialist). The goal of literature database search is to find relevant articles among millions in electronic literature databases. The literature search can be preliminary to understand the scope of the problem, or comprehensive to fully cover a specific clinical question. The main objective of literature search is to maximize the sensitivity to identify all relevant studies, while keeping the screening workload manageable. The 2011 ACCF/AHA's manual for clinical guideline development described the need for literature search to be comprehensive, and key to the development of valid guidelines [16]. The Cochrane handbook for systematic reviews highlights that "searches should seek high sensitivity, which may result in relatively low precision" [15, 17].

After determining a set of eligible studies, reviewers perform data extraction to generate evidence summary tables. The goal of data extraction is to collect all information that is relevant to the review question from original publication reports. The evidence summary table is a convenient place to validate the extracted information by seeking agreements from multiple authors. It also serves as a key data source for the quality assessment and meta-analysis steps. Data extraction conducted by humans has high prevalence of errors [18, 19]. Therefore, independent data extraction by at least 2 authors is recommended [15].

The systematic review development process, specifically the literature search and data extraction tasks, essentially rely on manual efforts, which are limited by time, knowledge, inconsistency and errors. In the age of information technology, such limitations can be overcome by computer technologies. In the following sections, the current researches on information retrieval and information extraction that can help in systematic reviews are discussed.

2.3 Information Retrieval (IR)

2.3.1 Medline, PubMed, and MeSH

MEDLINE is a database of citations and abstracts for the biomedical literature worldwide, maintained by the National Library of Medicine (NLM). At present, MEDLINE contains more than 22 million citations in 5,600 journals and 40 different languages [20]. PubMed is a free search engine that facilitates access to the MEDLINE database. Searchers can submit and get results via PubMed's web interface or programmatically via the Entrez Programming Utilities (E-utilities) [21]. To facilitate semantic search, the NLM maintained a comprehensive controlled vocabulary, called

Medical Subject Headings (MeSH), for indexing MEDLINE citations. Currently, MeSH has 27,455 unique headings/concepts and more than 220,000 entry terms that help with the assignment of those headings.

2.3.2 Related Work

Literature search often relies on querying IR systems such as PubMed, Google Scholar, and Scopus to obtain access to publications relevant to the research inquiry. Each IR system follows slightly different strategies to process the user's query, rank the results, and present information to searchers. IR systems have been successfully optimized to handle diverse specialized domains, such as computer science & engineering, biology, chemistry, and medicine. Similarly, IR methods optimized to narrow clinical systematic review topics have the potential to perform better than general IR approaches. Lu et al. [22] developed the automated query expansion algorithm Automatic Term Mapping (ATM). Using 2006 and 2007 TREC Genomics dataset, they demonstrated that query expansion using MeSH headings improved PubMed search over the word-based approach. Crespo et al. [23] developed a medical image retrieval system. They demonstrated that query expansion using MeSH headings and MeSH ontology significantly improved image retrieval. Damoni et al. [24] demonstrated that MeSH Concepts could significantly improve the precision of retrieval for PubMed searches related to rare and chronic diseases. Therefore, query expansion techniques have been demonstrated to improve information retrieval performance. However, the application of this technique to systematic review development has not been attempted. Unlike previous research, the development of systematic reviews imposes a much higher expectation for near perfect recall than precision.

Traditional information retrieval or question answering systems used vector space models to represent the queries and documents, and rank documents by similarities between vectors [25]. For short and generic queries such as guideline conditions (e.g., heart failure), thousands of retrieved citations can share the search keywords and cannot distinguish well using the vector-space model. For instance, citations that have keywords repeated once or twice might not match the review inclusion criteria. By default, PubMed sorts the citations by the time they were added to the MEDLINE database. PubMed also has a relevance-based ranking that uses a technique similar to the vector-space model [26]. Those ranking mechanisms do not consider the scientific quality of the studies, an important factor for study assessment in systematic reviews.

Haynes et al. [27, 28] developed PubMed Clinical Query filters using a set of manual rules to retrieve high-quality clinical studies. Filters are available for different topics, such as treatment, diagnosis, prognosis, systematic reviews, and medical genetics.

Kilicoglu et al. [29] implemented an ensemble approach combining several machine-learning classifiers (Naïve Bayes, support vector machine (SVM), and boosting) to identify scientifically sound studies. The classifier had five basic features: words, MEDLINE metadata, sematic predications, relations, and UMLS concepts. The classifier was trained on 10,000 randomly selected citations, and achieved 82.5% precision and 84.3% recall on an independent test set of 2000 citations.

Cohen et al. [30-33] also explored machine-learning approaches to help prioritize citations for screening in drug effectiveness reviews. They employed the Support Vector Machine (SVM) algorithm to train a number of features extracted from MEDLINE citations such as unigram, n-gram, MeSH terms, and UMLS concepts. Their classifiers achieved decent performance in cross-validation and prospective evaluations. Their ML classifiers have the potential to improve the ranking ability of information retrieval systems. However, in systematic reviews, new questions are often raised that do not have sufficient past data to train a competent machine-learning model. Therefore, standard ranking algorithms of search engine and general purpose ML classifiers (e.g., Kilicoglu's classifier) are more generalizable and scalable than topic-specific ML classifiers.

2.4 Information Extraction

2.4.1 Extraction From Full-Text Reports

Information extraction research in the past decade has investigated automated technique to extract study characteristics from the biomedical literature. However, the goal of automating or semiautomating data extraction for systematic reviews still needs further research. In a recent systematic review, Jonnalagadda et al. found there was no IE system that has been tailored to the systematic review process [34]. Previous IE systems focused on only 1 to 7 data elements out of 52 data elements commonly used in systematic reviews. To optimally benefit systematic reviews, IE systems need to extract data elements directly from full-text reports since this is a standard requirement for data extraction in the development of systematic reviews.

Kiritchenko et al. [35, 36] developed a tool called "ExaCT", which extracts 21 clinical trial data elements to help human reviewers in compiling a clinical trial database. Their method first used a machine-learning classifier to select top 5 relevant sentences for each element, then used hand-crafted weak extraction rules (i.e., regular expression matching) to collect values for each element. ExaCT takes full-text inputs from journal websites in HTML or XML formats.

Lin et al. [37] used a conditional random field approach to extract 3 publication metadata elements and 10 key study characteristics from full-text reports related to oncological and cardiovascular studies. They chose Rich Text Format (RTF) files downloaded from PubMed Central (PMC) as the sources of extraction. RTF is a proprietary document file format developed by Microsoft Corporation. At present, PMC no longer supports RTF download. Zhu et al. [38] employed rule-based approach based on constituency-tree parser to extract patient-related attributes. This study collected texts from Trip Answers and PubMed websites.

2.4.2 Extraction From PDF Reports

To our knowledge, there has been no research on PDF extraction to support systematic reviews. Most previous studies on PDF extraction focused on recognition of logical document structure, or extraction of basic elements that are not used in systematic reviews. Table 2.1 summarizes a representative set of PDF extraction studies and the target data elements.

Table 2.1 - Summary of data elements used in PDF extraction studies.

Authors/Software name	Data elements
Chao and Fan [39]	Text Layer, Image Layer, Vector Graphic Layer
Constantin et al. [40]	Title, author, abstract, body text, section, figure, table, reference, url, email, page number, side note
Kboubi [41]	Table, Cell
Klampfl et al. [42, 43]	Table, Cell, Metadata, Decorations, Captions, Main Texts, Headings,
Kern et al. [44]	Title, Journal, Abstract, Author , E-Mail, Affiliation
Luong [45]	Title, table, figure, headers, references, page number, note, keywords, equation, email, copyright, author, affiliation.
Oro et al. [46]	Table, Cell

Studies that extract data elements from PDF reports that might potentially benefit systematic reviews are underinvestigated in the current state of text mining research. Garcia-Remesal et al. [47] developed an algorithm using the finite-state machine to extract amino acids from PDF documents. They used PDFBox to extract raw texts and applied heuristic rules to recognize the amino acid sequences. Their method is specific to amino acid recognition, which lacks the generalizability to systematic reviews. In the data extraction step of systematic review development, common data elements such as patient population, intervention, and outcome mentions require more advanced natural language processing methods. In this dissertation research, we attempt to extract systematic review data elements from PDF reports for better supporting the data extraction process in the development of systematic reviews (Chapter 5 and Chapter 6).

2.5 References

[1] M. Ware, M. Mabe, An Overview of Scientific and Scholarly Journal Publishing, The STM Report, 2009.

[2] Statistical Reports on MEDLINE®/PubMed® Baseline Data, <https://www.nlm.nih.gov/bsd/licensee/baselinestats.html>.

[3] Trends, Charts, and Maps: ClinicalTrials.gov, <http://clinicaltrial.gov/ct2/resources/trends>.

[4] A.G. Fraser, F.D. Dunstan, On the impossibility of being expert, BMJ (Clinical Research ed). 341 (2010) c6815. PubMed PMID: Medline:21156739.

[5] K. Davies, J. Harrison, The information-seeking behaviour of doctors: a review of the evidence, Health Inform. Libr. J. 24 (2) (2007) 78–94. June. PubMed PMID: 17584211.

[6] D.G. Covell, G.C. Uman, P.R. Manning, Information needs in office practice: are they being met?, Ann. Intern. Med. 103(4) (1985) 596-599.

[7] G. Del Fiol, T.E. Workman, P.N. Gorman, Clinical questions raised by clinicians at the point of care: a systematic review, JAMA Int. Med. (2014).

[8] J.S. Burgers, R. Grol, N.S. Klazinga, M. Makela, J. Zaat, A. Collaboration, Towards evidence-based clinical practice: an international survey of 18 clinical guideline programs, Int. J. Qual. Health Care: J. Int. Soc. Qual. Health Care/ ISQua 15 (1) (2003) 31–45. February. PubMed PMID: 12630799.

[9] H.J. Schünemann, A. Fretheim, A.D. Oxman, Research WACoH. Improving the use of research evidence in guideline development: 1. Guidelines for guidelines, Health Res. Policy Syst. 4 (13) (2006) 1–6.

[10] T. Turner, M. Misso, C. Harris, S. Green, Development of evidence-based clinical practice guidelines (CPGs): comparing approaches, Implement. Sci.: IS 3 (45) (2008) 1–8.

[11] M.E. Schaafsma, The Cochrane Collaboration Treasurer's Report, 2012.

[12] Cochrane Collaboration, Directors' Reports and Financial Statements, 2013.

[13] K.G. Shojania, M. Sampson, M.T. Ansari, J. Ji, S. Doucette, D. Moher, How quickly do systematic reviews go out of date? A survival analysis, Ann. Intern. Med. 147 (4) (2007) 224–233.

[14] P. Bragge, O. Clavisi, T. Turner, E. Tavender, A. Collie, R.L. Gruen, The global evidence mapping initiative: scoping research in broad topic areas, BMC Med. Res. Methodol. 11 (1) (2011) 92.

[15] J.P. Higgins, S. Green, Cochrane Handbook for Systematic Reviews of Interventions, Wiley Online Library, 2008.

[16] ACCF/AHA TaskForce on Practice Guidelines, Methodology Manual and Policies From the ACCF/AHA Task Force on Practice Guidelines 2010, <http://assets.cardiosource.com/Methodology_Manual_for_ACC_AHA_Writing_Committees.pdf>.

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

[18] A.P. Jones, T. Remmington, P.R. Williamson, D. Ashby, R.L. Smyth, High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews, J. Clin. Epidemiol. 58 (7) (2005) 741–742.

[19] P.C. Gotzsche, A. Hrobjartsson, K. Maric, B. Tendal, Data extraction errors in meta-analyses that use standardized mean differences, JAMA : The journal of the American Medical Association 298 (4) (2007) 430-437.

[20] Fact Sheet MEDLINE® <http://www.nlm.nih.gov/pubs/factsheets/medline.html>

[21] E. Sayers, Entrez Programming Utilities Help, <http://www.ncbi.nlm.nih.gov/books/NBK25499. 2009>.

[22] Z. Lu, W. Kim, W.J. Wilbur, Evaluation of query expansion using MeSH in PubMed, Inf. Retr. Boston. 12 (1) (2009) 69–80.

[23] M. Crespo Azcarate, J. Mata Vazquez, M. Mana Lopez, Improving image retrieval effectiveness via query expansion using MeSH hierarchical structure, J. Am. Med. Inform. Assoc.: JAMIA 20 (6) (2013) 1014–1020. PubMed PMID: Medline:22952301. English.

[24] S.J. Darmoni, L.F. Soualmia, C. Letord, M.-C. Jaulent, N. Griffon, B. Thirion, A. Neveol, Improving information retrieval using Medical Subject Headings Concepts: a test case on rare and chronic diseases, J. Med. Libr. Assoc. 100 (3) (2012) 176-183.

[25] C.D. Manning, P. Raghavan, H. Schütze, An introduction to information retrieval, Cambridge University Press, 2008.

[26] National Library of Medicine, PubMed Help, 2015, <http://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.Sorting_your_search_results>.

[27] R.B. Haynes, K.A. McKibbon, N.L. Wilczynski, S.D. Walter, S.R. Werre, Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey, BMJ 330 (7501) (2005) 1179.

[28] R.B. Haynes, N.L. Wilczynski, Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey, BMJ 328 (7447) (2004) 1040.

[29] H. Kilicoglu, D. Demner-Fushman, T.C. Rindflesch, N.L. Wilczynski, R.B. Haynes, Towards automatic recognition of scientifically rigorous clinical research evidence, J. Am. Med. Inform. Assoc. 16 (1) (2009) 25–31. January–February. PubMed PMID: 18952929. Pubmed Central PMCID: 2605595.

[30] A.M. Cohen, K. Ambert, M. McDonagh, A prospective evaluation of an automated classification system to support evidence-based medicine and systematic review, in: AMIA Annual Symposium Proceedings/AMIA Symposium AMIA Symposium, 2010, pp. 121–125. PubMed PMID: 21346953. Pubmed Central PMCID: 3041348.

[31] A.M. Cohen, W.R. Hersh, K. Peterson, P.Y. Yen, Reducing workload in systematic review preparation using automated citation classification, J. Am. Med. Inform.

Assoc. 13 (2) (2006) 206–219. March–April. PubMed PMID:16357352. PubMed Central PMCID: 1447545.

[32] A.M. Cohen, Optimizing feature representation for automated systematic review work prioritization, in: AMIA Annual Symposium Proceedings: American Medical Informatics Association, 2008.

[33] A.M. Cohen, K. Ambert, M. McDonagh, Cross-topic learning for work prioritization in systematic review creation and update, J. Am. Med. Inform. Assoc. 16 (5) (2009) 690–704.

[34] S.R. Jonnalagadda, P. Goyal, M.D. Huffman, Automating data extraction in systematic reviews: a systematic review, Syst. Rev. 4(1) (2015) 78.

[35] S. Kiritchenko, B. de Bruijn, S. Carini, J. Martin, I. Sim, ExaCT: automatic extraction of clinical trial characteristics from journal publications, BMC Med. Inform. Decis. Mak. 10 (2010) 56.

[36] B. de Bruijn, S. Carini, S. Kiritchenko, J. Martin, I. Sim, Automated information extraction of key trial design elements from clinical trial publications, in: AMIA Annual Symposium Proceedings/AMIA Symposium AMIA Symposium, 2008, pp. 141–145.

[37] S. Lin, J.-P. Ng, S. Pradhan, J. Shah, R. Pietrobon, M.-Y. Kan, Extracting formulaic and free text clinical research articles metadata using conditional random fields, In: Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents, 2010, Association for Computational Linguistics, 2010, 90-95.

[38] H. Zhu, Y. Ni, P. Cai, Z. Qiu, F. Cao, Automatic extracting of patient-related attributes: disease, age, gender and race, Stud. Health Technol. Inform. 180 (2012) 589–593.

[39] H. Chao, J. Fan, Layout and content extraction for pdf documents, in: Document Analysis Systems VI, Springer, 2004, pp. 213–224.

[40] A. Constantin, S. Pettifer, A. Voronkov, PDFX: fully-automated PDF-to-XML conversion of scientific literature, In: Proceedings of the 2013 ACM symposium on Document engineering, ACM, 2013, 177-180.

[41] F. Kboubi, A.H. Chabi, M.B. Ahmed (Eds.), Table recognition evaluation and combination methods, Document Analysis and Recognition, 2005 Proceedings Eighth International Conference on, IEEE , 2005.

[42] S. Klampfl, K. Jack, R. Kern, A comparison of two unsupervised table recognition methods from digital scientific articles, D-Lib Mag. 20 (11) (2014) 7.

[43] S. Klampfl, R. Kern, An unsupervised machine learning approach to body text and table of contents extraction from digital scientific articles, In: Research and Advanced Technology for Digital Libraries, Springer , 2013 , pp. 144–155

[44] R. Kern, K. Jack, M. Hristakeva, M. Granitzer, TeamBeam meta-data extraction from scientific literature, D-Lib Magazine 18 (7) (2012) 1.

[45] M.T. Luong, T.D. Nguyen, M.Y. Kan, Logical structure recovery in scholarly articles with rich document features, Multimedia Storage Retrieval Innovat. Digital Lib. Syst. (2012) 270.

[46] E. Oro, M. Ruffolo (Eds.), PDF-TREX: An approach for recognizing and extracting tables from PDF documents, Document Analysis and Recognition, 2009 ICDAR'09 10th International Conference on, IEEE, 2009.

[47] M. Garcia-Remesal, V. Maojo, J. Crespo, A knowledge engineering approach to recognizing and extracting sequences of nucleic acids from scientific literature, Conf Proc IEEE Eng Med Biol Soc 2010, 2010, 1081-1084.

CHAPTER 3

AUTOMATICALLY FINDING RELEVANT CITATIONS FOR CLINICAL GUIDELINE DEVELOPMENT

Duy Duc An Bui, Siddhartha Jonnalagadda, Guilherme Del Fiol

Journal of Biomedical Informatics[1]

3.1 Abstract

Literature database search is a crucial step in the development of clinical practice guidelines and systematic reviews. In the age of information technology, the process of literature search is still conducted manually, therefore it is costly, slow, and subject to human errors. In this research, we sought to improve the traditional search approach using innovative query expansion and citation ranking approaches.

We developed a citation retrieval system composed of query expansion and citation ranking methods. The methods are unsupervised and easily integrated over the PubMed search engine. To validate the system, we developed a gold standard consisting of citations that were systematically searched and screened to support the development of cardiovascular clinical practice guidelines. The expansion and ranking methods were evaluated separately and compared with baseline approaches.

Compared with the baseline PubMed expansion, the query expansion algorithm

[1] D.D.A. Bui, S. Jonnalagadda, G. Del Fiol, Automatically finding relevant citations for clinical guideline development, J. Biomed. Inform. 57 (2015) 436-445.

improved recall (80.2% vs. 51.5%) with small loss on precision (0.4% vs. 0.6%). The algorithm could find all citations used to support a larger number of guideline recommendations than the baseline approach (64.5% vs. 37.2%, p<0.001). In addition, the citation ranking approach performed better than PubMed's "most recent" ranking (average precision +6.5%, recall@k +21.1%, p<0.001), PubMed's rank by "relevance" (average precision +6.1%, recall@k +14.8%, p<0.001), and the machine-learning classifier that identifies scientifically sound studies from MEDLINE citations (average precision +4.9%, recall@k +4.2%, p<0.001).

Our unsupervised query expansion and ranking techniques are more flexible and effective than PubMed's default search engine behavior and the machine-learning classifier. Automated citation finding is promising to augment the traditional literature search.

3.2. Introduction

The practice of evidence-based medicine requires integrating individual clinical expertise and the best available evidence in making decisions about patient care. However, health care practitioners have little time to keep up with the rapid growth in the biomedical literature. In 2009, there were about 25,400 peer-reviewed journals, and the number increases 3.5% a year [1]. Citations indexed in PubMed have grown from 4 million (pre 1975) to 22 million today [2]. Each year, about 3000 clinical trial studies have posted results in ClinicalTrial.gov [3]. Fraser and Dunstan showed that it is almost impossible to keep up with the medical literature even within a narrow specialty [4]. In a review of information-seeking behavior, Davies showed that clinicians' lack of time, issues with information technology, and limited search skills are top barriers for information searching

[5]. As a result, most clinical questions raised by clinicians at the point of care remain unanswered. In a recent systematic review, Del Fiol et al. showed that clinicians raised roughly one question out of every two patients seen and over 60% of these questions were not answered [6]. To cope with information overload, clinicians rely on existing expert-compiled resources such as clinical practice guidelines (CPG) to fulfill their information needs [7]. However, the development and update of CPGs is costly, slow, and unable to keep up with the rate of new evidence in the medical literature. In a 2003 survey of guideline developers, the average cost for CPGs development was $200,000 per guideline in the United States [8]. High-quality guidelines that meet strict quality criteria [9, 10] require more time and resources. Time required for finishing peer-review for a cardiology guideline published by The American College of Cardiology (ACC) and American Heart Association (AHA) was from 12 to 18 months [11]. In summary, the rapid pace of new published literature can quickly make the CPGs outdated and suboptimal for clinical decision-making.

In guideline development, experts perform systematic reviews of the available evidence, which involves a series of scientifically rigorous steps [11]. The two first and important steps are a systematic literature search followed by screening for relevant citations. Literature search involves identifying possibly relevant studies from electronic literature databases. Citation screening involves quickly scanning abstract and full-text manuscripts to assess the eligibility of studies. Informatics research has investigated automated and semiautomated methods to aid with citation screening [12-16]. Fiszman et al. were among the first research groups introducing informatics solutions to support clinical guideline development [15, 16]. They developed a semantic filter to automatically classify relevant citations. Similarly, Cohen et al. investigated a machine-learning approach to solve a classification problem in drug effectiveness reviews [12, 17]. To meet the needs of citation screening, those methods aimed for a balance between recall and precision. However, recall is more important than precision in systematic literature search. The 2011 ACCF/AHA's manual for clinical guideline development described the need for literature search to be comprehensive, and key to the development of valid guidelines [11]. The Cochrane handbook for systematic reviews highlights that "searches should seek high sensitivity, which may result in relatively low precision." [18]. In the present study, we investigated the literature search stage and aimed to maximize recall while controlling the impact on precision. We developed and assessed query expansion and ranking methods to enhance information retrieval performance in the context of clinical guideline development. The solution was based on an extension of PubMed's search engine, optimized to retrieve and rank relevant studies for cardiovascular guidelines.

There have been previous works that we leveraged to inform our system [15-17, 19-21]. Fiszman's gold standard included citations that were used to support 30 clinical questions [16]. Our work sought for a larger gold standard, which includes citations to support more than 600 guideline recommendations. Research on query expansion showed that using MeSH concepts and MeSH hierarchy can improve performance of image retrieval and biological question retrieval [19, 20]. Our query expansion method was also based on finding relevant MeSH concepts, but was optimized to retrieve guideline conditions.

Traditional information retrieval or question answering systems rank documents by relevance or similarity to the user query. Generic queries (e.g., "heart failure") can generate thousands of documents that share the search keywords. PubMed by default sorts the results by recently added date, without considering relevancy and scientific quality. Informatics research has investigated machine-learning approaches to prioritize citation screening in systematic reviews [14, 22, 23]. Yet machine-learning approaches are arguably not flexible, since they require sufficient high-quality training data and often do not generalize well to new domains. Unsupervised ranking methods have been investigated in the citation retrieval studies by Jonnalagadda et al. [24, 25]. Their method assigned weights based on journal impact measures; however, the method validation was limited to the "heart failure" topic. In the present research, we developed novel unsupervised query expansion and citation ranking methods with a larger gold standard that includes cardiovascular conditions. We then compared the performance of these methods with PubMed's query expansion and ranking, and a machine-learning classifier.

3.3. Materials and Methods

Our study design consisted of three main parts: (1) development of a gold standard composed of studies used in the development of cardiovascular guidelines; (2) iterative development of a citation finding system composed of two main components: query expansion and citation ranking; and (3) evaluation of each system component using standard information retrieval metrics and comparison with baseline approaches. Figure 3.1 depicts the summarization of our system architecture and study design.

3.3.1 Gold Standard

The gold standard consisted of citations that have been used to support guideline practice recommendations. We focused on the cardiovascular guidelines published by the

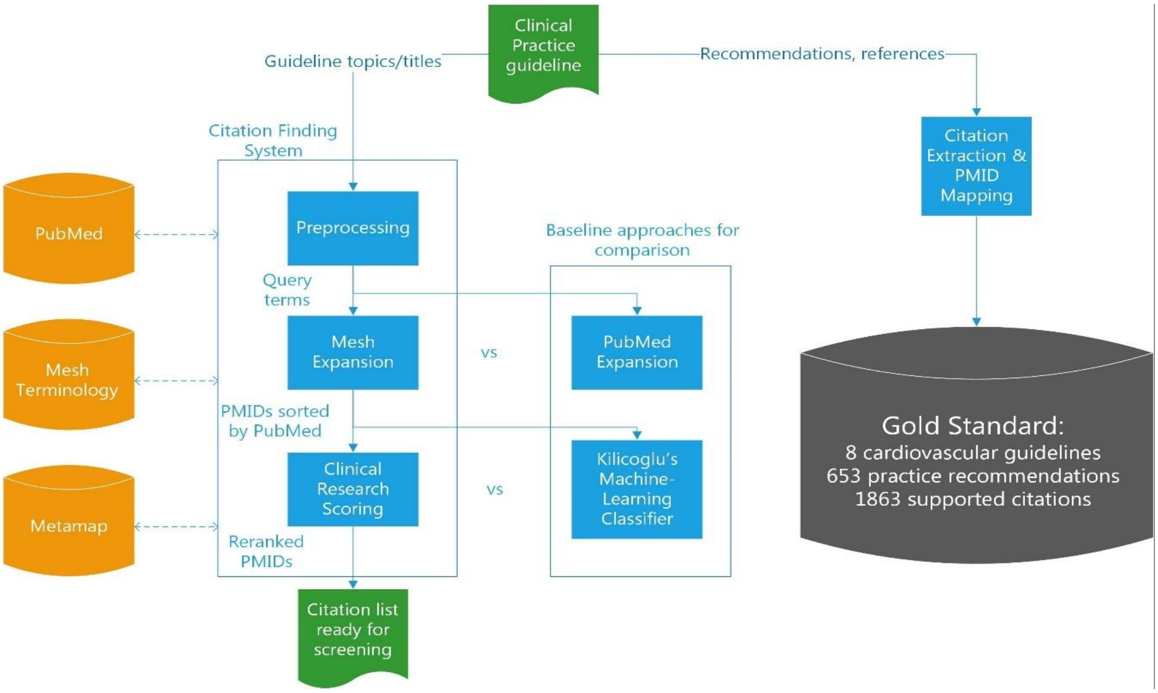

Clinical Practice guideline
Guideline topics/titles
Recommendations, references
Citation Finding System
Citation Extraction & PMID Mapping
PubMed
Preprocessing
Baseline approaches for comparison
Query terms
Mesh Terminology
Mesh Expansion
vs
PubMed Expansion
Gold Standard:
8 cardiovascular guidelines
653 practice recommendations
1863 supported citations
PMIDs sorted by PubMed
Clinical Research Scoring
vs
Kilicoglu's Machine-Learning Classifier
Metamap
Reranked PMIDs
Citation list ready for screening

Figure 3.1 - Overview of the citation finding system and the study design.

American College of Cardiology (ACC) and the American Heart Association (AHA). The full revision cardiovascular guidelines developed by the ACC/AHA and published from 2010 to 2014 were retrieved using a PubMed search. Since the majority of guideline topics are about complete management of a condition, we focused on retrieving condition topics in this study. Topics about interventions or diagnostic procedures are reserved for future research. For those guidelines discussing the comprehensive management of cardiovascular conditions, we performed the following steps to build the gold standard: (1) Extracted all the citations listed in the "References" section of the guideline; (2) extracted the guideline recommendations whose evidence sources were provided in the guideline and the citations that were used as evidence sources to support each recommendation; and (3) automatically mapped those citations in free-text to PubMed IDs using the NCBI Batch Citation Matcher tool [26]. Table 3.1 shows examples of guideline recommendations, supporting citations, and their corresponding PMIDs.

Table 3.1 - Examples of extracted guideline recommendations, supported citations, and PMID mappings for the "Guideline for the Management of Patients With Atrial Fibrillation" (2014).

Guideline recommendations	Supported citations
Selection of antithrombotic therapy should be based on the risk of thromboembolism irrespective of whether the AF pattern is paroxysmal, persistent, or permanent (167-170).	167. New oral anticoagulants for stroke prevention in atrial fibrillation: impact of gender, heart failure, diabetes mellitus and paroxysmal atrial fibrillation [27]. PMID: 23253272 168. Distribution and risk profile of paroxysmal, persistent, and permanent atrial fibrillation in routine clinical practice: insight from the real-life global survey evaluating patients with atrial fibrillation international registry [28]. PMID: 22787011 169. Efficacy and safety of dabigatran compared to warfarin in patients with paroxysmal, persistent, and permanent atrial fibrillation: results from the RE-LY (Randomized Evaluation of Long-Term Anticoagulation Therapy) study [29]. PMID: 22361407 170. Prevention of stroke in patients with atrial fibrillation: current strategies and future directions [30]. PMID: 25534093
Control of the ventricular rate using a beta blocker or nondihydropyridine calcium channel antagonist is recommended for patients with paroxysmal, persistent, or permanent AF (267-269).	267. Ventricular rate control in chronic atrial fibrillation during daily activity and programmed exercise: a crossover open-label study of five drug regimens [31]. PMID: 9973007 268. Efficacy of oral diltiazem to control ventricular response in chronic atrial fibrillation at rest and during exercise [32]. PMID: 3805530 269. The Atrial Fibrillation Follow-up Investigation of Rhythm Management (AFFIRM) study: approaches to control rate in atrial fibrillation [33]. PMID: 15063430

Table 3.1 - continued

Guideline recommendations	Supported citations
Intravenous administration of a beta blocker or nondihydropyridine calcium channel blocker is recommended to slow the ventricular heart rate in the acute setting in patients without preexcitation. In hemodynamically unstable patients, electrical cardioversion is indicated (270-273).	270. Efficacy and safety of esmolol vs propranolol in the treatment of supraventricular tachyarrhythmias: a multicenter double-blind clinical trial [34]. PMID: 3904379 271. A placebo-controlled trial of continuous intravenous diltiazem infusion for 24-hour heart rate control during atrial fibrillation and atrial flutter: a multicenter study [35]. PMID: 1894861 272. Intravenous diltiazem is superior to intravenous amiodarone or digoxin for achieving ventricular rate control in patients with acute uncomplicated atrial fibrillation [36]. PMID: 19487941 273. Esmolol versus verapamil in the acute treatment of atrial fibrillation or atrial flutter [37]. PMID: 2564725

3.3.2 System Overview

The system is an extension of PubMed's search engine to enhance the ability to retrieve citations for clinical guideline development. The system has a preprocessing stage and two other main stages: query expansion and document ranking. The query expansion stage aims to improve recall, while the document ranking aims to improve precision on top-ranked documents.

3.3.2.1 Preprocessing

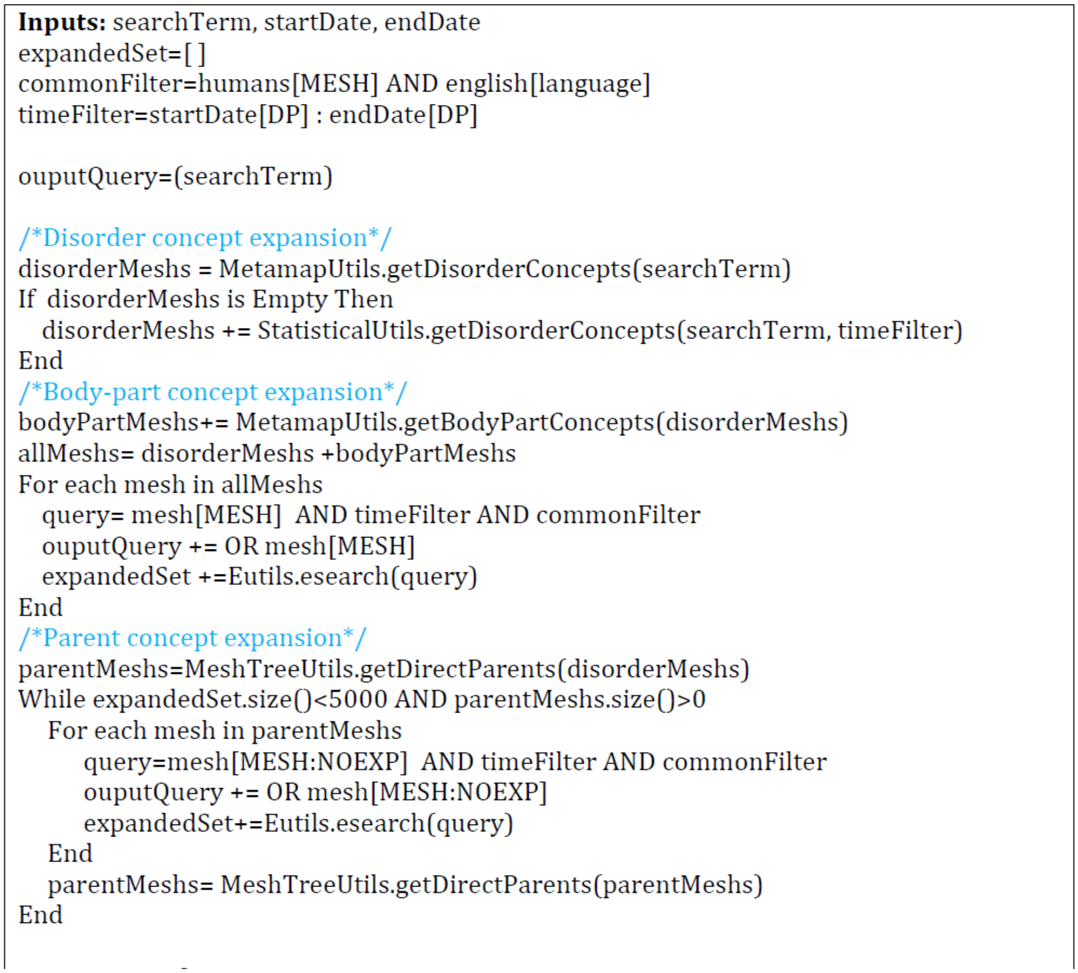
This step takes the title of the guideline as input and extracts the conditions of interest. Since there is little variation among guideline titles, we used simple regular expression rules such as words following "Patients With", "diagnosis and treatment of", and "management of" to extract main conditions from guideline titles (e.g., "Guideline for the Management of Patients With Atrial Fibrillation", "Guideline for the diagnosis and treatment of hypertrophic cardiomyopathy", "Guidelines for the diagnosis and management of patients with Thoracic Aortic Disease"). This step also detects whether a particular guideline focuses on one or more conditions. For instance, the phrase "Extracranial Carotid and Vertebral Artery Disease" was broken into two conditions: "Extracranial Carotid Disease" and "Vertebral Artery Disease".

3.3.2.2 Query Expansion

Based on the extracted condition terms, we conducted a search using PubMed's default search behavior. When entering a query on the PubMed search interface, PubMed automatically expands the query to maximize recall. For instance, PubMed expands the query "atrial fibrillation" by injecting additional MeSH terms and keywords: "atrial fibrillation"[MeSH concepts] OR ("atrial"[All Fields] AND "fibrillation"[All Fields]) OR "atrial fibrillation"[All Fields]. We used the results of PubMed expansion as the baseline to compare with our expansion approach. Our approach aims to find relevant and meaningful MeSH terms of the condition topics. Additional MeSH terms were injected to original query using the Boolean OR operator.

We consistently applied a set of filters (i.e., publication date, human study, and English language) for all queries generated. We considered other filters, such as hasabstract and the Haynes clinical filters[38], but those filters led to missing important eligible studies.

We developed an algorithm (Figure 3.2) to expand the seed query using MeSH



```
Inputs: searchTerm, startDate, endDate
expandedSet=[ ]
commonFilter=humans[MESH] AND english[language]
timeFilter=startDate[DP] : endDate[DP]

ouputQuery=(searchTerm)

/*Disorder concept expansion*/
disorderMeshs = MetamapUtils.getDisorderConcepts(searchTerm)
If  disorderMeshs is Empty Then
   disorderMeshs += StatisticalUtils.getDisorderConcepts(searchTerm, timeFilter)
End
/*Body-part concept expansion*/
bodyPartMeshs+= MetamapUtils.getBodyPartConcepts(disorderMeshs)
allMeshs= disorderMeshs +bodyPartMeshs
For each mesh in allMeshs
   query= mesh[MESH]  AND timeFilter AND commonFilter
   ouputQuery += OR mesh[MESH]
   expandedSet +=Eutils.esearch(query)
End
/*Parent concept expansion*/
parentMeshs=MeshTreeUtils.getDirectParents(disorderMeshs)
While expandedSet.size()<5000 AND parentMeshs.size()>0
    For each mesh in parentMeshs
         query=mesh[MESH:NOEXP]  AND timeFilter AND commonFilter
         ouputQuery += OR mesh[MESH:NOEXP]
         expandedSet+=Eutils.esearch(query)
    End
    parentMeshs= MeshTreeUtils.getDirectParents(parentMeshs)
End
```

Figure 3.2 - Pseudo-code for the query expansion algorithm.

resources (MeSH descriptors, MeSH Tree) and a natural language processing application (MetaMap [39]). The algorithm takes input as a single search query and outputs the expanded query. If there are multiple queries (multiple conditions), they were joined by the Boolean OR operator. Eventually, the query is adjusted by the common filter and applied the PubMed sorting mechanisms. To conduct a PubMed query, we formulated the PubMed query into the URL syntax and used the Entrez Programming Utilities (E-utilities) [40] to submit and retrieve results from the NCBI servers. The algorithm uses the following methods to find relevant MeSH concepts.

Disorder concept expansion attempts to find MeSH concepts that best describe the condition of interest using a concept-mapping method. We used MetaMap [41] to map narrative terms found in the Preprocessing stage into UMLS concepts. MetaMap was restricted to the MeSH terminology. The UMLS concepts were translated to MeSH concepts by querying the MRCONSO table [42]. We used the MeSH descriptors and MeSH Tree [43] to populate MeSH metadata and select concepts that have the semantic type "Disease or Syndrome". Concepts whose ancestors have this semantic type were also extracted.

Statistical expansion is based on the assumption that documents are likely relevant to a query if the extracted terms are mentioned in the document titles. The statistical expansion method first retrieved all articles that include the exact search term in the title. MeSH concepts of those articles were retrieved, aggregated, and sorted by frequency. The highest frequency concept having the semantic type "Disease or Syndrome" was selected. The statistical expansion is triggered if the concept-mapping approach does not recognize any concepts.

In some guidelines, the condition of interest is related to abnormalities in specific anatomical locations (e.g., heart valves, aortic valve). In exploratory work, we observed that using body-part concepts could improve recall in some queries. To find body-part concepts, we run MetaMap on the disorder concept entry terms, filter out the generic concepts, and select concepts having the semantic type "Body Part, Organ, or Organ Component".

Parent expansion looks for direct parent concepts by iteratively traversing the MeSH Tree. Using parent concepts in some circumstances can improve recall, but may substantially impact precision. Hence, the algorithm only uses parent expansion when the expansion set has not reached a specific threshold, and disables expansion to other MeSH children (e.g., using tag [MESH: NOEXP]).

We maintain a stop list of MeSH concepts to be filtered out from expansion. The list contains three general concepts for cardiovascular topics: *Disease*, *Heart Diseases*, and *Heart*. We investigated the technique to generate the stop list automatically, but it was not quite as successful as constructing manually. Our strategy is to test the algorithm in more diverse topics until we identify a pattern for a successful stop list.

3.3.2.3 Document Ranking

We present three ways searchers can obtain a ranked list of citations: (1) Use PubMed's sorting functionalities, (2) Use a general-purpose machine-learning classifier to identify clinical sound studies, and (3) Use our proposed scoring approach for clinical research studies.

3.3.2.3.1 PubMed sorting functionalities

PubMed offers 7 ways to sort order for search results: Most Recent, Relevance, Publication Date, First Author, Last Author, Journal, and Title. Most Recent is PubMed's default sorting and ranks citations by the time they were added to the MEDLINE database. The Relevance sort uses PubMed's internal algorithm to assign weight to citations depending on the frequency search terms are found and the fields in which they are found [44]. We used and evaluated the Most Recent and Relevance sorts to compare with our proposed ranking approach. The other sorts based on publication time and alphabetical orders are less likely to identify relevant citations.

3.3.2.3.2 A machine-learning approach

In 2009, Kilicoglu et al. implemented an ensemble approach combining several machine-learning classifiers (Naïve Bayes, support vector machine (SVM), and boosting) to identify scientifically rigorous studies [45]. The classifier was built on five basic features: words, MEDLINE metadata, semantic predications, relations, and UMLS concepts. In the original study, the classifier trained on 10,000 citations could achieve 82.5% precision and 84.3% recall on an unseen test set of 2000 citations. The classifier outputs the probability a citation is scientifically rigorous. We used this classifier as the baseline ranking approach.

3.3.2.3.3 Clinical research scoring approach

We propose an alternative method for ranking MEDLINE citations using three dimensions: MeSH majority, study design, and journal ranking. These dimensions attempt to capture three characteristics that are desirable for retrieved studies: relevancy, study quality, and study impact.

A PubMed document can be indexed with multiple MeSH concepts, but only a small subset are indexed as "major topic." Using the expanded MeSH concepts from the query expansion stage, we assigned a MeSH Majority score of 2.0 if one of the MeSH concepts or any of its children was tagged as a major topic. Otherwise, a MeSH score of 1.0 was applied.

We assign a Study Design (SD) score to a study based on the publication type of the retrieved document (score 4.0: Practice Guideline, Guideline, Review with Meta-Analysis; score 3.0: Randomized Controlled Trial; score 2.0: Clinical Trial, Controlled Clinical Trial, Case-Control Studies, Cohort Studies, Longitudinal Studies, Cross-Sectional Studies, Cross-Over Studies, Observational Study, Evaluation Studies, Validation Studies, Comparative Study; and score 1.0: any other types). The rationale for the SD scoring was adapted from the GRADE system [18]. If a study has multiple publication types, the maximum SD score found on the matrix is chosen. The SD score is increased with the presence of blinding methods (single-blinded method +0.1, double-blinded method +0.2) and setting (multicenter study +0.1). The adjustment for specific randomization method and setting is a simple way to resolve tiebreaker if multiple studies shared the general study design. Such values could be adjusted by the subjective rating in different systematic review projects.

Journal ranking is an estimation of the scientific quality and clinical impact of the study based on the popularity of the publishing source. We used the open-access SCImago Journal Rank (SJR), an impact factor metric, published by Scopus in 2012. The National

Library of Medicine's (NLM) journal records were mapped to Scopus' records using the journal's ISSN number, from which we retrieve the SJR metric.

Finally, the ranking score is calculated by multiplying all three metrics (ranking score= MeSHMajorScore * SD score * SJR). Since those metrics are independent, multiplication was considered to be the most appropriate method to aggregate the three metrics.

3.3.3 Evaluation

We used the gold standard described above to evaluate the query expansion and the ranking algorithms. We tested the following hypotheses: H1: the query expansion algorithm retrieves a perfect set of citations for a larger number of guideline recommendations than the PubMed expansion approach; and H2: the citation scoring approach has better recall at k than the machine-learning classifier and the standard PubMed sort mechanisms.

In addition, we compared the algorithm performance in terms of standard information retrieval metrics. For the query expansion task, we measured recall and precision. The query expansion task was aimed to maximize recall while controlling impact on precision. We define the metric "Seeding Recall" to measure the ability of finding seed studies used to generate guideline recommendations. A practice recommendation can be synthesized from one or multiple studies. In the initial literature search, finding seed studies appeared in as many recommendations as necessary to understand the scope of the problem and guide future literature search.

$$recall = \frac{number\ of\ \ relevant\ retrieved\ documents}{number\ of\ relevant\ documents} \tag{1}$$

$$precision = \frac{number\ of\ \ relevant\ retrieved\ documents}{number\ of\ retrieved\ documents} \tag{2}$$

$$Seeding\ recall = \frac{number\ of\ recommendations\ for\ which\ at\ least\ one\ relevant\ document\ is\ retrieved}{number\ of\ recommendations} \quad (3)$$

To evaluate the ranking algorithms, we used the average precision metric. For a ranked list of documents, average precision is calculated by

$Average\ Precision = \frac{1}{r}\sum_{k=1}^{r} precision(R_k)$ where r is the number of relevant documents, and R_k is the position of the kth relevant document in the ranked list.

Precision at k (precision@k) and recall at k (recall@k) are defined as follows:

$$precision(k) = precision@k\ = \frac{number\ of\ \ relevant\ documents\ in\ top\ kth\ list}{k} \quad (4)$$

$$recall(k) = recall@k\ = \frac{number\ of\ \ relevant\ documents\ in\ top\ kth\ list}{number\ of\ relevant\ documents} \quad (5)$$

To test the H1 hypothesis, we convert the data to a binary outcome. We assigned TRUE if all citations for a recommendation were retrieved, and FALSE otherwise. The chi-square statistical test was used to assess the significance of the differences. To test the H2 hypothesis, we measured recall@k in all k positions and used the Wilcoxon signed rank test to assess the significance of the differences found.

3.4. Results

From 2010 to 2014, the American College of Cardiology (ACC) published 17 guidelines about cardiovascular topics. Four of them are Focus Update releases. We excluded those releases since the development process for the Focus Updates does not include a systematic search. Five guidelines were not on the comprehensive management of a condition and were also excluded. These guidelines covered narrower subtopics of diagnosis or treatment such as Secondary Prevention, Blood Cholesterol Treatment, and Coronary Artery Bypass Graft Surgery. Although it is possible to develop filters to target those subtopics, we decided not to cover them in this research. Eight guidelines met our inclusion criteria as summarized in Table 3.2. We were able to extract 653 practice recommendations, which cited 1863 citations. Of those, we were able to find PubMed IDs (PMIDs) in 1848 citations (99.2 %). A small portion of citations such as book chapters, online resources (e.g., FDA site), and studies not indexed in MEDLINE did not have PMIDs.

The query expansion performance and comparison are summarized in Table 3.3. Overall, the query expansion algorithm achieved recall of 80.2% and seeding recall of 90.1%. In comparison with the default PubMed expansion, the algorithm improved recall

Table 3.2 - Included cardiovascular guidelines along with their recommendations and the citations used to support recommendations.

Authors	Year	Title	Recommendations	Citations w/ PMID	Citations w/o PMID
January et al. [46]	2014	Guideline for the Management of Patients With Atrial Fibrillation	62	132	1
Brott et al. [47]	2010	Guideline on the Management of Patients With Extracranial Carotid and Vertebral Artery Disease	34	70	1
Gersh et al. [48]	2011	Guideline for the diagnosis and treatment of hypertrophic cardiomyopathy	74	175	0
Yancy et al. [49]	2013	Guideline for the management of heart failure	97	317	1
O'Gara et at. [50]	2013	Guideline for the management of ST-elevation myocardial infarction	83	216	0
Fihn et al. [51]	2012	Guideline for the diagnosis and management of patients with stable ischemic heart disease	123	407	4
Hiratzka et al. [52]	2010	Guidelines for the diagnosis and management of patients with Thoracic Aortic Disease	63	156	4
Nishimura et al. [53]	2014	Guideline for the Management of Patients With Valvular Heart Disease	117	375	4

Table 3.3 - Comparison between PubMed expansion and MeSH expansion algorithm.

	Default PubMed Expansion	MeSH Expansion	Mean Difference
Recall % (SD)	51.5 (35.5)	80.2 (5.1)	28.7 (31.7)
Seeding recall % (SD)	63.5 (31.6)	90.1 (6.1)	26.5 (29.6)
Precision % (SD)	0.6 (0.5)	0.4 (0.5)	-0.2 (0.3)
Recommendations for which all citations were found %	37.2	64.5	27.3

by 28.7% and seeding recall by 26.5% with a 0.2% drop in precision. The ability to find seed studies (seeding recall) improved by 26.6%. Our query expansion algorithm could find all citations for more guideline recommendations than the default PubMed expansion (64.5% vs. 37.2%, p<0.0001).

For citation ranking, the clinical research scoring approach had the best average precision of 7% compared to 2.1% machine-learning classifier, 0.9% PubMed's sort by relevance, 0.5% PubMed's sort by Most Recent (Table 3.4). Similarly, the scoring approach had the highest average recall@k, improved 4.2% over the machine-learning classifier (66.2% vs 62%, p<0.001), 14.8% over PubMed's sort by Relevance (66.2% vs. 51.4%, p<0.001), and 21.1% over PubMed's sort by Most Recent (66.2% vs. 45.1%,

Table 3.4 - Performance comparison among various ranking approaches.

	PubMed's sorting by Most Recent	PubMed's sorting by Relevance	Kilicoglu's Machine-learning classifier	Clinical research scoring approach
Average precision % (SD)	0.5 (0.7)	0.9 (1.0)	2.1(1.7)	7.0 (4.8)
Recall@k % (SD)	45.1 (26.2)	51.4 (23.5)	62 (18.6)	66.2 (15.6)

p<0.001). In Figure 3.3, we illustrate the recall@k at various kth position in the ranked list. Overall, PubMed's sorts essentially performed worse than machine-learning classifier and the scoring approach. The curve of the scoring approach outperformed the machine-learning curve for most of the guidelines, especially at lower levels of k. However, the difference was significant in some guidelines (Hypertrophic Cardiomyopathy, Heart Failure, Thoracic Aortic Disease, Valvular Heart Disease), while only nonsignificantly improved in other guidelines.

3.5. Discussion

3.5.1. Significance

We developed and evaluated an automated approach to retrieve relevant and high-quality citations from PubMed. The approach can be used to assist the development of

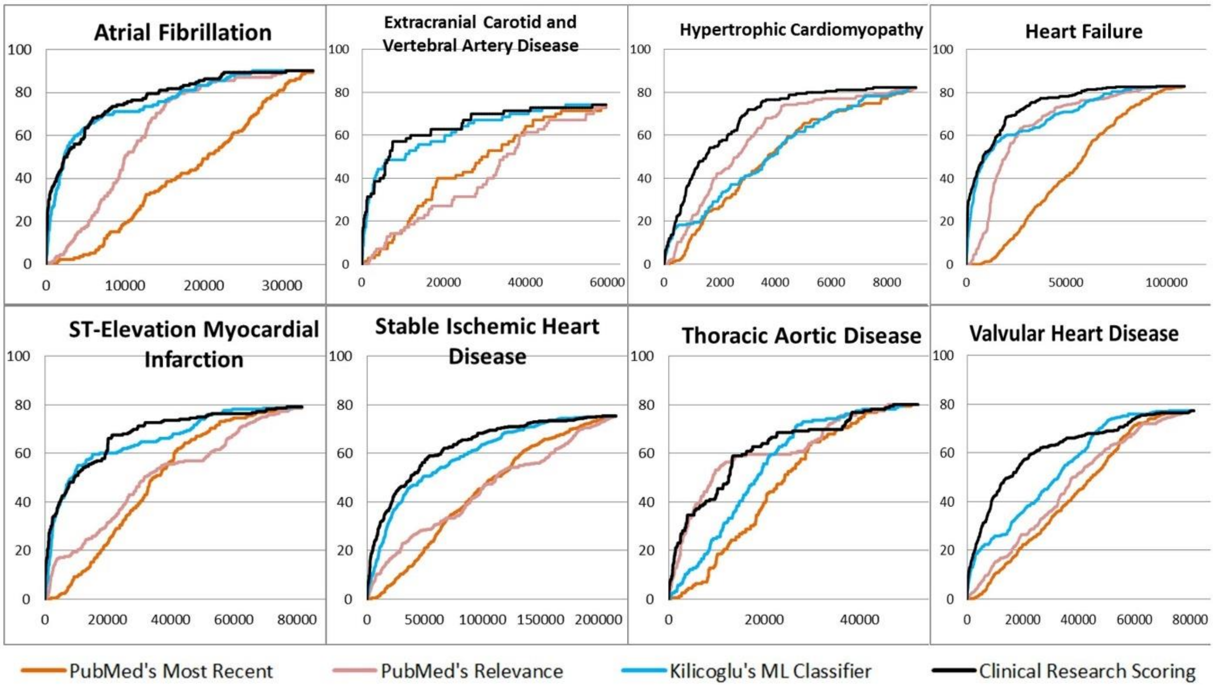


Figure 3.3 - Recall@k at various kth positions of 4 ranking methods in each of the cardiology guideline.

clinical guidelines and systematic reviews. The results showed that our proposed method outperformed the default PubMed query expansion in terms of recall (80.2% vs. 51.5%) and seeding recall (90% vs. 63.5%), with a nonsignificant loss in precision (0.6% vs. 0.4%; p=0.09). In addition, the method could find all citations for a larger number of guideline recommendations than the PubMed expansion (64.5% vs. 37.2%, p<0.0001). The results reflect the goal of systematic search, that is, to maximize recall to identify all relevant studies while controlling impact on precision to keep the results manageable.

We experienced a stable recall variance on all guideline topics (stddev = 5.1), however, the improving effect variance was high (stddev = 31.7). A subsequent analysis showed that three topics “atrial fibrillation”, “hypertrophic cardiomyopathy”, and “heart failure” had no improvement on recall, partially because the baseline PubMed expansion achieved good performance (avg recall 85.1%). All other topics had improvements in recall. The greatest improvement was seen in the topic “Extracranial Carotid and Vertebral Artery Disease”, in which PubMed expansion did not perform well. The query expansion algorithm was able to find supporting MeSH terms such as “Carotid Artery Diseases”, “Vertebrobasilar Insufficiency”, ” Brain Ischemia”, and “Cerebrovascular Disorders”, and recall was improved by 70%.

The system achieved precision of 0.6% versus 0.8% with PubMed expansion. Therefore, we deem the system’s precision performance was acceptable and comparable with existing methods. Achieving good precision is difficult and secondary for systematic search. In fact, the manual search approach achieved precision below 1% [54-56]. The poor precision can be attributed to the main goal of systematic search, which is to be exhaustive. Therefore, the queries were generally designed to be able to capture all potentially relevant candidates. In addition, some systematic reviews had specific inclusion/exclusion criteria which are not easily represented in the search queries without risking loss of recall. Further efforts to improve precision relate to previous works on document classification, in which training data to predict the inclusion/exclusion patterns is required [12, 17].

The citation ranking method proposed in this research used a simple light-weight approach that is independent of training data. Furthermore, the proposed approach improved ranking performance of the standard PubMed's ranking by "most recent" (average precision +6.5%, recall@k +21.1%, p<0.001), PubMed's ranking by "relevance" (average precision +6.1%, recall@k +14.8%, p<0.001), and the general purpose machine-learning classifier (average precision +4.9%, recall@k +4.2%, p<0.001).

3.5.2. Implications

In the development of systematic reviews, manual search is considered the state-of-the-art approach, but it does not guarantee perfect recall. The quality of the search is essentially impacted by skills, experience, and domain knowledge of searchers on the review topics. A common approach to improve recall is to gather results from multiple sources either from different search strategies or from domain experts. The American College of Cardiology Foundation (ACCF) recommends clinicians to perform their own search along with systematic search by skilled librarians [11]. Our method is not intended to completely replace the manual process. However, it can serve as starting point or as a reference list to augment the manual search approach. For example, taking our dataset, if reviewers screen the top 100 citations retrieved by our system, they would be able to find 16.2% of the citations included in the guidelines and seed citations for 24.4% of the guideline recommendations. Another potential approach is to use citation tracking by examining articles that cite or are cited by seed citations. The seeding recall metric used in our study provides a measure of algorithm performance in this respect. The system was able to find the seed studies for 90% of the guideline recommendations.

Ranking studies by relevancy and scientific rigor might be useful to help prioritize early stages in the development of systematic reviews. A good ranking mechanism increases the odds of finding relevant studies with less effort. Previous studies on work prioritization [22, 23] favor using machine-learning methods, which use previous manual screening as labeled data to train classifiers. However, in systematic search, new questions are often raised that have insufficient historical data to train a competent machine-learning model. As a result, searchers often rely on standard functionalities of search engines, or ML classifiers that were trained on broad topics. Our experiments showed that standard ranking methods of biomedical search engines and a general purpose ML classifier can be further improved using heuristics such as MeSH majority, research design, and journal ranking. These heuristics are independent of the training data and not specific to any particular guideline topic or domain.

This study focuses on cardiovascular guideline as our domain of interest; however, the proposed techniques are applicable to a literature search in general. First, the system employed reusable expansion techniques to identify relevant MeSH concepts (concept-mapping, statistical, and MeSH Tree traversing) that are not specific to cardiology and should work well in any other domain related to treatment. For areas other than diseases (e.g., procedures), the algorithm could be adapted by using different semantic types. For example, a review topic focused on an intervention procedure could use semantic type “Therapeutic or Preventive Procedure”. Secondly, our ranking approach was based on three factors: MeSH Majority, Study Design, and Journal Ranking. MeSH Majority and Journal Ranking information can always be found in MEDLINE and Scopus. The assignment of the study design (SD) score is adapted from the GRADE approach [18], which is widely used in the assessment of evidence quality independent of clinical domain.

3.5.3. Limitations

This study has five main limitations. First, our gold standard consists of eight guidelines, which limits the generalizability of our findings. However, the guidelines we selected represent a broad coverage in the important field of cardiovascular diseases. In 2010, ACCF/AHA published a methodology manual that mandated all practice recommendations grade A and B to be accompanied with citations to the evidence sources. This practice will help expand the size and breadth of gold standards in future studies. Second, our research was limited to guidelines on the treatment of cardiovascular diseases, so it is unknown whether the results generalize to other domains. Yet our approach did not use any methods that were specific to cardiovascular diseases, so it is expected that the methods will generalize to other domains and topics. Third, our query expansion algorithm uses an ad-hoc threshold (5000) for triggering parent concept expansion. The selection of this threshold was somewhat arbitrary and can be improved further based on heuristics such as the descriptive statistics of retrieved documents. Fourth, our system achieved low precision that is common and secondary in systematic search. Previous techniques based on automated and semiautomated document classification to support citation screening could be used to improve precision. For example, Cohen et al. showed that optimize machine-learning algorithm parameters improved the classification performance [12, 23]. However, their approach needs to customize classifier for specific drug questions, which made the solution difficult to scale given the diversity of clinical systematic review queries. Last, we did not retrain Kilicoglu's classifier with our dataset and use the classifier developed in their original research [45]. In the early stage of literature search, the lack of labeled data made it difficult to train a competent machine-learning classifier.

3.5.4. Future Studies

Areas that warrant further investigation include improving overall precision using automated and semiautomated document classification techniques; expanding the gold standard beyond cardiovascular topics; improving the method to distinguish diagnosis and treatment topics; and applying the method to other types of systematic review, such as Cochrane systematic reviews, and drug effectiveness reviews.

3.6. Conclusion

We present informatics solutions to improve the retrieval performance of high-quality studies to support the development of clinical guidelines in the cardiovascular domain. Overall, our methods are unsupervised and integrated over a widely used biomedical search engine (PubMed). The methods showed improved recall over standard PubMed's query expansion and rankings and a general-purpose machine-learning classifier. The proposed approach could be used to aid the systematic search and screening process in the development of systematic reviews and clinical guidelines.

3.7 Acknowledgements

This project was supported in part by grants 1R01LM011416-02 and 4R00LM011389-02 from the National Library of Medicine.

3.8 References

[1] M. Ware, M. Mabe, An Overview of Scientific and Scholarly Journal Publishing, The STM Report, 2009.

[2] Statistical Reports on MEDLINE®/PubMed® Baseline Data, <https://www.nlm.nih.gov/bsd/licensee/baselinestats.html>.

[3] Trends, Charts, and Maps: ClinicalTrials.gov, <http://clinicaltrial.gov/ct2/resources/trends>.

[4] A.G. Fraser, F.D. Dunstan, On the impossibility of being expert, BMJ (Clinical Research ed). 341 (2010) c6815. PubMed PMID: Medline:21156739. English.

[5] K. Davies, J. Harrison, The information-seeking behaviour of doctors: a review of the evidence, Health Inform. Libr. J. 24 (2) (2007) 78–94. June. PubMed PMID: 17584211.

[6] G. Del Fiol, T.E. Workman, P.N. Gorman, Clinical questions raised by clinicians at the point of care: a systematic review, JAMA Int. Med. (2014).

[7] R. Smith, Strategies for coping with information overload, BMJ 341 (2010) c7126. PubMed PMID: 21159764.

[8] J.S. Burgers, R. Grol, N.S. Klazinga, M. Makela, J. Zaat, A. Collaboration, Towards evidence-based clinical practice: an international survey of 18 clinical guideline programs, Int. J. Qual. Health Care: J. Int. Soc. Qual. Health Care/ ISQua 15 (1) (2003) 31–45. February. PubMed PMID: 12630799.

[9] H.J. Schünemann, A. Fretheim, A.D. Oxman, Research WACoH. Improving the use of research evidence in guideline development: 1. Guidelines for guidelines, Health Res. Policy Syst. 4 (13) (2006) 1–6.

[10] T. Turner, M. Misso, C. Harris, S. Green, Development of evidence-based clinical practice guidelines (CPGs): comparing approaches, Implement. Sci.: IS 3 (45) (2008) 1–8.

[11] ACCF/AHA TaskForce on Practice Guidelines, Methodology Manual and Policies From the ACCF/AHA Task Force on Practice Guidelines 2010, <http://assets.cardiosource.com/Methodology_Manual_for_ACC_AHA_Writing_Committees.pdf>.

[12] A.M. Cohen, W.R. Hersh, K. Peterson, P.Y. Yen, Reducing workload in systematic review preparation using automated citation classification, J. Am. Med. Inform. Assoc. 13 (2) (2006) 206–219. March–April. PubMed PMID:16357352. PubMed Central PMCID: 1447545.

[13] B.C. Wallace, T.A. Trikalinos, J. Lau, C. Brodley, C.H. Schmid, Semi-automated screening of biomedical citations for systematic reviews, BMC Bioinformatics 11 (2010) 55. PubMed PMID: 20102628. Pubmed Central PMCID: 2824679.

[14] S. Jonnalagadda, D. Petitti, A new iterative method to reduce workload in systematic review process, Int. J. Comput. Biol. Drug Des. 6 (1–2) (2013) 5–17. PubMed PMID: 23428470. Pubmed Central PMCID: 3787693.

[15] M. Fiszman, B.E. Bray, D. Shin, H. Kilicoglu, G.C. Bennett, O. Bodenreider, et al., Combining relevance assignment with quality of the evidence to support guideline development, Stud. Health Technol. Inform. 160 (Pt 1) (2010) 709–713. PubMed PMID: 20841778.

[16] M. Fiszman, E. Ortiz, B.E. Bray, T.C. Rindflesch, Semantic processing to support clinical guideline development, in: AMIA Annual Symposium proceedings/AMIA Symposium AMIA Symposium, 2008, pp. 187–191. PubMed PMID: 18999127. Pubmed Central PMCID: 2656081.

[17] A.M. Cohen, K. Ambert, M. McDonagh, A prospective evaluation of an automated classification system to support evidence-based medicine and systematic review, in: AMIA Annual Symposium Proceedings/AMIA Symposium AMIA Symposium, 2010, pp. 121–125. PubMed PMID: 21346953. Pubmed Central PMCID: 3041348.

[18] J.P. Higgins, S. Green, Cochrane Handbook for Systematic Reviews of Interventions, Wiley Online Library, 2008.

[19] M. Crespo Azcarate, J. Mata Vazquez, M. Mana Lopez, Improving image retrieval effectiveness via query expansion using MeSH hierarchical structure, J. Am. Med. Inform. Assoc. 20 (6) (2013) 1014–1020. PubMed PMID: Medline:22952301. English.

[20] Z. Lu, W. Kim, W.J. Wilbur, Evaluation of query expansion using MeSH in PubMed, Inf. Retr. Boston. 12 (1) (2009) 69–80.

[21] D. Bui, D. Redd, T. Rindflesch, Q. Zeng-Treitler, An Ensemble Approach for Expanding Queries, DTIC Document, 2012.

[22] A.M. Cohen, Optimizing feature representation for automated systematic review work prioritization, in: AMIA Annual Symposium Proceedings: American Medical Informatics Association, 2008.

[23] A.M. Cohen, K. Ambert, M. McDonagh, Cross-topic learning for work prioritization in systematic review creation and update, J. Am. Med. Inform. Assoc. 16 (5) (2009) 690–704.

[24] S. Moosavinasab, M. Rastegar-Mojarad, H. Liu, S.R. Jonnalagadda, Towards transforming expert-based content to evidence-based content, AMIA Summits Trans. Sci. Proc. 2014 (2014) 83.

[25] S.R. Jonnalagadda, S. Moosavinasab, D. Li, M.D. Abel, C.G. Chute, H. Liu, Prioritizing journals relevant to a topic for addressing clinicians' information needs, in: 2013 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2013, IEEE.

[26] NCBI Batch Citation Matcher, <http://www.ncbi.nlm.nih.gov/pubmed/batchcitmatch>.

[27] Y. Ahmad, G.Y. Lip, S. Apostolakis, New oral anticoagulants for stroke prevention in atrial fibrillation: impact of gender, heart failure, diabetes mellitus and paroxysmal atrial fibrillation, Expert Rev. Cardiovasc. Ther. 10 (12) (2012) 1471–1480. December. PubMed PMID: 23253272.

[28] C.E. Chiang, L. Naditch-Brule, J. Murin, M. Goethals, H. Inoue, J. O'Neill, et al., Distribution and risk profile of paroxysmal, persistent, and permanent atrial fibrillation in routine clinical practice: insight from the real-life global survey evaluating patients with atrial fibrillation international registry, Circulat. Arrhythmia Electrophysiol. 5 (4) (2012) 632–639. August 1. PubMed PMID: 22787011.

[29] G. Flaker, M. Ezekowitz, S. Yusuf, L. Wallentin, H. Noack, M. Brueckmann, et al., Efficacy and safety of dabigatran compared to warfarin in patients with paroxysmal, persistent, and permanent atrial fibrillation: results from the RELY (Randomized Evaluation of Long-Term Anticoagulation Therapy) study, J. Am. Coll. Cardiol. 59 (9) (2012) 854–855. February 28. PubMed PMID: 22361407.

[30] S.H. Hohnloser, G.Z. Duray, U. Baber, J.L. Halperin, Prevention of stroke in patients with atrial fibrillation: current strategies and future directions, Eur. Heart J. Suppl. 10 (suppl H) (2008) H4–H10.

[31] R. Farshi, D. Kistner, J.S. Sarma, J.A. Longmate, B.N. Singh, Ventricular rate control in chronic atrial fibrillation during daily activity and programmed exercise: a

crossover open-label study of five drug regimens, J. Am. Coll. Cardiol. 33 (2) (1999) 304–310. February. PubMed PMID: 9973007.

[32] J.S. Steinberg, R.J. Katz, G.B. Bren, L.A. Buff, P.J. Varghese, Efficacy of oral diltiazem to control ventricular response in chronic atrial fibrillation at rest and during exercise, J. Am. Coll. Cardiol. 9 (2) (1987) 405–411.

[33] B. Olshansky, L.E. Rosenfeld, A.L. Warner, A.J. Solomon, G. O'Neill, A. Sharma, et al., The Atrial Fibrillation Follow-up Investigation of Rhythm Management (AFFIRM) study: approaches to control rate in atrial fibrillation, J. Am. Coll. Cardiol. 43 (7) (2004) 1201–1208. April 7. PubMed PMID: 15063430.

[34] J. Abrams, J. Allen, D. Allin, J. Anderson, S. Anderson, L. Blanski, et al., Efficacy and safety of esmolol vs. propranolol in the treatment of supraventricular tachyarrhythmias: a multicenter double-blind clinical trial, Am. Heart J. 110 (5) (1985) 913–922. November. PubMed PMID: 3904379.

[35] K.A. Ellenbogen, V.C. Dias, V.J. Plumb, J.T. Heywood, D.M. Mirvis, A placebocontrolled trial of continuous intravenous diltiazem infusion for 24-hour heart rate control during atrial fibrillation and atrial flutter: a multicenter study, J. Am. Coll. Cardiol. 18 (4) (1991) 891–897. October. PubMed PMID: 1894861.

[36] C.-W. Siu, C.-P. Lau, W.-L. Lee, K.-F. Lam, H.-F. Tse, Intravenous diltiazem is superior to intravenous amiodarone or digoxin for achieving ventricular rate control in patients with acute uncomplicated atrial fibrillation, Crit. Care Med. 37 (7) (2009) 2174–2179.

[37] E.V. Platia, E.L. Michelson, J.K. Porterfield, G. Das, Esmolol versus verapamil in the acute treatment of atrial fibrillation or atrial flutter, Am. J. Cardiol. 63 (13) (1989) 925–929. April 15. PubMed PMID: 2564725.

[38] R.B. Haynes, K.A. McKibbon, N.L. Wilczynski, S.D. Walter, S.R. Werre, Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey, BMJ 330 (7501) (2005) 1179.

[39] A.R. Aronson, Metamap: Mapping Text to the UMLS Metathesaurus, NLM, NIH, DHHS, Bethesda, MD, 2006, pp. 1–26.

[40] E. Sayers, Entrez Programming Utilities Help, <http://www.ncbi.nlm.nih.gov/books/NBK25499. 2009>.

[41] A.R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, in: Proceedings/AMIA Annual Symposium AMIA Symposium, 2001, pp. 17–2. PubMed PMID: 11825149. Pubmed Central PMCID: 2243666.

[42] National Library of Medicine, UMLS Reference Manual, 2009, <http://www.ncbi.nlm.nih.gov/books/NBK9685/>.

[43] National Library of Medicine, Medical Subject Headings, 2014, <http://www.nlm.nih.gov/mesh/filelist.html>.

[44] National Library of Medicine, PubMed Help, 2015, <http://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.Sorting_your_search_results>.

[45] H. Kilicoglu, D. Demner-Fushman, T.C. Rindflesch, N.L. Wilczynski, R.B. Haynes, Towards automatic recognition of scientifically rigorous clinical research evidence, J. Am. Med. Inform. Assoc. 16 (1) (2009) 25–31. January–February. PubMed PMID: 18952929. Pubmed Central PMCID: 2605595.

[46] C.T. January, L.S. Wann, J.S. Alpert, H. Calkins, J.C. Cleveland Jr., J.E. Cigarroa, et al., 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society, J. Am. Coll. Cardiol. (2014). March 28. PubMed PMID: 24685669.

[47] T.G. Brott, J.L. Halperin, S. Abbara, J.M. Bacharach, J.D. Barr, R.L. Bush, et al., 2011 ASA/ACCF/AHA/AANN/AANS/ACR/ASNR/CNS/SAIP/SCAI/SIR/SNIS/SVM/S VS guideline on the management of patients with extracranial carotid and vertebral artery disease. A report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines, and the American Stroke Association, American Association of Neuroscience Nurses, American Association of Neurological Surgeons, American College of Radiology, American Society of Neuroradiology, Congress of Neurological Surgeons, Society of Atherosclerosis Imaging and Prevention, Society for Cardiovascular Angiography and Interventions, Society of Interventional Radiology, Society of NeuroInterventional Surgery, Society for Vascular Medicine, and Society for Vascular Surgery, Circulation 124 (4) (2011) e54–e130. July 26 PubMed PMID: 21282504.

[48] B.J. Gersh, B.J. Maron, R.O. Bonow, J.A. Dearani, M.A. Fifer, M.S. Link, et al., 2011 ACCF/AHA guideline for the diagnosis and treatment of hypertrophic cardiomyopathy: a report of the American College of Cardiology Foundation/ American Heart Association Task Force on Practice Guidelines, Circulation 124 (24) (2011) e783–e831. December 13 PubMed PMID: 22068434.

[49] C.W. Yancy, M. Jessup, B. Bozkurt, J. Butler, D.E. Casey Jr., M.H. Drazner, et al., 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines, J. Am. Coll. Cardiol. 62 (16) (2013) e147–e239. October 15 PubMed PMID: 23747642.

[50] P.T. O'Gara, F.G. Kushner, D.D. Ascheim, D.E. Casey Jr., M.K. Chung, J.A. de Lemos, et al., 2013 ACCF/AHA guideline for the management of STelevation myocardial infarction: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines, Circulation 127 (4) (2013) e362–e425. January 29 PubMed PMID: 23247304.

[51] S.D. Fihn, J.M. Gardin, J. Abrams, K. Berra, J.C. Blankenship, A.P. Dallas, et al., 2012 ACCF/AHA/ACP/AATS/PCNA/SCAI/STS guideline for the diagnosis and management of patients with stable ischemic heart disease: executive summary: a report of the American College of Cardiology Foundation/American Heart Association task force on practice guidelines, and the American College of Physicians, American Association for Thoracic Surgery, Preventive Cardiovascular Nurses Association, Society for Cardiovascular Angiography and Interventions, and Society of Thoracic Surgeons, Circulation 126 (25) (2012) 3097–3137. December 18 PubMed PMID: 23166210.

[52] L.F. Hiratzka, G.L. Bakris, J.A. Beckman, R.M. Bersin, V.F. Carr, D.E. Casey Jr., et al., 2010 ACCF/AHA/AATS/ACR/ASA/SCA/SCAI/SIR/STS/SVM guidelines for the diagnosis and management of patients with Thoracic Aortic Disease: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines, American Association for Thoracic Surgery, American College of Radiology, American Stroke Association, Society of Cardiovascular Anesthesiologists, Society for Cardiovascular Angiography and Interventions, Society of Interventional Radiology, Society of Thoracic Surgeons, and Society for Vascular Medicine, Circulation 121 (13) (2010) e266–e369. April 6 PubMed PMID: 20233780.

[53] R.A. Nishimura, C.M. Otto, R.O. Bonow, B.A. Carabello, J.P. Erwin 3rd, R.A. Guyton, et al., 2014 AHA/ACC Guideline for the Management of Patients With Valvular Heart Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines, J. Am. Coll. Cardiol. 63 (22) (2014) e57–e185. June 10 PubMed PMID: 24603191.

[54] F. Wiesbauer, H. Domanovits, O. Schlager, B. Wildner, M. Schillinger, H. Blessberger, Perioperative Beta-Blockers for Preventing Surgery Related Mortality and Morbidity, The Cochrane Library, 2003.

[55] L.G. De Lima, H. Saconato, Á.N. Atallah, E.M. da Silva, Beta-Blockers for Preventing Stroke Recurrence, The Cochrane Library, 2014.

[56] E. Lopez-Briz, V. Ruiz Garcia, J.B. Cabello, S. Bort-Marti, R. Carbonell Sanchis, A. Burls, Heparin versus 0.9% sodium chloride intermittent flushing for prevention of occlusion in central venous catheters in adults, The Cochrane Database of Systematic Reviews, vol. 10, 2014.

CHAPTER 4

PDF TEXT CLASSIFICATION TO LEVERAGE INFORMATION EXTRACTION FROM PUBLICATION REPORTS

Duy Duc An Bui, Guilherme Del Fiol, Siddhartha Jonnalagadda

Journal of Biomedical Informatics[2]

4.1 Abstract

Data extraction from original study reports is a time-consuming, error-prone process in systematic review development. Information extraction (IE) systems have the potential to assist humans in the extraction task, however majority of IE systems were not designed to work on Portable Document Format (PDF) document, an important and common extraction source for systematic review. In a PDF document, narrative content is often mixed with publication metadata or semistructured text, which add challenges to the underlining natural language processing algorithm. Our goal is to categorize PDF texts for strategic use by IE systems.

We used an open-source tool to extract raw texts from a PDF document and developed a text classification algorithm that follows a multipass sieve framework to automatically classify PDF text snippets (for brevity, texts) into TITLE, ABSTRACT, BODYTEXT, SEMISTRUCTURE, and METADATA categories. To validate the algorithm, we

[2] D.D.A. Bui, G. Del Fiol, S. Jonnalagadda, PDF text classification to leverage information extraction from publication reports, J. Biomed. Inform. (2016).

developed a gold standard of PDF reports that were included in the development of previous systematic reviews by the Cochrane Collaboration. In a two-step procedure, we evaluated (1) classification performance and compared it with machine-learning classifier, and (2) the effects of the algorithm on an IE system that extracts clinical outcome mentions.

The multipass sieve algorithm achieved an accuracy of 92.6%, which is 11.9% ($p<0.001$) higher than that of the machine-learning classifier. F-measure improvements were observed in the classification of TITLE (+2.1%), ABSTRACT (+19.8%), BODYTEXT (+4.7%), SEMISTRUCTURE (+37.8%), and METADATA (+23.7%). In addition, use of the algorithm to filter semistructured texts and publication metadata improved performance of the outcome extraction system (F-measure +4.3%, $p<0.001$). It also reduced of number of sentences to be processed by 46.5%, which reduced processing time without causing performance loss.

The rule-based multipass sieve framework can be used effectively in categorizing texts extracted from PDF documents. Text classification is an important prerequisite step to leverage information extraction from PDF documents.

4.2. Introduction

Systematic reviews (SRs) are important expert-synthesized information sources to enable evidence-based medicine practice [1]. However, the production and updating of SRs are often costly, slow, and unable to keep pace with the rapid growth of the biomedical literature [2, 3]. The total expenditure of the Cochrane Collaboration, a prominent SR development organization, in fiscal year 2011 was $2.4 million [4], and that number increased to $3.9 million in 2013 [5]. Citations indexed in PubMed have grown from 4 million (pre-1975) to 22 million today [3]. It takes 2.5 to 6.5 years for a primary study publication to be included and published in a new SR [6]. About 23% of SRs have not been updated with new evidence within 2 years after the first publication [6], and many clinical questions are not addressed in existing SRs [7]. As a result, SRs quickly become outdated and suboptimal for patient care. This is partially because the SR process involves many labor-intensive manual steps, which face human limitations such as limited time and resources, and human inconsistency and errors. This situation highlights the need for investigating computer techniques to aid humans in SR development.

The SR process involves a series of scientifically rigorous steps, such as citation searching, abstract screening, full-text screening, data extraction, and article appraisal. Data extraction to generate evidence summaries is one of the most important and time-consuming steps in SR development [8]. Natural language processing (NLP) research in the past decade has investigated techniques to extract study characteristics from biomedical publications [9-18]. Those techniques have the potential to optimize the manual data extraction process; however, there are research gaps that have not been filled. One of the gaps that we choose to address in the present study concerns the heterogeneity of the digital document format. Present information extraction (IE) studies select sources of extraction from MEDLINE abstracts, PubMed Central (PMC) archives, and journal websites [11, 16, 18] available in HyperText Markup Language (HTML) format. However, the data extraction practice requires that the source of extraction be the original full-text study reports [8], and the most common format for full-text reports is the Portable Document Format (PDF).

NLP research on PDF documents faces several challenges. In a PDF document, narrative content is often mixed with publication metadata (header, footer, author information, journal information, etc) and semistructured text (tables or figures). Publication metadata are often not relevant to the extraction goal and can add noise to the NLP system. Semistructured text can contain relevant information, but it does not adhere to grammatical rules and requires different extraction strategies than narrative text. Therefore, categorizing the text snippets (texts) in a PDF document is a necessary first step to design an optimal extraction strategy.

There have been studies on document structure recognition that sought to recover the logical structure from PDF documents. Commonly used approaches were machine-learning [19-22] and rule-based or heuristics [23-28]. A rich number of PDF features have been used, including text pattern, format, spatial coordinates, and page boundary. Those methods used different approaches to recognize the PDF structure, and their performances also varied. None of the previous studies have been evaluated with practical real-world applications; therefore, the usefulness of PDF structure recognition for IE or text mining has not been validated. In the present research, we present an alternative approach to recognizing PDF structure. We used an open-source PDF library (PDFBox) to extract raw texts and metadata from PDF files and applied the novel rule-based multipass sieve approach for text classification. The algorithm was evaluated against PDF reports used in the development of Cochrane reviews. A two-step evaluation was performed. First, classification performance was measured and compared against a machine-learning classifier. Then, the algorithm's impact on an IE system was evaluated.

4.3. Materials and Methods

Our study has three main parts: (a) development of a gold standard for PDF text classification and outcome extraction task, (b) development of a multipass sieve algorithm for PDF text classification, and (c.1) evaluation of the performance of the multipass sieve algorithm and comparison with a machine-learning approach, and (c.2) evaluation of the impact of PDF text classification on IE performance. The system architecture and the study design are summarized in Figure 4.1.

4.3.1. Gold Standard

Cochrane reviews on the subject "heart and circulation" that were published after October 2014 were retrieved from the Cochrane Library web interface. In each review, we located the included primary studies and searched for the PDF reports. To narrow the research focus to clinical trials and facilitate the IE task, we excluded nonrandomized control trials and studies that had been reported in multiple publications.

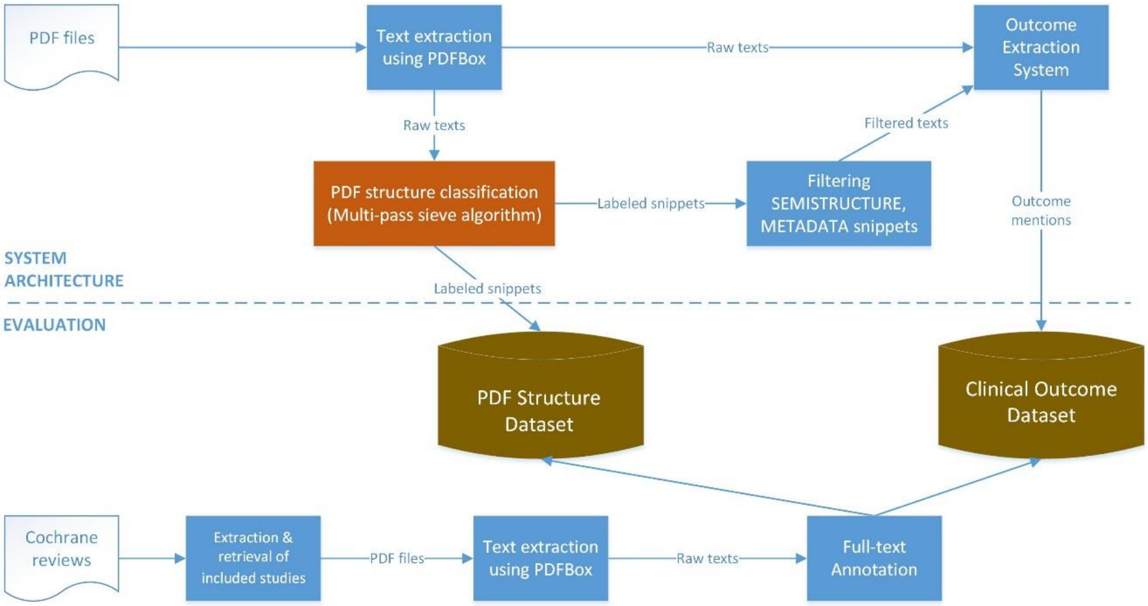

PDF files
Text extraction using PDFBox
Raw texts
Outcome Extraction System
Raw texts
Filtered texts
PDF structure classification (Multi-pass sieve algorithm)
Labeled snippets
Filtering SEMISTRUCTURE, METADATA snippets
Outcome mentions
SYSTEM ARCHITECTURE
Labeled snippets
EVALUATION
PDF Structure Dataset
Clinical Outcome Dataset
Cochrane reviews
Extraction & retrieval of included studies
PDF files
Text extraction using PDFBox
Raw texts
Full-text Annotation

Figure 4.1 - The system architecture and the study design.

To build the text classification corpus, we used the PDFBox tool to extract raw texts from the PDF reports. Texts extracted by using PDFBox are similar to the characteristics of text copied and pasted directly from a PDF reader. The principal structure is lost and texts are broken into multiple lines of text snippets. We then used the GATE annotation tool [29] to annotate text snippets into five categories: TITLE, ABSTRACT, BODYTEXT, SEMISTRUCTURE, and METADATA. The METADATA labels were assigned to text snippets related to citation information, such as authorship, journal name, header/footer, and references. The SEMISTRUCTURE labels were assigned to text snippets that consisted of tables or figures. The TITLE and ABSTRACT labels were assigned to snippets that were the title and abstract of the document, respectively. The remainder of the text snippets were assigned the BODYTEXT label.

In the next stage, we developed the gold standard for IE of study outcomes. Study outcomes are the extracted data elements commonly reported in the evidence summary of Cochrane reviews. They are the measurements used to assess a study hypothesis. We started with the outcome values reported in the Cochrane evidence summary and extended by reviewing full-text manuscripts to validate and supplement the gold standard. Specifically, we looked for synonymous mentions (e.g., abbreviations), and outcome mentions such as adverse events and side effects that were not completely reported. High-level measurements (e.g., "antihypertensive efficacy") were replaced with specific measures that described how a measurement was operationalized (e.g., "blood pressure"). This effort served to correctly estimate recall of the system, because sometimes the evidence summary table reported only measures relevant to the review questions.

4.3.2. The MultiPass Sieve Algorithm

We followed a multipass sieve framework to classify text snippets. In previous studies, the multipass sieve framework has been successfully applied to solve co-reference resolution problems [30, 31]. The framework favors applying multiple independent sieves to solve the problem rather than using a singlepass model. The multipass sieve model includes a succession of multiple independent sieves, and each contains a set of rules to target specific data elements. To implement effective rules, we used standard features of PDFBox to extract raw texts, font type, font size, page boundary, and paragraph boundary. In addition, we used the named-entity recognition feature of the Stanford NLP tool to identify whether text snippets contain PERSON, ORGANIZATION, or LOCATION entities. MEDLINE resources such as abstract, title, and author names were also used. To facilitate rule maintenance, we implemented the following sieve configurations: begin condition, directionality, pass condition, stop condition, and repetition. The configurations are varied in different sieves and optimized to recognize specific types of text snippet.

If one of the begin condition rules is met, the sieve triggers the discovery process. Many of the begin conditions are dictionary matching rules such as looking up a prebuilt section-heading collection. For instance, to identify author metadata, we look for snippets having patterns such as “correspondence to:”, “author affiliations”, and “financial disclosures”.

Directionality defines the direction to which the sieve moves to compute the next snippets in the document. Typically, the directionality for specific labels is statically configured to either the UP or DOWN direction. For TABLE and FIGURE labels, the directionalities are dynamically configured. The sieve chooses direction dynamically based on an examination of the surrounding contexts.

The sieve assigns the target label to the snippet if the pass condition is met. The sieve stops the discovery process if the condition is met. Stop condition prevents the sieve from aggressively expanding to other sections. A frequently used stop condition is the first failure of the Pass Condition, but there are other rules, such as matching common content headings and maximum page number limit.

Repetition defines the number of times the sieve is repeated. The sieve can be repeated one or many times. A difficult case involves the sieve recognition of ABSTRACT snippets. Abstract texts are sometimes divided into two clusters of texts. Therefore, we configured the sieve repetition of two times to capture those clusters.

Table 4.1 describes the full multipass sieve algorithm. The recognition of METADATA labels is subdivided into recognition of HEADER, FOOTER, KEYWORD, AUTHOR, JOURNAL, and REFERENCE. We built FIGURE and TABLE sieves to recognize SEMISTRUCTURE text snippets. Last, all unlabeled snippets were assigned to BODYTEXT labels.

4.3.3. Machine-learning Classifier

We implemented a baseline machine-learning approach to compare with the multipass sieve algorithm. To train the ML classifier, we used a Support Vector Machine algorithm with the Sequential Minimal Optimization (SMO) implementation. The Support Vector Machine SMO has been commonly and successfully applied in many text classification studies [32-34]. The SMO algorithm was implemented with the WEKA data mining software with a linear kernel; default values were used for all other SMO parameters. We used eight features: length of the snippet, paragraph number, page number, whether the snippet was in the document’s main font, font size, whether the snippet was contained in

Table 4.1 - The full description of multipass sieve algorithm (R=Rule).

Target Element	Begin Condition	Pass Condition	Stop Condition	Direction/ Repetition
HEADER	**R1:** Begin of page AND **R2:** Repeat more than 1 times.	**R3:** Same paragraph with previous line.	**R4:** Fail Pass Condition **R5:** Match common section headings	Direction: DOWN Repeat: UNLIMITED
FOOTER	**R5:** End of page AND **R2**	**R3**	**R4** **R5**	Direction: UP Repeat: UNLIMITED
KEYWORD	**R6:** Match keywords headings	**R3**	**R4** **R5**	Direction: DOWN Repeat: 1
TABLE	**R7:** Match table common headings. **R7.1:** Treat the following lines in the paragraph as the Table captions.	**R3** **R8:** Same font with previous line AND **R9.1:** NOT contain document main font **R10:** Contain sequence of number pattern (e.g., Mean age 46 87) **R11** Contain mathematical and reporting symbols (±*<>†‡)	**R4** **R5** **R9.2:** Contain document main font AND **R12**: Contain predication/verb	Direction: DYNAMIC (e.g., choose the direction with the largest number of numeric patterns) Repeat: UNLIMITED
FIGURE	**R13:** Match figure common headings. **R13.1:** Treat the following lines in paragraph as the Figure captions	**R3** **R8** AND **R9.1** **R11**	**R4** **R5** **R9.2** AND **R12**:	Direction: DYNAMIC Repeat: UNLIMITED
TITLE	**R13:** Begin of paragraph AND **R14:** Contain in MEDLINE's title	**R3** **R14**	**R4** **R5** **R15:** Page number > 2	Direction: DOWN Repeat: 1
ABSTRACT	**R16:** Match the Abstract heading. **R17:** Contain in MEDLINE's abstract	**R3** **R17**	**R4** **R5** **R6** **R15**	Direction: DOWN Repeat: 2
REFERENCE	**R18:** Match reference common headings. **R19:** Prefix by a number AND **R20:** Contain PERSON or ORGANIZATION entities.	**R3** **R8**	**R4**	Direction: DOWN Repeat: UNLIMITED
AUTHOR	**R20:** Contain in MEDLINE's authors **R21:** Match common Authors information headings (e.g., correspondence to:, author affiliations, financial disclosures, etc)	**R3** **R8** AND **R9** **R22:** Contain LOCATION, PERSON, ORGANIZATION entities **R23:** Contain publication predications (submitted, published, supported, received, accepted, etc)	**R4** **R5**	Direction: DOWN Repeat: UNLIMITED
JOURNAL	**R24:** Match Journal metadata headings (e.g., original article, print issn:, link available on, etc)	**R3** **R23** **R25:** Match URL, IP address, price, DOI patterns	**R4** **R5**	Direction: DOWN Repeat: UNLIMITED

the MEDLINE title, whether the snippet was contained in the MEDLINE abstract, and bag-of-words features in which each feature is a word's frequency. We encountered a scalability issue when treating each snippet as a single feature vector. The number of text snippets is significantly large, which caused memory and training time issues. Therefore, we treated each document as a single classifier, and the classification decision is determined by majority voting. A snippet was assigned to a class if it received the majority of votes from multiple document classifiers. If there were ties, random assignment following a uniform random distribution was conducted.

4.3.4. Outcome Extraction System

To measure the impact of the classification algorithm on an IE system, we used a homegrown PICO (Population, Intervention, Comparison, and Outcome) extraction system. The goal of this system is to extract PICO data elements from full-text PDF reports to aid in SR development. The full description of the system is beyond the scope of this report; therefore, we present a brief description of one of the most mature components, the outcome extraction system. In short, the outcome extraction system is composed of two main stages: sentence selection and noun phrase chunking and filtering. The first stage accepts raw text input from any source and splits the input into multiple sentences by using Stanford NLP's sentence splitter. From those sentences, we selected only sentences that potentially contain outcome information (e.g., contain definitive phrases such as "outcomes were" and "study end points were", or reporting phrases such as "statistically different", "was improved", and "was measured"). In the second stage, we used Stanford's parser to generate a Penn tree and extract all noun phrase mentions. Since noun phrase extraction might detect exceedingly long phrases in complex sentences, we filtered phrases that have more than 10 words. Last, we applied a set of regular expression rules and semantic tests to collect outcome mentions. Regular expression rule looks for surrounding contexts [e.g., rate of (\\S+), incidence of (\\S+), etc.] to determine candidate mentions. In semantic test, we used MetaMap [35] to map text snippets to UMLS concepts restricted to the following semantic types: “Finding,” “Sign or Symptom,” “Laboratory or Test Result,” and “Disease and Syndrome.”

4.3.5. Evaluation Approach

We used the gold standard and methods described to test two hypotheses: H1: In the classification of PDF texts, the rule-based multipass sieve approach is more accurate than the machine-learning approach; and H2: PDF text classification improves performance of information extraction from full-text publications when compared to off-the-shelf PDF Box extraction.

To test H1, we randomly divided the gold standard into a 50-50 random split of documents. The first half of the dataset was used to train the ML classifier and to develop rules for the multipass sieve algorithm; the other half was used for the evaluation. Standard text classification evaluation metrics such as accuracy, recall, precision, and F-measure were calculated at the token level, with accuracy used as the primary endpoint. We used a Wilcoxon signed-rank test to assess the significance of the accuracy difference between the two approaches.

To test H2, we setup the experiment with two study arms. The first arm used raw texts extracted from PDF reports by using PDFBox. The second arm used the multipass sieve algorithm to categorize raw texts and filter all SEMISTRUCTURE and METADATA snippets before passing them to the IE system. Both arms were tested against the evaluation set of the gold standard described earlier. Recall, precision, F-measure, and number of split sentences are reported, with F-measure being the primary endpoint. We considered a correct mention if it contained phrases that appeared in the gold standard. A Wilcoxon signed-rank test was used to test the significance of the performance difference between the two study arms.

4.4. Results

We constructed a gold standard composed of 48 published reports that were included in eight Cochrane reviews. Those reports represent the publication formats of 34 different journals. A follow-up analysis showed that only 64% of studies have contents available in HTML pages, while all of them can be downloaded as PDF reports. All of them are randomized controlled trials, but only 16% have posted structured results on ClinicalTrials.gov. Raw text extraction using PDFBox generated 33,307 lines of text snippets, from which we were able to annotate 157 (0.5%) TITLE snippets, 1230 (3.7%) ABSTRACT snippets, 17,711 (53.2%) BODYTEXT snippets, 5596 (16.8%) SEMISTRUCTURE snippets, and 8613 (25.9%) METADATA snippets. In the outcome extraction dataset, we were able to manually annotate 204 outcome mentions, with a rate of 4.2 outcome mentions per document.

Performance comparison of two classification algorithms is summarized in Table 4.2. The multipass sieve approach achieved an average accuracy of 92.6% over 24 documents, for a significant improvement of 11.9% ($p<0.001$) over the machine-learning classifier. According to a power analysis, to reach a statistical power of 80% for the effect size we found (11.9%), a sample of 16 documents would be needed. For specific data elements, the F-measures for the multipass sieve approach were better than those for the machine

Table 4.2 - Performance comparison of the multipass approach versus the machine-learning approach.

	Machine-learning classifier				Multipass sieve approach			
	Acc.	Recall	Precision	F1	Acc.	Recall	Precision	F1
TITLE (%)	80.7	85.2	**84.5**	83.2	**92.6** (p<0.001)	**95.8**	80.3	**85.3**
ABSTRACT (%)		58.1	70.8	63.3		**86.6**	**82.2**	**83.1**
BODYTEXT (%)		89.1	88.9	88.3		**95.9**	**90.6**	**93**
SEMISTRUCTURE (%)		46.4	67.8	51.1		**88.6**	**91.5**	**88.9**
METADATA (%)		73.9	61.7	63.6		**82**	**96.8**	**87.3**

learning classifier (TITLE +2.1%, ABSTRACT +19.8%; BODYTEXT +4.7%; SEMISTRUCTURE +37.8%; MEDADATA +23.7%). Overall, performance improvements were seen in almost all evaluation metrics, except precision for TITLE snippet (-4.2%).

For outcome extraction task, the IE system that operated on PDF texts after filtering out SEMISTRUCTURE and METADATA snippets had better performance than off-the-shelf PDF Box extraction (Table 4.3). The improvement on recall was not significant (+0.6%; p=0.16), while precision was significantly improved (+4.5%; p<0.001). F-measure increased significantly by 4.3% (p<0.001). Most notably, filtering publication metadata and semistructured texts reduced the number of sentences to be processed by 46.5%.

Table 4.3 - Comparison of IE performance operated on original PDF texts vs PDF texts after filtering SEMISTRUCTURE and METADATA snippets.

	Original texts extracted from PDF reports	PDF texts after filtering SEMISTRUCTURE and METADATA snippets
Recall (%)	95.8	**96.4** (p=0.16)
Precision (%)	45.1	**49.6** (p<0.001)
F Score (%)	60.4	**64.7** (p<0.001)
Average number of sentences (per document)	256	137

4.5. Discussion

4.5.1. PDF Text Classification

We designed and evaluated a rule-based multipass sieve approach to categorize texts extracted from PDF documents. The approach is an alternative to the machine-learning algorithms that are commonly used in text classification studies [36-38]. Overall, the multipass sieve classifier significantly outperformed the machine-learning classifier (accuracy 92.6% vs. 80.7%, $p<0.001$). Our results are significant because PDF is the preferred format for the data extraction of clinical studies in the development of systematic reviews. Our dataset showed that 36% of studies published in PDF format did not have an HTML version available. This is further confirmed with Cochrane Heart Group's systematic reviewers, who stated that PDF is often the preferred choice due to wider adoption and availability offline. Therefore, IE systems need to operate on PDF documents to effectively support systematic review development.

The annotation results showed that 26% of text snippets are publication metadata, and 17% are semistructured texts. These findings confirm the heterogeneity problem in PDF reports. There are off-the-shelf tools developed to help detect the PDF logical structure. However, our preliminary studies could not find one that meets the needs of data extraction, either because the classification schema did not match the needs of data extraction, or because tools did not perform well in our systematic review dataset.

The multipass sieve framework proposed in this study has several strengths. First, its accuracy outperformed a machine-learning approach by 11.9%. Second, the framework is flexible and extendable. Developers have the flexibility to create and add new sieves and rules to target new data elements. Rules are organized at different stages to facilitate maintainability and extension. Third, the algorithm is intuitive, i.e., it operates in a way similar to human screening, in which documents are scanned for the prominent signatures (e.g., heading, caption) and then examine the contents.

4.5.2. Clinical Outcome Extraction

Our baseline system achieved 96% recall and 45% precision. While the recall is adequate, precision needs further improvement. The use of the classification algorithm to filter publication metadata and semistructured texts improved recall by 0.6% and precision by 4.5%. The difference in recall was not significant, since the baseline recall was very high with little room for improvement. The precision improvement corresponds to a reduction of 15% in the number of false-positive mentions; therefore the algorithm would considerably reduce the number of mentions that reviewers would need to correct in a semiautomated data extraction process.

A subsequent analysis showed that texts without filtering sometimes have publication metadata and semistructured texts embedded within body text fragments, breaking up sentences. Detecting and filtering those nonprose texts improved the performance of the sentence splitter algorithm (e.g., Stanford sentence splitter). Defining correct sentence boundaries is an important prerequisite step for most NLP systems. Incorrect sentence boundaries negatively impact subsequent NLP pipelines such as syntactic parsing and phrase chunking. Moreover, the filtering reduced the number of sentences to be processed by 46.5%. This reduction improved the efficiency of the NLP approach by reducing processing time without causing performance loss.

This study evaluated impacts of the PDF structure recognition on an IE system. However, the proposed technique is also potentially useful in other areas, such as information retrieval, automated document classification, and library management. These areas share the PDF heterogeneity problem that might degrade the performance of any text processing approaches.

4.5.3. Limitations

The algorithm takes advantage of MEDLINE resources such as title, abstract, and author metadata, which reduces the applicability of the method to documents not indexed in MEDLINE. However, our focus was on biomedical research publications and the majority of these resources are openly available in the MEDLINE database. This study did not test an exhaustive list of machine-learning algorithms and their optimization parameters. While it may still be possible to further improve the accuracy of the machine-learning approach, we selected the default SVM baseline due to its popularity and to minimize the risk of overfitting. This study used the outcome extraction module to perform an intrinsic evaluation. The impacts of filtering on other data elements, such as sample size and intervention, are unknown. Our analysis confirmed that filtering affects the performance of the sentence-splitter, which impacts any extraction methods that rely on the assumption of a correct sentence boundary. We did not use semistructured texts as the source of extraction, although they might contain outcome mentions. Semistructured texts require different extraction strategies that rely more on pattern-matching and dictionary-matching than on syntactic parsing and chunking. Because we deemed recall to be satisfactory (96.4%), an additional source of extraction was considered unnecessary.

4.5.4. Future Work

Areas that demand future research include testing the classification algorithm on a diverse set of PDF documents, validating the usefulness in other text mining research such as information retrieval, document classification, and IE of other data elements, and using semistructured texts instead of excluding them.

4.6. Conclusion

We present an alternative approach for PDF structure recognition by using PDFBox to extract raw texts and a multipass sieve algorithm for classification. The multipass sieve algorithm achieved a higher accuracy than the more commonly used machine-learning classification approach. The multipass sieve algorithm also improved the performance of an IE system compared to off-the-shelf PDF extraction. PDF structure recognition unlocks the door to conduct text mining research on PDF files, an important information source for biomedical research.

4.7. Acknowledgments

This work was made possible by funding from National Library of Medicine grants (R00LM011389 and R01LM011416-01). We are also thankful to the guidance provided by the US Satellite of the Cochrane Heart Group, led by Dr. Mark Huffman.

4.8. References

[1] D.L. Sackett, W.M. Rosenberg, J.A. Gray, R.B. Haynes, W.S. Richardson, Evidence based medicine: what it is and what it isn't, BMJ (Clinical research ed). 312 (7023) (1996) 71–72.

[2] M. Ware, M. Mabe, An overview of scientific and scholarly journal publishing, The STM Report, 2009.

[3] Statistical Reports on MEDLINE®/PubMed® Baseline Data. Available from: <https://www.nlm.nih.gov/bsd/licensee/baselinestats.html>.

[4] M.E. Schaafsma, The Cochrane Collaboration Treasurer's Report, 2012.

[5] Cochrane Collaboration, Directors' Reports and Financial Statements, 2013.

[6] K.G. Shojania, M. Sampson, M.T. Ansari, J. Ji, S. Doucette, D. Moher, How quickly do systematic reviews go out of date? A survival analysis, Ann. Intern. Med. 147 (4) (2007) 224–233.

[7] P. Bragge, O. Clavisi, T. Turner, E. Tavender, A. Collie, R.L. Gruen, The global evidence mapping initiative: scoping research in broad topic areas, BMC Med. Res. Methodol. 11 (1) (2011) 92.

[8] J.P. Higgins, S. Green, Cochrane Handbook for Systematic Reviews of Interventions, Wiley Online Library, 2008.

[9] R.L. Summerscales, Automatic Summarization of Clinical Abstracts for Evidence-based Medicine, Illinois Institute of Technology, 2013.

[10] K.-C. Huang, I.J. Chiang, F. Xiao, C.-C. Liao, C.C.-H. Liu, J.-M. Wong, PICO element detection in medical text without metadata: are first sentences enough?, J Biomed. Inform. 46 (5) (2013) 940–946.

[11] F. Boudin, J.-Y. Nie, J.C. Bartlett, R. Grad, P. Pluye, M. Dawes, Combining classifiers for robust PICO element detection, BMC Med. Inform. Decis. Mak. 10 (2010) 29.

[12] H. Zhu, Y. Ni, P. Cai, Z. Qiu, F. Cao, Automatic extracting of patient-related attributes: disease, age, gender and race, Stud. Health Technol. Inform. 180 (2012) 589–593.

[13] D.P.A. Corney, B.F. Buxton, W.B. Langdon, D.T. Jones, BioRAT: extracting biological information from full-length papers, Bioinformatics 20 (17) (2004) 3206–3213.

[14] K. Verspoor, A. Mackinlay, J.D. Cohn, M.E. Wall, Detection of protein catalytic sites in the biomedical literature, In: Pac Symp Biocomput, 2013, pp. 433–444.

[15] J. Hakenberg, R. Leaman, N.H. Vo, S. Jonnalagadda, R. Sullivan, C. Miller, et al., Efficient extraction of protein–protein interactions from full-text articles, IEEE/ ACM Trans. Comput. Biol. Bioinform. 7 (3) (2010) 481–494.

[16] W. Hsu, W. Speier, R.K. Taira, Automated extraction of reported statistical analyses: towards a logical representation of clinical trial literature, in: AMIA Annual Symposium Proceedings/AMIA Symposium AMIA Symposium, 2012, pp. 350–359.

[17] B. de Bruijn, S. Carini, S. Kiritchenko, J. Martin, I. Sim, Automated information extraction of key trial design elements from clinical trial publications, in: AMIA Annual Symposium Proceedings/AMIA Symposium AMIA Symposium , 2008, pp. 141–145.

[18] S. Kiritchenko, B. de Bruijn, S. Carini, J. Martin, I. Sim, ExaCT: automatic extraction of clinical trial characteristics from journal publications, BMC Med. Inform. Decis. Mak. 10 (2010) 56.

[19] R. Kern, K. Jack, M. Hristakeva, M. Granitzer, TeamBeam meta-data extraction from scientific literature, D-Lib Magazine 18 (7) (2012) 1.

[20] M. Granitzer, M. Hristakeva, R. Knight, K. Jack, R. Kern (Eds.), A comparison of layout based bibliographic metadata extraction techniques, Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, ACM, 2012.

[21] H. Han, C.L. Giles, E. Manavoglu, H. Zha, Z. Zhang, E. Fox (Eds.), Automatic document metadata extraction using support vector machines, Digital Libraries 2003 Proceedings 2003 Joint Conference, IEEE, 2003.

[22] M.T. Luong, T.D. Nguyen, M.Y. Kan, Logical structure recovery in scholarly articles with rich document features, Multimedia Storage Retrieval Innovat. Digital Lib. Syst. (2012) 270.

[23] A. Constantin, S. Pettifer, A. Voronkov (Eds.), PDFX fully-automated PDF-to- XML conversion of scientific literature, Proceedings of the 2013 ACM symposium on Document Engineering, ACM , 2013.

[24] F. Kboubi, A.H. Chabi, M.B. Ahmed (Eds.), Table recognition evaluation and combination methods, Document Analysis and Recognition, 2005 Proceedings Eighth International Conference on, IEEE , 2005.

[25] H. Chao, J. Fan, Layout and content extraction for pdf documents, in: Document Analysis Systems VI, Springer, 2004, pp. 213–224.

[26] S. Klampfl, K. Jack, R. Kern, A comparison of two unsupervised table recognition methods from digital scientific articles, D-Lib Mag. 20 (11) (2014) 7.

[27] S. Klampfl, R. Kern, An unsupervised machine learning approach to body text and table of contents extraction from digital scientific articles, In: Research and Advanced Technology for Digital Libraries, Springer , 2013 , pp. 144–155

[28] E. Oro, M. Ruffolo (Eds.), PDF-TREX: An approach for recognizing and extracting tables from PDF documents, Document Analysis and Recognition, 2009 ICDAR'09 10th International Conference on, IEEE, 2009.

[29] T. Kenter, D. Maynard, Using Gate as an Annotation Tool, <http://www.ia.hiof.no/softengin/ias/literature/sw/annogate.pdf >

[30] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, et al., **A** multi-pass sieve for coreference resolution, Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2010.

[31] S.R. Jonnalagadda, D. Li, S. Sohn, S.T. Wu, K. Wagholikar, M. Torii, et al., Coreference analysis in clinical notes: a multi-pass sieve with alternate anaphora resolution modules, J. Am. Med. Inform. Assoc. 19 (5) (2012) 867– 874.

[32] M.-S. Ong, F. Magrabi, E. Coiera, Automated categorisation of clinical incident reports using statistical text classification, *Qual. Saf.* Health Care 19(6) (2010) e55.

[33] M.K. Ross, K.-W. Lin, K. Truong, A. Kumar, M. Conway, Text Categorization of Heart, Lung, and Blood Studies in the Database of Genotypes and Phenotypes (dbGaP) Utilizing n-grams and Metadata Features, Biomed. Inform. Insights 6 (2013) 35-45.

[34] O. Uzuner, I. Goldstein, Y. Luo, I. Kohane, Identifying patient smoking status from medical discharge records, J. Am. Med. Inform. Assoc. 15(1) (2008) 14-24.

[35] A.R. Aronson, f, NLM, NIH, DHHS., Bethesda, MD , 2006.

[36] M.H. Song, S.H. Kim, D.K. Park, Y.H. Lee, A multi-classifier based guideline sentence classification system, Healthcare Inform. Res. 17 (4) (2011) 224–231.

[37] S. Sohn,M. Torii, D. Li, K. Wagholikar, S. Wu, H. Liu, A hybrid approach to sentiment sentence classification in suicide notes, Biomed. Inform. Insights 5(Suppl. 1) (2012) 43-50.

[38] D.D. Bui, Q. Zeng-Treitler, Learning regular expressions for clinical text classification, J. Am. Med. Inform. Assoc. 21 (5) (2014) 850–857.

CHAPTER 5

TOWARDS A COMPUTER-AIDED DATA EXTRACTION APPROACH TO SUPPORT SYSTEMATIC REVIEW DEVELOPMENT

Duy Duc An Bui, Guilherme Del Fiol, Siddhartha Jonnalagadda

Journal of Biomedical Informatics (Submitted)

5.1. Abstract

Extracting data from publication reports is a standard process in systematic review development. However, the data extraction process still relies too much on manual effort, which is slow, costly and subject to human errors. In this study, we aimed to develop a computer-aided solution to enhance productivity and reduce errors in the traditional data extraction process.

We developed a system to help extracting sample size, group size and PICO values from publication reports. The system is composed of 2 main stages: prioritizing sentences for specific target element, and recommending key phrases from the sentence. To evaluate the system, we built a gold standard based on the data extraction summary developed by Cochrane review authors. The system was evaluated at sentence and fragment levels. We tested a hypothesis as to whether using machine-learning approach to prioritize full-text sentences was more effective than the manual abstract+title screening approach.

At sentence level, we consider recall as primary outcome and precision as secondary outcome. Our best sentence classifiers achieved 91.5% recall and 58.6% precision. In comparison with abstract+title screening, top ranked sentences proposed by our system achieved nonsignification recall improvement on SampleSize/GroupSize element (+8.4%, p=0.38), significant recall improvement on Outcome element (+23.5%, p<0.001). Both the system and manual approach achieved perfect recall on Population and Intervention/Control elements. Significant precision improvements were seen on all elements (SampleSize/GroupSize: +6.1%, p<0.001, Population: +20.7% , p=0.003, Intervention/Control: +21.8%, p<0.001, Outcome: 30.6%, p<0.001). At fragment level, the ensemble approach combining rule-based, concept-mapping, and dictionary-based methods performed better than the individual method, which achieved 85.1% F-measure.

Our system achieved decent performance for sentence ranking and key phrase extraction. Furthermore, we demonstrated the system performed equally or better than abstract screening and might reduce time and errors in full-text screening. The system has the potential to realize the prospect of computer-aided data extraction.

5.2. Introduction

Systematic reviews (SR) are important information sources for healthcare providers, researchers, and policy makers. SR attempts to comprehensively identify, appraise and synthesize best available evidence to find reliable answers to research questions [1]. SRs can be conducted by a single author or a large group of authors. Cochranne Collaboration is an internationally recognized nonprofit organization that developed SRs for health-related topics. Cochrane SRs were aimed for highest standard in evidence-based practice [2]. Cochrane usage data in 2009 showed that “Every day someone, somewhere searches

The Cochrane Library every second, reads an abstract every two seconds and downloads a full-text article every three seconds."[3]

The development of systematic reviews was criticized as resource-intensive and slow [4-6]. Data extraction is one of the SR development steps whose goal is to collect relevant information from published reports to perform the data analysis and quality assessment. Studies showed that the manual data extraction task had a high prevalence of errors [7, 8]. Therefore, this study aims to enhance the manual extraction processing using the computer-aided solutions. The ultimate goal is to enhance productivity and to reduce human errors.

There were previous computer systems that can be adapted to help with the data extraction task. Boundin et al. [9] and Huang et al. [10, 11], and Kim et al. [12] were interested in a machine-learning approach to classify sentences that contain PICO elements. PICO, which stands for Population, Intervention, Control, and Outcome, is a popular framework used to formulate and find answers to clinical questions. Demner-Fushman and Lin [13], Kelly and Yang [14], and Hansen et al. [15] employed rule-based and machine-learning approaches to extract PICO and patient-related attributes. Those studies selected extraction sources from study abstracts which might not contain sufficient information for the review question. Extraction from full-text reports is the typical requirement in systematic review development [16]. However, extraction from full-text documents is more challenging since we have to deal with a larger chunk of text with abundance of redundancies and noises. Kiritchenko et al. [17] and de Bruijin et al. [18] developed ExaCT to help extracting clinical trial characteristics, is considered as one of the most successful full-text extraction system for clinical elements. Their method first used a machine-learning classifier to select top 5 relevant sentences for each element, then used hand-crafted weak extraction rules to collect values for each element. ExaCT selected RCT studies from top 5 core clinical journals, and has full-texts available in HTML format. However, systematic review in practice might select studies outside of the top 5 clinical journals and many studies are not available in HTML format. In this research, we aim to enhance previous works on sentence classification and information extraction for better support systematic review. We proposed a computer system having two main goals: prioritization of relevant sentences and key phrase recommendation. We employed a bottom-up approach that developed gold standard very close to the development of Cochrane reviews. We enabled extraction from PDF format, a popular and common extraction source in systematic review. We enhanced the sentence classification model with contextual and semantic features, and introduced a concept-mapping approach to complement the rule-based approach in extracting literal data elements.

5.3. Methods

Our methods comprise three main parts: (1) development of data extraction gold standard; (2) development of a semiautomated extraction system that helps extracting key clinical trial characteristics from full-text PDF reports; and (3) evaluation of sentence ranking and key phrase recommendations from the extraction system.

5.3.1. Gold Standard

From the Cochrane Library, we retrieved systematic reviews on the subject "heart and circulation" that were published after October 2014. The included primary studies were extracted and the corresponding PDF reports were collected. We excluded studies that had been reported in multiple publications and nonrandomized trials since randomized controlled trial publications are the focus of our research.

To develop the data extraction gold standard, we started with evidence summary tables developed by Cochrane reviewers and reviewed the full-text reports to validate and supplement the gold standard. Specifically, we looked for mentions (e.g., synonyms, abbreviations, morphological variations) that are co-referred to the same entity in the documents. For each document, we built an extraction template including five data elements: sample size, group size, population, intervention/control, and outcome.

Sample size is defined as the total number of patients enrolled in the study and included in the statistical analysis. Group size is the number of participants in each study group. Sample size can be inferred by summing up all group sizes.

Population is defined as the main characteristics of the patient population included in the study. The population characteristics describe the group of patients sharing the same disease, demographics, or that underwent the same medical procedure.

Intervention/Control is defined as the name of a therapy or control treatment. We do not distinguish intervention and control groups, since this is a naming convention not always explicitly mentioned in texts. For instance, groups absent of an intervention treatment or placebo-treated groups are implicitly classified as control groups.

Outcome is defined as measurements used to assess a study hypothesis, including all clinical attributes and adverse effects on patients. We do not distinguish between primary outcome and secondary outcomes in individual studies, since the reviewers might select different primary outcomes relevant to the review question to perform the meta-analysis. Table 5.1 shows examples of the extraction template with data extracted from two publication reports.

Table 5.1 - Data extraction template with examples extracted from the Cochrane review "Primary prophylaxis for venous thromboembolism in ambulatory cancer patients receiving chemotherapy".

Cochrane ID	Klerk 2005
Study Title	The effect of low molecular weight heparin on survival in patients with advanced malignancy
Sample Size	302
Group Size	148, 154
Population	Advanced Malignancy
Intervention/Control	low molecular weight heparin|Nadroparin Placebo
Outcome	death from any cause|death as a result of any cause|death major bleeding|non-major bleeding|bleeding
Cochrane ID	Mitchell 2003
Study Title	Trend to efficacy and safety using antithrombin concentrate in prevention of thrombosis in children receiving l-asparaginase for acute lymphoblastic leukemia. Results of the PAARKA study.
Sample Size	85
Group Size	25, 60
Population	Children acute lymphoblastic leukaemia
Intervention/Control	antithrombin
Outcome	symptomatic or asymptomatic thrombotic event|thrombotic event major and minor bleeding|bleeding

5.3.2. System Overview

The overall system architecture and data flow are summarized in Figure 5.1. The system takes the PDF publication reports as input, and outputs a list of relevant sentences for each data element as well as recommended key phrases in each sentence. In a pipeline design, the system comprises of seven stages.

5.3.2.1. PDF Text Classification & Filtering

We used the system described in Chapter 4 to extract and categorize text snippets from PDF reports, then filtered all semistructured and publication metadata snippets.

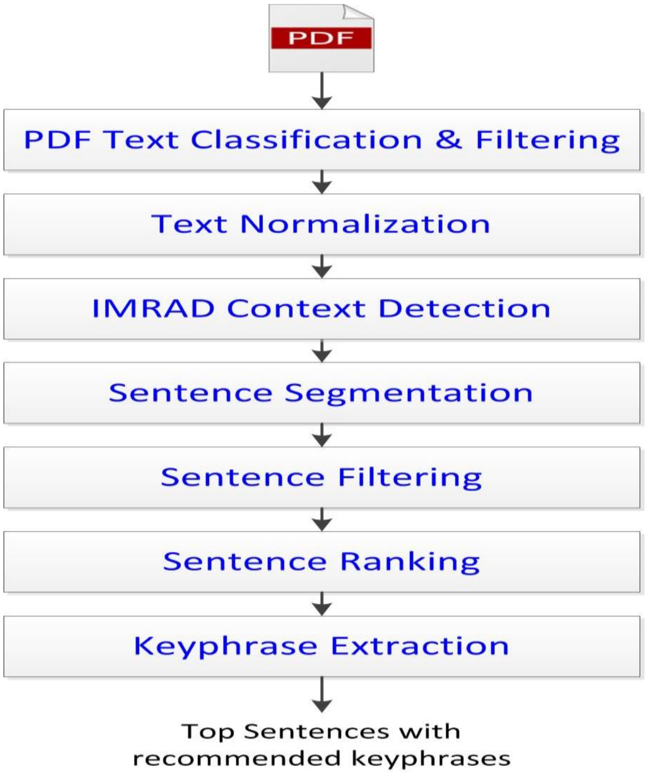


Figure 5.1 - System architecture overview.

5.3.2.2. Text Normalization

The goal of this step is to translate texts into canonical form. Specifically, we find and replace all numbers in literal expression to numeric format (e.g., “a hundred and three patients” → 103 patients). We developed an acronym normalization algorithm that reads full-text documents, detects, and replaces acronyms to their full form. The acronym normalization algorithm first checks all parenthetical expressions and the preceding text (e.g., “small cell lung cancer (SCLC)”) for candidate acronym pairs, then uses the pattern of initial letters for validation. This is the most frequently used pattern in publication reports. The acronym normalization increases the clarity of the sentences in manual review as well as improves performance of the concept-mapping approach in the subsequent stage.

5.3.2.3. IMRAD Context Detection

This step attempts to assign the common scientific organization structure IMRAD (introduction, methods, results, and discussion) to text snippets. It recognizes the common section headings from a prebuilt collection, and assigns the IMRAD context labels to text snippets placed between those headings. The text snippets are clustered into different context nodes, as illustrated in Figure 5.2.

5.3.2.4. Sentence Segmentation

We used the Stanford NLP sentence splitter [19] to perform sentence segmentation in different context nodes. With this approach, a sentence's contexts can be easily determined, which is useful for subsequent text mining steps.

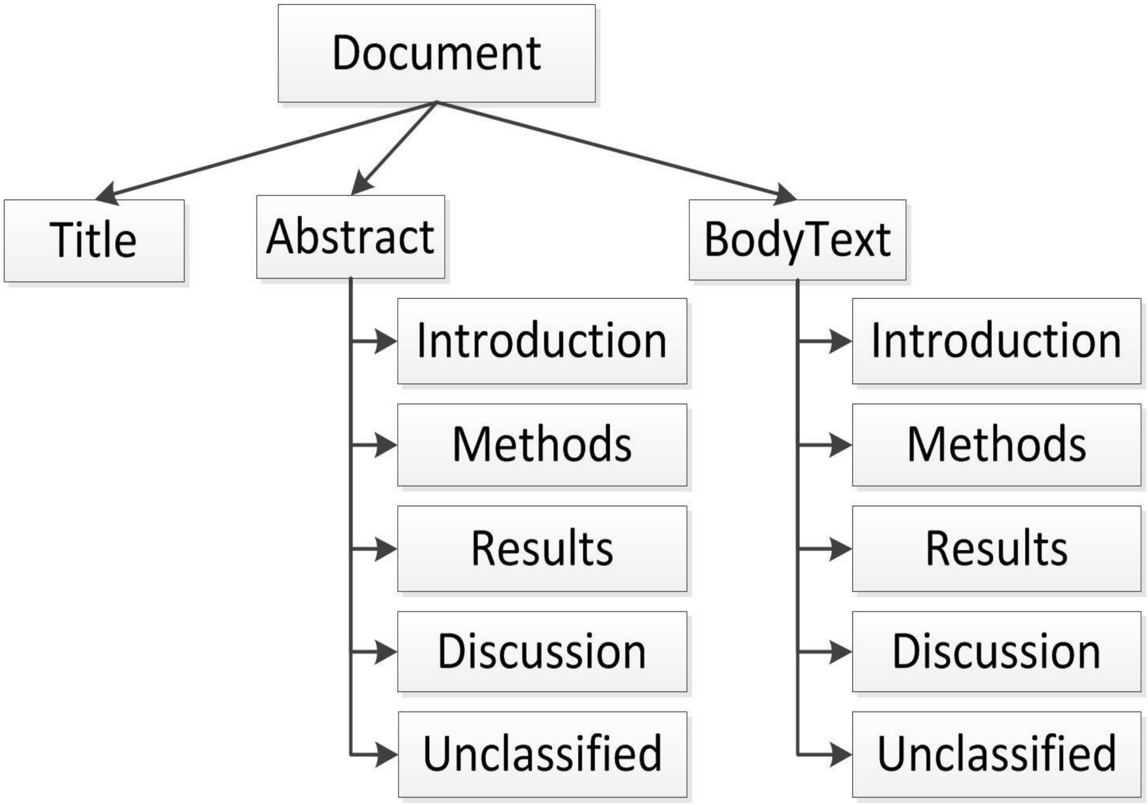


Figure 5.2 - The standard structure organization of text in a scientific article.

5.3.2.5. Sentence Filtering

In this step, we attempt to filter all sentences that discuss background knowledge and therefore are not relevant to the extraction goal. We filter sentences having the IMRAD context INTRODUCTION, sentences containing year and citation expressions, and sentences referring to other studies (e.g., containing phrases such as "these trials", "et al.", and "previous studies").

5.3.2.6. Sentence Ranking

The goal of this step is to prioritize sentences for each individual data element. We used the Support Vector Machine Regression (SVMR) implemented in Weka [20] with linear kernel and other Weka default parameters. To train the regression model, we used 50% of the sentences in the gold standard as the training set. The response variable is set to the number of times the target element appeared in the sentence. The predictor variables or features can be divided into Bag-Of-Term (BOT), Context, and Semantic groups.

The BOT group is based on words, terms, or patterns that appear in the sentence. (1) The top 100 most frequent words that are present in relevant sentences (i.e., sentences that contain at least one target element for data extraction) were selected as BOT features. We used the frequency of those words in the sentence to generate a feature vector. (2) We constructed a binary variable determining whether the sentence contains at least a true-positive mention in the training set. (3) We used regular expression-based features to capture text patterns that are strong indicators of relevant sentences. For the scope of the present study, we maintained a small set of regular expression features per data element. A complete set of regular expression and word based features is described in Table 5.2.

Table 5.2 - Complete set of regular expression and word based features.

Data Element	Bag Of Term Features
SampleSize /GroupSize	Regex: total of \\d+ patients n = \\d+ Words: results, patients, completed, the, trial, having, defaulted, of, control, maintained, sinus, tachy, cardia, compared, with, only, who, received, propranolol, postoperatively, -lrb-, p, -rrb-, and, methods, a, total, undergoing, bimaxillary, surgery, were, studied, in, prospective, randomized, double-blind, study, oral, mg, or, placebo, as, premedication, before, hypotensive, anesthesia, sodium, nitroprusside, peak, plasma, concentration, is, attained, within, to, minutes, after, administration, half-life, hours, although, duration, pharmacologic, effect, longer, population, consisted, all, figure, demonstrates, mean, heart, rates, both, groups, from, baseline, rate, time, every, infusion, for, shortest, was, given, induction, because, onset, drug, about, ingestion, which, appropriate, orthognathic, at, our, institute, where, we, start
Population	Regex: patients? with Words: university, department, of, dermatology, newcastle, upon, tyne, double-blind, tkial, zinc, sulphate, in, the, treatment, chronic, venous, leg, ulceration, m., w., greaves, and, f, results, a, trial, patients, with, are, reported, total, -lrb-, female, -rrb-, who, were, attending, either, royal, victoria, infirmary, time, or, dryburn, hospital, durham, studied, it, is, therefore, concluded, that, there, no, justification, for, administration, this, form, continued, propranolol, following, coronary, bypass, surgery, antiarrhythmic, effects, therapy, isolated, appears, to, be, safe, efficacious, method, decreasing, incidence, postoperative, supraventricular, tachycardias, although, operative, mortality, elective, major, centers, presently, below, significant, morbidity, remains, particularly, regard, early, arrhythmias, supraventricular, tachycardia, has, been

Table 5.2 - Continued

<table>
<tr><th>Data Element</th><th>Bag Of Term Features</th></tr>
<tr><td>Intervention /Control</td><td>Words:
university, department, of, dermatology, newcastle, upon, tyne, double-blind, tkial, zinc, sulphate, in, the, treatment, chronic, venous, leg, ulceration, m., w., greaves, and, f, results, a, trial, patients, with, are, reported, no, significantly, increased, rate, healing, was, observed, zinc-treated, group, this, paper, we, describe, oral, condition, were, randomly, allocated, to, either, placebo, or, capsules, by, pharmacy, royal, victoria, infirmary, systemic, other, than, prescribed, double-blind, op, mean, per, week, determined, for, each, patient, dividing, total, linear, re-epithelialization, time, taken, if, incomplete, end, although, greater, difference, did, not, reach, statistical, significance, their, study, designed, primarily, investigate, plasma, concentrations, before, after, side-effects, these, authors</td></tr>
<tr><td>Outcome</td><td>Regex:
(end points?|endpoints?| outcomes?) (was|were|included)
Words:
no, significantly, increased, rate, of, healing, was, observed, in, the, zinc-treated, group, an, approximate, estimate, made, by, us, -lrb-, m.w, mean, difference, between, pairs, corresponding, radii, represents, extent, linear, re-epithelialization, during, treatment, period, double-blinb, trial, op, zinc, sulphate, per, week, determined, for, each, patient, dividing, total, time, taken, to, or, if, incomplete, end, although, greater, did, not, reach, statistical, significance, present, double-blind, study, does, support, these, preliminary, observations, since, occurred, compared, with, control, groups, they, also, found, that, oral, accelerated, patients, venous, leg, ulcers, who, have, a, low, plasma, concentration, there, were, preoperative, differences, apparent, degree, coronary, arterial, disease, left</td></tr>
</table>

The Context group includes two features based on the contexts of the sentence. The document-structure feature is a nominal attribute that takes one of three categories: TITLE, ABSTRACT, or BODYTEXT. The IMRAD nominal feature accepts one of four categories: INTRODUCTION, METHODS, RESULTS, DISCUSSION. If the sentence contexts were not determined from the previous steps, they are treated as missing values.

The Semantic group uses 15 UMLS semantic groups [21] as features. We used MetaMap [22] to map text to UMLS concepts, from which we map the UMLS sematic types to sematic groups. Then we aggregate and compute the sematic group features based on their frequency.

Based on three feature groups, we created four machine-learning models for comparing and selecting the best model for each data element: BOT, BOT+Context, BOT+Semantic, and BOT+Context+Semantic.

5.3.2.7. Keyphrase Extraction

The goal of this step is to recognize key phrases from the sentences to help reviewers quickly identify parts of the sentence that are relevant to the extraction goal. Based on the type of data element, we employed a subset of the following techniques.

5.3.2.7.1. Regular Expression Matching

Since numbers are normalized to numeric expressions, regular expression (regex) pattern matching is one of the most useful techniques in the recognition of numeric values. We used a list of regular expressions rules (Table 5.3) to extract numeric values for sample size and group size. Each regex rule contains context expressions and capturing groups referring to target elements. Since each sentence might have a unique way to convey the

Table 5.3 - Regular expressions and semantic types used for extracting individual elements.

<table>
<tr><th>Data element</th><th>Extraction Methods</th></tr>
<tr><td>Sample Size/ Group Size</td><td>Regular expression:
• (\\d+) met|meet (?: \\S+){0,1} criteria
• randomized(?: \\S+)? (\\d+)(?: \\S+)? patients?
• (\\d+)(?: \\S+){0,2} were randomized
• patients , (\\d+) ,
• (\\d+)(?: \\S+){0,1} patients?
• n = (\\d+)
• -LRB- n (\\d+) -RRB-
• (\\d+)</td></tr>
<tr><td>Population</td><td>Semantic type:
• Disease or Syndrome
• Therapeutic or Preventive Procedure
• Finding
• Neoplastic Process
• Medical Device</td></tr>
<tr><td>Intervention/ Control</td><td>Semantic type:
• Pharmacologic Substance
• Inorganic Chemical
• Element, Ion, or Isotope
• Therapeutic or Preventive Procedure
• Clinical Drug
• Organic Chemical</td></tr>
<tr><td>Outcome</td><td>Regular expression:
• ^(the(?: \\S+){1,3} rate)$
• ^((?: the)?(?: \\S+){1,3} volume) + B
• outcome was((?: \\S+){1,5}) + B
• differences? in((?: \\S+){1,5}) or((?: \\S+){1,5}) + B
• (?:differences?|reductions?|improvements?) (?:in|of)((?: \\S+){1,5}) + B
• by((?: \\S+){1,5}) reduction + B
• (?:prolongs?|improves?|decreases?)((?: \\S+){1,5}) + B
• effects? of(?: \\S+){1,5} on((?: \\S+){1,5})" + B
• (anti-\\S+ effects?)" + B
• (length of(?: \\S+){1,3})" + B
Semantic type:
• Finding
• Disease or Syndrome
• Pathologic Function
• Laboratory or Test Result
• Molecular Function
Therapeutic or Preventive Procedure</td></tr>
</table>

B = BOUNDARY = (?: and| to| in| with| between| \\.| _| ,|$)

numeric value, only the best match was considered.

5.3.2.7.2. Noun Phrase Chunking and Regex Matching

For literal expression data elements (e.g., outcome), applying regular expression matching might detect very long phrases, which is less useful for key phrase recommendation. Therefore, we performed noun phrase chunking to restrict the matching to only noun phrases of the sentence. To perform the noun phrase chunking, we used the Stanford parser to generate a Penn tree and used the Tregex parser to collect all noun phrase expressions.

5.3.2.7.3. Concept-Mapping and Semantic Type Restriction

The majority of key terms can be found in controlled medical terminologies. Concept-mapping was found to be an effective approach. We used MetaMap [22] to detect medical terms from the sentence that can be mapped to UMLS concepts. To enhance precision, we restricted the mapping to few semantic types relevant to the target elements. The selection of optimal set of semantic types is based on experimental testing on the training set. At present, we maintained an optimal set of semantic types for each data element. A set of semantic types are shared in multiple elements.

5.3.2.7.4. Supplement Dictionary

This approach encourages the development of controlled terms for individual literal data elements. Since there are reuses of key terms in multiple publication reports, maintaining a good coverage, element-specific dictionary would help improve accuracy as well as save computing resources. We included all true positive terms in our training set to the dictionary and matched them against the sentence to extract the candidate terms. Since our training set is relatively small and not representative, we still need to combine this method with other generalizable methods.

5.3.2.7.5. Postprocessing

This step filters phrases that are lengthy (> 5 words), phrases contained in other phrases, and phrases contained in a stop list. The stop list was constructed using the top 20 most frequent false-positive terms upon evaluating the system on the training set, which were not recognized as true-positives in all training documents.

5.3.3. Evaluation Approach

The study evaluation has two parts: sentence-level evaluation and fragment-level evaluation. In sentence-level evaluation, we evaluate the top N sentences recommended by the machine-learning classifier. N is set to the number of MEDLINE abstract sentences plus 1 (title). The following metrics were used in the sentence-level evaluation:

$$recall = \frac{Number\ of\ unique\ true\ possitive\ mentions\ contained\ in\ N\ sentences}{Total\ number\ of mentions\ in\ document} \tag{6}$$

$$precision = \frac{Number\ of\ sentences\ contain\ at\ least\ one\ true\ possitive\ mentions}{N} \tag{7}$$

We used a different definition of true positive for sample size and group size versus other literal data elements. Sample size and group size values are not always presented separately in all sentences, so we merged them into a single element SampleSize/GroupSize to complement each other. We applied a binary rule: true if the sentence contains the sample size value or all group size values, and false otherwise.

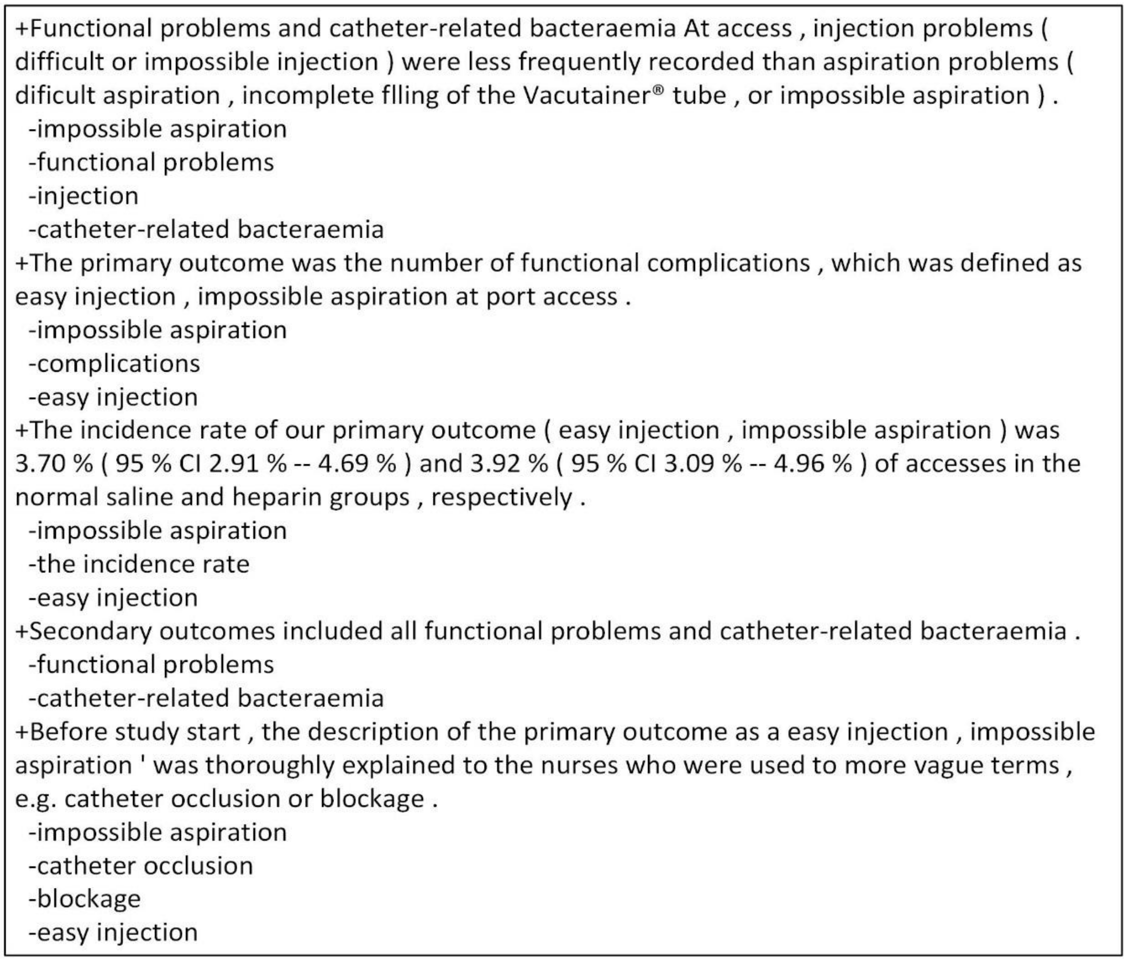
In the sentence-level evaluation, we tested the following hypothesis: the machine-learning classifier to prioritize sentences in full-text performed equally or better than manual title and abstract screening. To obtain the baseline abstract+title sentences, we applied step Text normalization and Sentence Segmentation on MEDLINE abstracts and titles. Those sentences followed the exact order of the title and abstract screening that is followed in the standard manual process. We performed the same evaluation method with MEDLINE sentences and ML classifier recommended sentences. Recall was selected as the primary outcome and precision as a secondary outcome. To test the significance of performance difference, we used the Chi-square test to assess the sample size element, and the Wilcoxon signed rank test to assess other literal elements.

In the fragment-level evaluation, we extracted a test set including the top N classified sentences containing at least one true positive mention. Recall, precision, and f-measure were measured at the sentence unit. To consider a recommendation as a true positive, an exact match was required for numeric elements (i.e., sample size and group size). For literal elements, phrases of up to five words that contain a correct mention or one of its synonyms were considered true positives. We evaluated and compared the performance of the following extraction methods: Regex Matching, Concept-Mapping, Supplement Terminology, and a combination of these three methods.

5.4. Results

The gold standard was composed of 48 publication reports included in 8 systematic reviews. Although all these studies are randomized controlled trials, only 16% of them have posted structured results in ClinicalTrials.gov. The annotation task found 48 sample sizes, 116 group sizes, 53 populations, 99 intervention/control groups, and 270 outcomes. Terms that co-referred to the same entity are counted once. In the sentence-level evaluation, 3166 sentences in the training set were used to train the regression model, and 3404 sentences in the test set were used for evaluation. In the fragment-level evaluation, we extracted from the gold standard the top relevant sentences that contain at least a target element. The number of testing sentences per data element is as follows: Sample size/Group size: 20; population: 65; intervention/control: 149; and outcome: 124. Figure 5.3 demonstrated an example of system outputs which included top relevant sentences and recommended key phrases.

Table 5.4 summarizes the performance of four machine-learning models compared with abstract screening. Different models performed best for different data elements. BOT+Context performed best for Sample/Group size and population; BOT+Context+Semantic performed best for interventions; and BOT+Semantic performed



```
+Functional problems and catheter-related bacteraemia At access , injection problems (
difficult or impossible injection ) were less frequently recorded than aspiration problems (
dificult aspiration , incomplete flling of the Vacutainer® tube , or impossible aspiration ) .
 -impossible aspiration
 -functional problems
 -injection
 -catheter-related bacteraemia
+The primary outcome was the number of functional complications , which was defined as
easy injection , impossible aspiration at port access .
 -impossible aspiration
 -complications
 -easy injection
+The incidence rate of our primary outcome ( easy injection , impossible aspiration ) was
3.70 % ( 95 % CI 2.91 % -- 4.69 % ) and 3.92 % ( 95 % CI 3.09 % -- 4.96 % ) of accesses in the
normal saline and heparin groups , respectively .
 -impossible aspiration
 -the incidence rate
 -easy injection
+Secondary outcomes included all functional problems and catheter-related bacteraemia .
 -functional problems
 -catheter-related bacteraemia
+Before study start , the description of the primary outcome as a easy injection , impossible
aspiration ' was thoroughly explained to the nurses who were used to more vague terms ,
e.g. catheter occlusion or blockage .
 -impossible aspiration
 -catheter occlusion
 -blockage
 -easy injection
```

Figure 5.3 - Example of recommended sentences and key phrases of outcome element.

Table 5.4 - Performance comparison of various machine-learning models and with abstract screening.

	Bag-Of-Term (BOT)		BOT+Context		BOT+Semantic		BOT+Context+Semantic		Abstract Screening	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
Sample Size/Group Size	83.3	14.7	**91.7***	**15.1***	83.3	13.3	75.0	12.5	83.3	9.0
Population	95.8	43.7	**100.0**	**48.1***	100.0	43.2	95.8	46.2	100.0	27.4
Intervention/ Control	95.8	82.0	100.0	82.0	95.8	82.9	**100.0**	**85.2***	100.0	63.4
Outcome	73.7	87.3	73.5	85.2	**74.2***	**86.0***	73.1	86.0	50.7	55.4

* indicates statistically significant improvement over abstract screening (p<0.05).

best for outcomes. Context features and semantic features mildly contributed to improvement over the Bag-of-Term features. Recall of our best ML models was not significantly different from abstract screening in the extraction of sample/group size (recall +8.4%, p=0.38). Both methods achieved perfect recall on population and intervention/control elements. The best ML model significantly outperformed abstract screening in the extraction of outcome (recall +23.5%, p<0.001). Statistically significant improvements on precision were seen on all elements (SampleSize/GroupSize: +6.1%, p<0.001, Population: +20.7% , p=0.003, Intervention/Control: +21.8%, p<0.001, Outcome: 30.6%, p<0.001).

On fragment-level evaluation (Table 5.5), regular expression-matching approach achieved F-measure 90% on sample size and group size extraction. This confirms regex matching is the most commonly used and effective approach in extracting numeric values. For literal elements, each individual extraction method underperformed the combined method. Our combined extraction method achieved decent performance (avg f-measure: 83.4). Recall performance is better than precision (avg recall 93.2% vs. avg precision 75.6%).

Table 5.5 - Fragment-level performance of various extraction methods.

	Regex Matching			Concept-Mapping			Supplement Dictionary			Combined Method		
	R	P	F1	R	P	F1	R	P	F1	R	P	F1
SampleSize /GroupSize	92.5	87.5	90	NA	NA	NA	NA	NA	NA	92.5	87.5	90
Population	NA	NA	NA	87.7	67.7	76.4	48.5	47.7	48.1	93.8	71.5	81.2
Intervention /Control	NA	NA	NA	86.8	76.7	81.5	71.9	78.3	75	91.7	80	85.4
Outcome	15.8	17.3	16.5	54.2	54.9	49.7	60.1	63.2	61.6	94.1	75.2	83.6

5.5. Discussion

In this paper, we describe and evaluate a system that can help SR developers in extracting sample size and PICO values from full-text reports. The system is a contribution to previous work on PICO extraction, which primarily focused on extraction from study abstracts. For better support of SR development, our system is designed to operate on full-text PDF documents.

In sentence ranking, the best ML model varies for different data elements, which highlights the need to have an optimized ML model for each data element. Using context features and semantic features improved the performance of ML models that use Bag-Of-Term features alone. In comparison with abstract+title screening, the recall in top-ranked sentences was not significantly better for the extraction of the SampleSize/GroupSize element (+8.4%, p=0.38). A significant improvement in recall was obtained for the Outcome element (+23.5%, p<0.001). Both the system and abstract+title screening achieved perfect recall on the Population and Intervention/Control elements. These findings confirm that the majority but not all information can be found in the abstract and title of the studies. Therefore, information extraction systems supporting the development of systematic reviews need to be able to operate on full-text reports to maximize the comprehensiveness of data extraction and to comply with systematic review development requirements [16].

In addition, the system significantly improved precision for all data elements. Better precision corresponds to a higher number of relevant sentences in the top-ranked list. In full-text documents, information can be repeated in multiple sections. The ML system did a better job in collecting repeated relevant sentences, which offers reviewers multiple sources to validate the extraction results. In summary, we demonstrated the superiority of using an ML approach in prioritizing full-text sentences over the manual abstract+title screening. The technique has the potential to replace abstract screening when searching for specific data elements.

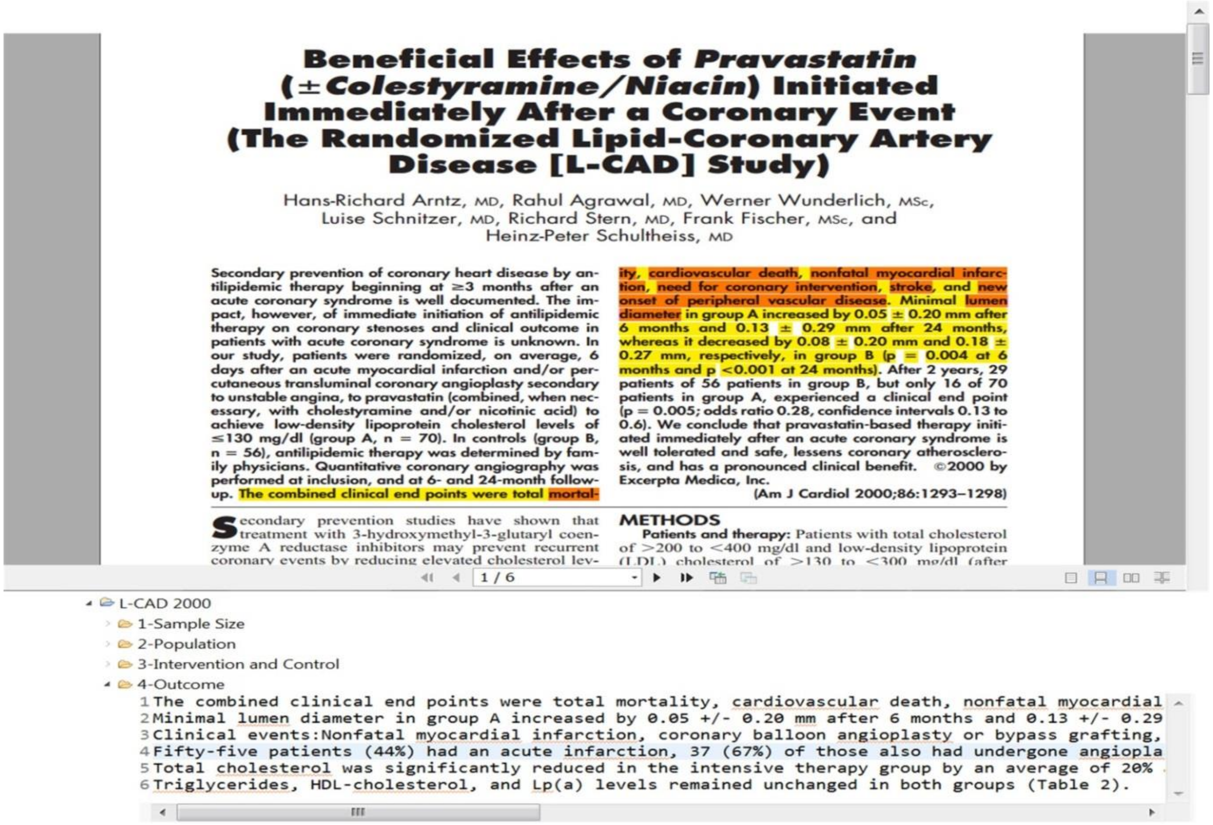
In fragment-level extraction, we proposed three extraction methods: regular expression matching, mapping to UMLS concepts, and element-specific dictionary. Regular expressions are most useful for extracting templates or numerical values. Designing and implementing a regular expression approach requires considerable manual work unless regular expression learning techniques are effectively applied [23, 24]. Mapping text to UMLS concepts is one of the extraction methods commonly used in clinical and biomedical NLP studies [25-27]. MetaMap tends to perform well in the recognition of texts that can be mapped to medical terms. However, there are more than 3 million UMLS concepts (2015AA Release), and classifying them to the data element of interest is challenging. In this study, we employed the simple semantic type restriction approach to categorize concepts to a specific element. There are other approaches to categorize UMLS concepts, such as heuristics using UMLS concept relationships [28], semantic distribution [29], or machine-learning [30], which demand additional investigation and optimization to perform well on sample size and PICO elements. The last approach (element-specific dictionary) was motivated from the fact that the UMLS Metathesaurus might not fully cover medical terms for specific extraction needs. Element-specific terms are needed to complement the UMLS concept-mapping approach. In this study, we utilized true-positive terms that appeared in the training set and that achieved a good coverage (60% recall) on the test set. The experiment results showed that an ensemble approach combining the three methods performed better than any of the individual methods. For PICO elements, the system's recall was better than precision (93.2% vs. 75.6%), which meets our performance goal. Recall is often more important for semiautomated extraction, since humans are effective at judging whether recommended phrases are true positives but tend to miss information while screening large textual contents.

5.5.1. Proposal of Computer-Assisted Data Extraction

The proposed data extraction pipeline could be closely integrated with a PDF reader interface for the optimal support of the data extraction process. The enhanced PDF reader has the ability to select top relevant sentences suggested by a text-mining system, and automatically navigate to the sentence location in the document. This way, reviewers have at least two strategies to conduct the computer-assisted data extraction: (1) Rely on the computer-suggested sentences and follow the normal full-text screening if a relevant sentence cannot be found; or (2) follow the normal extraction process and use the tool for verification. We suggest using the first strategy for data elements for which a conclusion can be reached from one or two relevant sentences, such as sample size, intervention, and population. For elements that are often mentioned in multiple sentences (e.g., clinical outcome), the second strategy might be more useful to reduce human errors. Figure 5.4 depicts a proposal for an enhanced PDF reader that can assist the manual extraction process.

5.5.2. Limitations

This study focused on sample size and PICO elements that were commonly reported in randomized controlled trial studies. There are data elements suggested by the Cochrane Collaboration that were not covered. Some of those elements, such as funding sources, study design, and study authors, can be easily retrieved from Medline metadata. Data elements such as age, sex distribution, and number of participants in each group are usually reported in table structure, which requires a specialized table-parsing algorithm. Other



Beneficial Effects of *Pravastatin* (±*Colestyramine/Niacin*) Initiated Immediately After a Coronary Event (The Randomized Lipid-Coronary Artery Disease [L-CAD] Study)

Hans-Richard Arntz, MD, Rahul Agrawal, MD, Werner Wunderlich, MSc,
Luise Schnitzer, MD, Richard Stern, MD, Frank Fischer, MSc, and
Heinz-Peter Schultheiss, MD

Secondary prevention of coronary heart disease by antilipidemic therapy beginning at ≥3 months after an acute coronary syndrome is well documented. The impact, however, of immediate initiation of antilipidemic therapy on coronary stenoses and clinical outcome in patients with acute coronary syndrome is unknown. In our study, patients were randomized, on average, 6 days after an acute myocardial infarction and/or percutaneous transluminal coronary angioplasty secondary to unstable angina, to pravastatin (combined, when necessary, with cholestyramine and/or nicotinic acid) to achieve low-density lipoprotein cholesterol levels of ≤130 mg/dl (group A, n = 70). In controls (group B, n = 56), antilipidemic therapy was determined by family physicians. Quantitative coronary angiography was performed at inclusion, and at 6- and 24-month follow-up. The combined clinical end points were total mortality, cardiovascular death, nonfatal myocardial infarction, need for coronary intervention, stroke, and new onset of peripheral vascular disease. Minimal lumen diameter in group A increased by 0.05 ± 0.20 mm after 6 months and 0.13 ± 0.29 mm after 24 months, whereas it decreased by 0.08 ± 0.20 mm and 0.18 ± 0.27 mm, respectively, in group B (p = 0.004 at 6 months and p <0.001 at 24 months). After 2 years, 29 patients of 56 patients in group B, but only 16 of 70 patients in group A, experienced a clinical end point (p = 0.005; odds ratio 0.28, confidence intervals 0.13 to 0.6). We conclude that pravastatin-based therapy initiated immediately after an acute coronary syndrome is well tolerated and safe, lessens coronary atherosclerosis, and has a pronounced clinical benefit. ©2000 by Excerpta Medica, Inc.

(Am J Cardiol 2000;86:1293–1298)

Secondary prevention studies have shown that treatment with 3-hydroxymethyl-3-glutaryl coenzyme A reductase inhibitors may prevent recurrent coronary events by reducing elevated cholesterol lev-

METHODS

Patients and therapy: Patients with total cholesterol of >200 to <400 mg/dl and low-density lipoprotein (LDL) cholesterol of >130 to <300 mg/dl (after

- L-CAD 2000
  - 1-Sample Size
  - 2-Population
  - 3-Intervention and Control
  - 4-Outcome

```
1 The combined clinical end points were total mortality, cardiovascular death, nonfatal myocardial
2 Minimal lumen diameter in group A increased by 0.05 +/- 0.20 mm after 6 months and 0.13 +/- 0.29
3 Clinical events:Nonfatal myocardial infarction, coronary balloon angioplasty or bypass grafting,
4 Fifty-five patients (44%) had an acute infarction, 37 (67%) of those also had undergone angiopla
5 Total cholesterol was significantly reduced in the intensive therapy group by an average of 20%
6 Triglycerides, HDL-cholesterol, and Lp(a) levels remained unchanged in both groups (Table 2).
```

Figure 5.4 - Proposed user interface of an enhanced PDF reader to assist the data extraction task.

elements, such as detailed inclusion/exclusion criteria, study duration, randomization method, and blinding, can be extracted with an extension of our proposed method. There are other machine-learning models, such as linear regression, multilayer perceptron, and Gaussian processes, that were not evaluated in this study and could be investigated in future research. For comparison of feature groups, we only used support vector machine regression, given its popularity and effectiveness in data mining research [31-33].

5.5.3. Future Work

To fully support the vision of computer-assisted data extraction, automated systems need to support diverse systematic review data elements and have an interactive user interface well integrated into the traditional data extraction workflow. Additional innovative approaches in sentence ranking and phrase extraction can be explored to find optimal strategies for each individual data element.

5.6. Conclusion

We presented a system that can help human reviewers in extracting sample size and PICO values from full-text PDF reports. The system is composed of two main components: sentence ranking and key phrase extraction. In sentence ranking, we demonstrated that using a machine-learning classifier to prioritize full-text sentences performed equally or better than the manual abstract screening approach. That highlights the potential of using a machine-learning approach to replace the traditional abstract screening in searching for specific information. For fragment-level extraction, we showed that using an ensemble approach combining three different extraction methods improved extraction performance. The system is a key component for a computer-aided data extraction application. Future research is needed to integrate the data extraction system with an effective and usable user interface.

5.7. Acknowledgments

This study was supported in part by grants 1R01LM011416-02, 4R00LM011389-02 and R01-LM010981 from the National Library of Medicine.

5.8. References

[1] D.J. Cook, C.D. Mulrow, R.B. Haynes, Systematic reviews: synthesis of best evidence for clinical decisions, Ann. Intern. Med. 126(5) (1997) 376-380.

[2] S. Tong, D. Koller, Restricted bayes optimal classifiers, AAAI/IAAI, 2000, 658-664.

[3] D. Dahlmeier, H.T. Ng, Domain adaptation for semantic role labeling in the biomedical domain, Bioinformatics 26(8) (2010) 1098-1104.

[4] Cochrane Collaboration, Directors' reports and financial statements, 2013.

[5] K.G. Shojania, M. Sampson, M.T. Ansari, J. Ji, S. Doucette, D. Moher, How quickly do systematic reviews go out of date? A survival analysis, Ann. Intern. Med. 147 (4) (2007) 224–233.

[6] P. Bragge, O. Clavisi, T. Turner, E. Tavender, A. Collie, R.L. Gruen, The global evidence mapping initiative: scoping research in broad topic areas, BMC Med. Res. Methodol. 11 (1) (2011) 92.

[7] A.P. Jones, T. Remmington, P.R. Williamson, D. Ashby, R.L. Smyth, High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews, J. Clin. Epidemiol. 58 (7) (2005) 741–742.

[8] P.C. Gotzsche, A. Hrobjartsson, K. Maric, B. Tendal, Data extraction errors in meta-analyses that use standardized mean differences, JAMA : The journal of the American Medical Association 298(4) (2007) 430-437.

[9] F. Boudin, J.-Y. Nie, J.C. Bartlett, R. Grad, P. Pluye, M. Dawes, Combining classifiers for robust PICO element detection, BMC Med. Inform. Decis. Mak. 10 (2010) 29.

[10] D.H. Wolpert, Stacked generalization, Neural Networks 5(2) (1992) 241-259.

[11] K.-C. Huang, I.J. Chiang, F. Xiao, C.-C. Liao, C.C.-H. Liu, J.-M. Wong, PICO element detection in medical text without metadata: are first sentences enough?, J. Biomed. Inform. 46(5) (2013) 940-946.

[12] S.R. Eddy, Hidden markov models, Curr. Opin. Struc. Biol. 6(3) (1996) 361-365.

[13] D. Demner-Fushman, J. Lin, Answering clinical questions with knowledge-based and statistical techniques, Comput. Linguist. 33(1) (2007) 63-103.

[14] M. Ware, M. Mabe, An Overview of Scientific and Scholarly Journal Publishing, The STM Report, 2009.

[15] M.J. Hansen, N.O. Rasmussen, G. Chung, A method of extracting the number of trial participants from abstracts describing randomized controlled trials, J. Telemed. Telecare 14(7) (2008) 354-358.

[16] J.P. Higgins, S. Green, Cochrane Handbook for Systematic Reviews of Interventions, Wiley Online Library, 2008.

[17] S. Kiritchenko, B. de Bruijn, S. Carini, J. Martin, I. Sim, ExaCT: automatic extraction of clinical trial characteristics from journal publications, BMC Med. Inform. Decis. Mak. 10 (2010) 56.

[18] B. de Bruijn, S. Carini, S. Kiritchenko, J. Martin, I. Sim, Automated information extraction of key trial design elements from clinical trial publications, AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium, 2008, 141-145.

[19] C.D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S.J. Bethard, D. McClosky, The Stanford CoreNLP Natural Language Processing Toolkit, In: ACL (System Demonstrations), 2014, 55-60.

[20] M.A.M. García, R.P. Rodríguez, L.E.A. Rifón, Biomedical literature classification using encyclopedic knowledge: a Wikipedia-based bag-of-concepts approach, PeerJ 3 (2015) e1279.

[21] A.T. McCray, A. Burgun, O. Bodenreider, Aggregating UMLS semantic types for reducing conceptual complexity, Stud. Health Technol. Inform. 84 (Pt 1) (2001) 216-220.

[22] A.R. Aronson, Metamap: Mapping text to the umls metathesaurus, Bethesda, MD: NLM, NIH, DHHS, 2006, 1-26.

[23] D.D. Bui, Q. Zeng-Treitler, Learning regular expressions for clinical text classification. J. Am. Med. Inform. Assoc. 21(5) (2014) 850-857.

[24] Y. Li, R. Krishnamurthy, S. Raghavan, S. Vaithyanathan, H.V. Jagadish, Regular expression learning for information extraction, Proceedings of the Conference on Empirical Methods in Natural Language Processing. Honolulu, Hawaii, Association for Computational Linguistics, 2008, 21-30.

[25] C. Soguero-Ruiz, K. Hindberg, J. Rojo-Alvarez, S.O. Skrovseth, F. Godtliebsen, K. Mortensen, A. Revhaug, R.-O. Lindsetmo, K.M. Augestad, R. Jenssen, Support Vector Feature Selection for Early Detection of Anastomosis Leakage from Bag-of-Words in Electronic Health Records, IEEE J. Biomed. Health Inform. (2014).

[26] R. Xu, Y. Hirano, R. Tachibana, S. Kido, Classification of diffuse lung disease patterns on high-resolution computed tomography by a bag of words approach, Med. Image Comput. Comput. Assist. Interv. 14(Pt 3) (2011) 183-90.

[27] D.D. Bui, S. Jonnalagadda, G. Del Fiol, Automatically finding relevant citations for clinical guideline development, J. Biomed. Inform. (2015).

[28] L.I. Kuncheva, Combining pattern classifiers: methods and algorithms, 2004.

[29] Apache PDFBox, A Java PDF Library <https://pdfbox.apache.org/>

[30] J. Beel, B. Gipp, A. Shaker, N. Friedrich, SciPlore Xtract: Extracting Titles from Scientific PDF Documents by Analyzing Style Information (Font Size), Lect. Notes Comput. Sc. (2010) 413-416.

[31] U. Schäfer, B. Kiefer, Advances in deep parsing of scholarly paper content, Springer (2011).

[32] N. Japkowicz, S. Stephen, The class imbalance problem: A systematic study, Intell. Data Anal. 6(5) (2002) 429-449.

[33] J. Van Hulse, T.M. Khoshgoftaar, A. Napolitano, Experimental perspectives on learning from imbalanced data, Proceedings of the 24th international conference on Machine learning 2007, ACM, 2007, 935-942.

CHAPTER 6

DISCUSSION

6.1 Summary

Systematic review (SR) development typically relies on human experts to find, appraise, and synthesize the best available evidence to produce reliable answers to scientific questions. Due to the high reliance on a manual-labor-intensive process, systematic review development is affected by human constraints such as limited time, scarce resources, inconsistencies, errors, and biases. In the 21st century, when computer technology has become a part of everyday life, it is worth investigating whether systematic review development can be enhanced through the use of computing resources in order to meet the needs of the health care community.

In this dissertation, we explored the feasibility of computer-aided solutions to enhance the traditional systematic review process. Two SR steps, literature search and data extraction, were selected as the focus for technology improvement since they are standard, labor-intensive, and highly associated with the quality of the final outcomes. The technology interventions are not intended to replace the human process, but rather provided as an option to integrate with the traditional workflow, aiming to enhance productivity and reduce errors.

The proposed systems were designed to be applied in a real work setting. For system implementation, we leveraged technology from basic computer science research:

information retrieval, machine-learning, and information extraction. To enhance the practicality of the evaluations, we developed gold standards from real-world systematic reviews such as ACCF/AHA practice guidelines and Cochrane reviews, and adapted evaluation metrics to reflect the expectation of system performance in actual practice. In the course of three studies, we proposed and realized innovative computer-aided methods to assist human experts perform literature database search and extract relevant information from original study reports.

First, we developed an information retrieval method based on extending PubMed, and optimized to retrieve relevant high-quality studies for clinical guideline topics (Chapter 3). In comparison with PubMed, the system significantly improved recall with non-significant loss on precision. The proposed scientific quality ranking outperformed the standard PubMed ranking and a general purpose machine-learning classifier. Due to better recall and ranking performance, the system has the potential to replace PubMed as a tool for the systematic search of relevant articles for systematic review development. Searchers can use the system as a starting point to further expand the search, or as a reference list to complement the results of manual search.

Second, we proposed the rule-based multipass sieve algorithm to help extract and categorize PDF text snippets into high-level document structure metadata (Chapter 4). The algorithm can serve as a preprocessing step in any text-mining pipeline that attempts to unlock information from PDF documents. In this study, we demonstrated that the rule-based multipass sieve approach is more effective than a popular machine-learning approach in categorizing PDF document structure. Furthermore, filtering nonprose texts such as publication metadata and semistructures improved accuracy and efficiency of an information extraction system.

Third, we developed a system to assist in the data extraction process (Chapter 5). The system accepts PDF publication reports as input, prioritizes sentences for specific data elements, and recommends key phrases in the sentence. In this study, we demonstrated that using a machine-learning classifier to prioritize full-text sentences performed equally or better than the manual abstract screening approach. This finding encourages the use of the machine-learning system in sentence prioritization to replace the traditional abstract screening approach when searching for specific data elements. There are also other useful findings. For example, using contextual and semantic features improved the ML model that uses Bag-Of-Term features alone. Combining several individual extraction methods might have a cumulative improvement effect.

6.2 Limitations

This dissertation research has several limitations. First, we selected systematic review topics from the cardiovascular domain, so it is unknown whether the findings are generalizable to other domains. Yet, since our methods did not use knowledge specific to the cardiovascular domain, they are expected to generalize to other domains with minimal adaptation efforts. The machine-learning approach (Chapter 5) requires developing the training corpus for other domains or employing domain adaptation approaches [1]. Second, the frequently used approach, concept-mapping using MetaMap, often results in mapping generic concepts that are not useful for the NLP task. In this study we resolved the issue using a manually (Chapter 3) or automatically (Chapter 5) generated stop list. However, this approach is sometimes not optimal and not generalizable, which demands additional investigation to enhance concept-mapping for specific information needs. Third, when evaluating impacts of PDF text classification on an information extraction system (Chapter 4), we only tested one use case focused on filtering publication metadata and semistructured texts before submission to the outcome extraction system. There are other meaningful uses of semistructured texts for information extraction. However, they require specialized table/figure parsing algorithms, which go beyond the scope of this dissertation research. Fourth, the information extraction system (Chapter 5) only focused on sample size and PICO elements. There are other data elements needed in systematic reviews that were not covered in this research. However, the proposed framework can be extended to almost all elements that are embedded in narrative texts. Information unlocked in tables or figures typically requires different extraction techniques, such as regular expression matching, image processing, and optical character recognition.

6.3 Future Research

This dissertation research could lead to several directions for future studies, as summarized below:

- Future studies could take advantage of information retrieval techniques (query expansion and scientific quality ranking) developed in this research, and enhance and adapt to diverse clinical evidence summaries: clinical guidelines developed by medical societies [2], Cochrane systematic reviews [1], drug effectiveness reviews [3], and ad-hoc reviews.
- The PDF text classification algorithm could be investigated in other text-mining research such as information retrieval, document classification, and information extraction. Those branches of research typically select texts from convenient

sources (e.g., HTML pages, EMR, unformatted texts). Digital documents like PDFs are popular information sources for humans, however rarely used in text-mining research.

- There is also a need to extend the information extraction system to diverse types of systematic review data elements, and to extract information from semistructures. There are ongoing studies on table structure recognition [4, 5], which might be the solution to the semistructures extraction problem; however, further work is required to verify and adapt to the needs of systematic review development.
- The proposed techniques have the potential to apply to the systematic review updating process. The SR updating process attempts to identify and appraise evidence that is clinically significant enough to induce changes to an existing review. While the retrieval algorithm needs further improvement to be able to compare and recognize the knowledge gaps, the uses of automated citation retrieval and computer-aided data extraction to the updating process is essentially similar to the full SR process.
- Computer-aided approaches in both literature search and data extraction tasks could be investigated in extrinsic studies in simulated or real work settings. This is an important step to assess the technology impacts in real task performance. Randomized controlled trial studies could be conducted to measure the impact of the technology intervention on productivity, time savings, error reduction, and user acceptance.

6.4 References

[1] D. Dahlmeier, H.T. Ng, Domain adaptation for semantic role labeling in the biomedical domain. Bioinformatics 26(8) (2010) 1098-1104.

[2] L.A. Stokowski, National Guideline Clearinghouse, Adv. Neonatal Care 5(5) (2005) 239.

[3] M.S. McDonagh, D.E. Jonas, G. Gartlehner, A. Little, K. Peterson, S. Carson, et al, Methods for the drug effectiveness review project, BMC Med. Res. Methodol. 12(1) (2012) 140.

[4] F. Kboubi, A.H. Chabi, M.B. Ahmed, Table recognition evaluation and combination methods, Document Analysis and Recognition 2005 Proceedings Eighth International Conference on 2005 IEEE, 2005, 1237-1241.

[5] E. Oro, M. Ruffolo (Eds.), PDF-TREX: An approach for recognizing and extracting tables from PDF documents, Document Analysis and Recognition, 2009 ICDAR'09 10th International Conference on, IEEE, 2009.