MEDICAL KNOWLEDGE ACQUISITION USING BIOMEDICAL KNOWLEDGE RESOURCES FOR DISEASE-SPECIFIC ONTOLOGIES

by

Liqin Wang

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 2017

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Liqin Wang**

has been approved by the following supervisory committee members:

Peter J. Haug	, Chair	**03/14/2017** Date Approved
Bruce E. Bray	, Member	**03/14/2017** Date Approved
Guilherme Del Fiol	, Member	**03/13/2017** Date Approved
Olivier Bodenreider	, Member	**03/07/2017** Date Approved
Wendy W. Chapman	, Member	**03/13/2017** 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

Disease-specific ontologies, designed to structure and represent the medical knowledge about disease etiology, diagnosis, treatment, and prognosis, are essential for many advanced applications, such as predictive modeling, cohort identification, and clinical decision support. However, manually building disease-specific ontologies is very labor-intensive, especially in the process of knowledge acquisition. On the other hand, medical knowledge has been documented in a variety of biomedical knowledge resources, such as textbook, clinical guidelines, research articles, and clinical data repositories, which offers a great opportunity for an automated knowledge acquisition. In this dissertation, we aim to facilitate the large-scale development of disease-specific ontologies through automated extraction of disease-specific vocabularies from existing biomedical knowledge resources. Three separate studies presented in this dissertation explored both manual and automated vocabulary extraction. The first study addresses the question of whether disease-specific reference vocabularies derived from manual concept acquisition can achieve a near-saturated coverage (or near the greatest possible amount of disease-pertinent concepts) by using a small number of literature sources. Using a general-purpose, manual acquisition approach we developed, this study concludes that a small number of expert-curated biomedical literature resources can prove sufficient for acquiring near-saturated disease-specific vocabularies. The second and third studies introduce automated techniques for extracting disease-specific vocabularies from both

MEDLINE citations (title and abstract) and a clinical data repository. In the second study, we developed and assessed a pipeline-based system which extracts disease-specific treatments from PubMed citations. The system has achieved a mean precision of 0.8 for the top 100 extracted treatment concepts. In the third study, we applied classification models to reduce irrelevant disease-concepts associations extracted from MEDLINE citations and electronic medical records. This study suggested the combination of measures of relevance from disparate sources to improve the identification of true-relevant concepts through classification and also demonstrated the generalizability of the studied classification model to new diseases. With the studies, we concluded that existing biomedical knowledge resources are valuable sources for extracting disease-concept associations, from which classification based on statistical measures of relevance could assist a semi-automated generation of disease-specific vocabularies.

To Jinsong, Siyu, and Yiran.

TABLE OF CONTENTS

LIST OF FIGURES

Figures

LIST OF TABLES

Tables

ACKNOWLEDGEMENTS

I would like to thank some of the many people who have helped me through the long process of graduate school. First of all, I would like to express my deepest gratitude to my advisor, Peter J. Haug, for his unending support and guidance over the last six years. He is an excellent mentor, a remarkable scientist, and a great person, whom I will always admire.

Special thanks also to the other members of the dissertation committee, Guilherme Del Fiol, Bruce E. Bray, Olivier Bodenreider, and Wendy Chapman, for their many contributions to this work. Dr. Del Fiol has provided strong support to this dissertation work. He has led me to be critical in thinking and writing as a scientific researcher through many discussions and manuscript revisions. Dr. Bray served as an advisor for my first-year graduate school and a committee member thereafter. He has contributed countless hours to my first study as an expert and is the very person who filled up my gap of medical knowledge and made me confident in applying informatics to the medicine. Dr. Bodenreider was an excellent mentor when I interned at the U.S. National Library of Medicine. He is knowledgeable, passionate, and always helpful with my questions regarding the medical terminologies and vocabularies. I also appreciate Dr. Chapman for the mentorship I received when I interned at UC San Diego and also her devotion to our department as the chair. She is collaborative, cares for people, and a great leader but with whom you also want to be a friend.

I wish to express thanks to Qing Zeng-Treitler for serving as a member of my committee before she moved to the George Washington University.

Specially thanks also to my dear colleagues and friends, Jeffery Ferraro, Jianlin Shi, Mike Conway, Ming-Chin (Mark) Lin, Shan He, Susan Matney, Rick Bradshaw, Jingran Wen, Jiantao Bian, and Joseph Plasek for their help, advice, and many discussions encountered over the past years.

I wish to thank the administrative staff at the department of Biomedical Informatics, especially JoAnn Thompson, Barbara Saffel, and Kate Handziuk, for their help over the past years.

I wish to express my special thanks to the members of the Homer Warner Center at Intermountain Healthcare, particularly Susan Rea, Bart Dodds, and Philips Jackson who provided help in querying the AHR database at Intermountain Healthcare; also to the administrative staff, particularly Jason Gagner and Rose Wirthlin, for help and kindness while working at Intermountain Healthcare as a student.

I acknowledge the Homer Warner scholarship from the Department of Biomedical Informatics for my first-year graduate school and the financial support from the Homer Warner Research Center at Intermountain Healthcare for the past six years.

Lastly, I am immensely grateful to my parents and husband for their unconditional love, encouragement, and support, and to my two lovely daughters for their healthy distraction from school and for keeping me enjoying the happy moments of life.

CHAPTER 1

INTRODUCTION

The work described below represents an effort to develop tools and processes that can ease the work necessary to develop collections of disease-specific medical concepts that will support the curation of computer-accessible medical ontologies. The medical knowledge managed in ontologies using these concepts will contribute to promote efficient and effective patient-centered care using automated health information systems

1.1 The Need of Disease-Specific Medical Knowledge

The following examples help to recognize a variety of informatics areas demanding **disease-specific medical knowledge** for supporting better informed healthcare and research activities.

1.1.1 Case 1: Physicians Facing Information Overload

A cardiologist is seeing his patient who has congestive heart failure (CHF) but also other problems, such as diabetes mellitus, depression, and rheumatoid arthritis. He has to manually collect and distill the information relevant to the CHF as the system has stored a large number of records for this individual. He desires a system that could automatically collect, distill, and summarize information that is relevant to the problem of the current visit.

1.1.2 Case 2: Consumers' Online Information Seeking

A middle-age lady was recently diagnosed with hepatic cirrhosis. Before her upcoming visit to her healthcare provider, she would like to know the available means of treatment and compare them regarding the cons and pros. She started searching on the Internet with the broad words "treatment" plus "hepatic cirrhosis", but ended up with frustration by the broad, irrelevant, or even incorrect information received online. She wishes for automated assistance in forming more precise searches directed at trusted sources.

1.1.3 Case 3: Researchers' Analysis of Healthcare Data

A data scientist aims to build a predictive model for lung cancer prognosis. Facing massive amounts of both structured and unstructured data, he has to consult medical experts throughout the entire process of data analysis, including understanding and cleaning the data, extracting important data elements as inputs for building the predictive model, and conducting comprehensive data analysis and evaluation. He wishes there were a well-developed disease model which would assist the entire process and reduce the dependence on medical experts.

1.2 Objectives and Hypothesis

The primary goal of this dissertation is to enable the large-scale development of disease-specific ontologies which could serve as a fundamental component of those advanced clinical applications (e.g., problem-oriented summarization of medical records, question answering, diagnostic and predictive modeling) for better informed healthcare. More specifically, we have studied methods to extract disease-specific, assertional

medical knowledge from existing biomedical knowledge resources for the development of disease-specific ontologies. Disease-specific ontologies are computer-understandable and human-readable knowledge bases that have been designed to explicitly support representations of the knowledge of disease etiology, diagnosis, treatment and prognosis for each kind of disease. The underlying assumption behind disease-specific ontologies is the belief that they can be useful for the representation, sharing, and computation of domain-specific medical knowledge. The main research question of this study is "can disease-specific vocabularies required for building disease-specific ontologies be extracted from existing biomedical knowledge resources."

In the three studies to be presented, three specific research questions were explored:

1. Is it practical to use only a small number of expert-curated textual knowledge sources to acquire disease-specific vocabularies that reach a saturated coverage (Chapter 3)?
2. Is it feasible to automatically acquire disease-specific treatment vocabularies from the biomedical literature using a pipeline-based approach (Chapter 4)? (Specifically, we hypothesize that *there is no difference in precision at the top 100 extracted concepts among the rankings produced by the four measures of relevance in the pipeline-based approach*. We also hypothesize that *there is no difference in precision at the top 100 extracted concepts among the rankings produced by the pipeline-based system and two baseline approaches.)*
3. Can classifiers, generated using machine learning techniques, be used to reduce the manual effort necessary to review noisy collections of proposed disease-

specific concepts extracted from both biomedical literature and clinical data repositories (Chapter 5)? (Specifically, we hypothesize that *using the features (e.g., measures of relevance) from both the biomedical literature and clinical data repositories would improve the classifiers' performance compared to using features from the individual sources*. We also hypothesize that *the classifiers initially built for specific diseases would be generalizable to other disease(s)*.)

1.3 Rationale for Analysis

Computers have been introduced to the medical field to assist healthcare activities since the 1950s [1]. As the complexity of the domain of medicine continuously increases, comprehensive computer-understandable knowledge bases (KBs) are needed. The term "knowledge bases" can refer to different things such as vocabularies, ontologies, collections of rules, semantic networks, or probabilistic models. In this dissertation, we choose the ontology as a medium to represent a kind of medical knowledge. Ontologies represent an explicit specification of a conceptualization [2] which allow sharing and reuse and have been commonly used by the healthcare informatics communities to represent medical semantics. In addition, among the kinds of medical knowledge, we discerned that disease-specific medical knowledge (i.e., disease's etiology, diagnosis, therapy, and prognosis) is particularly important. As a clinician, a comprehensive understanding of the disease in all its different aspects can lead to better medical practice and desired patient outcomes. Similarly, having such kinds of knowledge available to the computer can empower and support healthcare activities through many advanced applications. Specifically, several disease-specific ontologies have been demonstrated to be useful for clinical applications such as diagnostic modeling [3], reminder systems [4],

and text annotation [5,6]. We assume that they will also be useful for other applications, like problem-oriented summaries of patient EHRs [7–9], clinical question-answering [10], query expansion [11], and treatment recommendation [12].

However, the development of this kind of ontologies is very labor-intensive. One of the main challenges for the large-scale development of disease-specific ontologies is the acquisition of disease-relevant medical knowledge. It is expensive to build ontologies that rely heavily on human experts, and this effort becomes impractical when building ontologies for thousands of diseases. As the majority of the medical knowledge is well documented in the biomedical knowledge resources, such as textbooks, clinical guidelines, research articles, and clinical notes, the sources offer great opportunities for an automated knowledge extraction. Therefore, we aimed to extract disease-specific medical knowledge from existing biomedical knowledge resources using approaches in which the involvement of human experts or knowledge engineers could be minimal.

1.4 Overview of the Dissertation

Chapter 2 of this dissertation provides the background for the body of research and contains two parts. Part one introduces the disease-specific ontologies and potential applications. Part two describes state-of-the-art techniques for medical knowledge acquisition from existing knowledge resources.

Chapter 3 of this dissertation investigates the manual acquisition of disease-specific reference vocabularies from expert-curated documents (e.g., textbooks, clinical practice guidelines, and journal articles) [13]. We described a complete process of manual acquisition including document selection, manual annotation and adjudication, mapping, and assessment of vocabulary saturation.

In Chapter 4 of this dissertation, we develop and assess a pipeline-based system which automatically extracts disease-specific treatments from PubMed citations. The research question is addressed in Chapter 4 with a detailed description of a pipeline-based vocabulary extraction approach and the analysis of automated extracted results with a comparison to the manual acquired reference vocabularies from Chapter 3. Two corresponding hypotheses were tested.

Chapter 5 of this dissertation describes an effort to solve a challenge remaining from several prior studies of knowledge extraction, where the automated generated vocabularies from the biomedical literature and electronic medical records have low signal-to-noise ratio, and therefore require considerable manual review and selection. We tested several classifications models to automatically determine the relevance of those extracted concepts to the disease of interest. The research question is addressed and two corresponding hypotheses were assessed.

Chapter 6 summarizes our findings from three studies, and discusses the limitations, range of applicability, and future directions.

1.5 References

[1] Collen MF. Origins of medical informatics. West J Med 1986;145:778–85.

[2] Gruber TR. A translation approach to portable ontology specifications. Knowl Creat Diffus Util 1993;5:199–220.

[3] Haug PJ, Ferraro JP, Holmen J, Wu X, Mynam K, Ebert M, et al. An ontology-driven, diagnostic modeling system. J Am Med Inform Assoc 2013;20:e102-10.

[4] Chalortham N, Buranarach M, Supnithi T. Ontology development for type II diabetes mellitus clinical support system. Proc 4th Int Conf Knowl Inf Creat Support Syst 2009.

[5] Younesi E, Malhotra A, Gündel M, Scordis P, Kodamullil AT, Page M, et al. PDON: Parkinson's disease ontology for representation and modeling of the Parkinson's disease knowledge domain. Theor Biol Med Model 2015;12:20.

[6] Malhotra A, Younesi E, Gündel M, Müller B, Heneka MT, Hofmann-Apitius M. ADO: a disease ontology representing the domain knowledge specific to Alzheimer's disease. Alzheimer's Dement 2014;10:238–46.

[7] Zeng Q, Cimino JJ. A knowledge-based, concept-oriented view generation system for clinical data. J Biomed Inform 2001;34:112–28.

[8] Feblowitz JC, Wright A, Singh H, Samal L, Sittig DF. Summarization of clinical information: A conceptual model. J Biomed Inform 2011;44:688–99.

[9] Mccoy AB, Wright A, Laxmisan A, Ottosen MJ, Mccoy JA, Butten D, et al. Development and evaluation of a crowdsourcing methodology for knowledge base construction: Identifying relationships between clinical problems and medications. J Am Med Inform Assoc 2012;19:713-18.

[10] Demner-fushman D, Lin J. Answering clinical questions with knowledge-based and statistical techniques. Computational Linguistics 2007; 33(1):63-103.

[11] Liu Z, Chu WW. Knowledge-based query expansion to support scenario-specific retrieval of medical free text. Inf Retr Boston 2007;10:173–202.

[12] Kazdin AE. Evidence-based treatment and practice: new opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. Am Psychol 2008;63:146–59.

[13] Wang L, Bray BE, Shi J, Del Fiol G, Haug PJ. A method for the development of disease-specific reference standards vocabularies from textual biomedical literature resources. Artif Intell Med 2016.

[14] Wang L, Wang L, Zhang M, Conway M, Haug P, Chapman W. Using cKASS to

facilitate knowledge authoring and sharing for syndromic surveillance. Emerg Health Threats J 2011:11147.

CHAPTER 2

BACKGROUND

2.1 Disease-Specific Ontologies

2.1.1 What Is Ontology?

Ontology is originally defined 'as the branch of metaphysics which investigates and explains the nature of all things or existences.' In the world of information science, the view of ontology is somewhat narrower. A classic definition of ontology was given by Gruber [1] that ontology is 'an explicit, formal specification of a shared conceptualization of a domain of interest'. To expand this definition, the *conceptualization* is 'an abstract, simplified view of the world that we wish to present for some purpose'[2], while the *specification* is 'the representation of this conceptualization in a concrete form'.

In the aspect of conceptualization, Noy [3] provided details about what is inside of an ontology. Four main components were defined: concepts, properties, restrictions, and instances. A concept represents a set or class of entities or 'things' within a domain. For instance, 'car' is a class which has subclasses like 'SUV' and 'Minivan'. An instance of the 'car' would be the car that you drive to work or home. 'Car' has properties, such as 'door', 'manufacturer', 'window', and 'wheels', and also restrictions, such as a car 'has four wheels'. In the process of specification, one goal is to encode the classes of entities

in the domain of interest with relations, properties, and restrictions, and organize them using semantic structure. Many ontology specification languages have been developed, among which several are very popular including KIF (Knowledge Interchange Format) [4], OWL (Web Ontology Language) [5], RDF+RDF(S) [6], and DAML+OIL [7]. In the meantime, a number of tools for developing and maintaining ontologies were also developed, such as Ontolingua, WebOnto, WebODE, Protégé, OntoEdit, etc. [8,9] Each of the tools or languages has its own strengths and weakness; therefore, the choice of the tools and languages are dependent on the users' needs.

Ontologies have been made popular in many areas, for example, knowledge representation [10,11], Semantic Web [12], and bioinformatics [13]. This could be attributed to several possible reasons [3]. First, ontology's enable the sharing of common knowledge among either people or software agents. Many ontologies have been developed for all kinds of domains or purposes, and stored in open repositories (e.g., Swoogle, NCBO BioPortal [14], OBO Foundry [15]). They can be easily accessed by and shared with those people who are interested in the same domain. The ontology, with the formal, explicit representation of the knowledge, also enables the reuse of domain knowledge. Computers that understand the languages of ontologies can parse the ontologies and read information from the ontologies. Further, ontology has the potential to enable many advanced applications, such as ontology-based reasoning [16], data integration [17], information retrieval [18], and question answering [19].

2.1.2 Ontologies in the Biomedical Domain

The history of "ontology" in the computer science domain starts from Gruber's definition of ontology in the early 1990s [1,2], and the history of the "ontology" in the

biomedical domain probably dates to the beginning of the 2000s with typical work including Gene Ontology [20], Foundational Model of Anatomy (FMA) [21], and Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [22]. Since then, ontologies have been widely developed and used in the biomedical domain [14,23]. Hundreds of biomedical ontologies have been developed, tracked from several well-known biomedical ontology repositories. For example, BioPortal (https://bioportal.bioontology.org/) [14] currently hosts up to 538 ontologies (accessed by Jan 2017). OBO foundry (http://www.obofoundry.org/) [24] also hosts over 140 biomedical ontologies (accessed by Jan 2017). Those biomedical ontologies have not only given the possibility of sharing and reuse of domain knowledge, but have also played a fundamental role in many biomedical informatics research projects, including the annotation of biomedical datasets, the biomedical literature and patient records, information retrieval, data integration, knowledge discovery, and decision support and reasoning [23,25].

2.1.3 Disease-Specific Ontologies and Applications

Each ontology has a scope of focus, which can be broad or narrow. For example, SNOMED CT broadly covers several clinical subdomains (e.g., anatomy, clinical finding, medications, procedure, etc.), while nurse administrator ontology (https://bioportal.bioontology.org/ontologies/ADMIN) has a narrow scope which typically focuses on nursing administration. In this dissertation, the scope of the ontologies we intend to construct are focused on specific diseases, which we called disease-specific ontologies (DSOs). Each DSO takes one disease or condition as a scope of focus. They are knowledge bases intended to structure and represent the medical

knowledge about disease etiology, diagnosis, treatment and/or prognosis, etc.

Although hundreds of biomedical ontologies were made available to the public, they primarily contain definitional knowledge which is considered as universally true. For example, in terms of disease-specific information, SNOMED CT essentially contains definitional knowledge about disease categories (e.g., myocardial infarction is_a cardiovascular disease) and body location (e.g., congestive heart failure finding_site_of cardiac ventricle). However, assertion knowledge, a kind of knowledge considered as true in a given context (e.g., aspirin treats headache), is usually missing from those existing ontologies. For example, SNOMED CT has little information specifying the relationships between drugs and diseases. As the importance of assertion knowledge is being increasingly recognized [26–28], we intend to integrate it with definitional knowledge to create DSOs to support different kind of applications, such as knowledge discovery, problem-oriented summarization of medical records, information retrieval, etc.

As we review the literature, researchers have made different attempts to develop DSOs for specific diseases and applied them in several informatics areas. For example, Haug *et al.* [29] developed a pneumonia ontology for their "ontology-driven diagnostic modeling system". The pneumonia ontology contains (1) the relationships among diseases, (2) the relationships between diseases and relevant observations, (3) the relationships between diseases and typical therapeutic interventions, and (4) the relationships between diseases and anticipated outcomes. By linking the concepts in the ontology to the data stored in Intermountain Healthcare's enterprise data warehouse (EDW), the ontology was used to identify diseased and non-diseased patients, and choose data elements to be useful in diagnosing pneumonia. Thereafter, the extracted data was

fed into the Bayes Network for building a diagnostic model. In another example, Malhotra *et al*. constructed an Alzheimer's disease ontology (ADO) which covers clinical, etiological, molecular, and cellular mechanism aspects of AD [30]. The authors of ADO also created two additional disease-specific ontologies using the same approach as they built the ADO. Younesi *et al.* built a Parkinson's disease ontology (PDO) to model the domain knowledge of Parkinson's disease [31]. This ontology covers the clinical aspects, etiology, model, neuropathology, disease categories, as well as associated familial diseases of Parkinson's disease. Malhotra *et al.* also created a multiple sclerosis ontology (MSO) which covers similar aspects as the PDO [32]. These three ontologies (i.e., ADO, PDO, and MSO) were all applied to the semantic mining of patient records and literature for effective retrieval and extraction of accurate disease-related information. In addition, Chalortham *et al.* developed an ontology for type II diabetes mellitus (DM), which contains DM-relevant information such as sign and symptom, treatment, assessment, and follow-ups activities [33]. The type II diabetes mellitus ontology was applied to a reminding system that provided patients' useful information to hospital providers [33] and also to identify a patient cohort from the EHR [34]. Similarly, El-Sappagh and Farman published a diabetes mellitus diagnosis ontology (DDO), which covers diabetes-related complications, symptoms, drugs, lab tests, etc. [35] Most of the development of these disease-specific ontologies mentioned above happened in the last several years. And most of these ontologies are currently available in the Bioportal.

Besides the actual development of DSOs for several specific diseases as mentioned above, there are some works closely related to the DSOs. Hadzic proposed generic human disease ontology (GHDO) that was designed for the representation of

knowledge regarding human disease [36]. It organizes the concepts of existing ontologies into four dimensions: disease types, symptoms, causes, and treatments. The top hierarchy of GHDO could be a useful guide for developing disease-specific ontologies. Bertaud-Gounot *et al.* argued that diagnostic criteria (such as signs and symptoms) should be included as part of the operational definition of diseases in the ontology for supporting the diagnostic modeling and reasoning [37]. Mendonca *et al.* proposed a model for accessing evidence from a digital library to answer clinical questions, where a major component is the knowledge bases that contain clinical concepts derived from clinical settings and relations (e.g., "is-caused-by") [38]. Another parallel work is the building of the diseases symptoms ontology [39] which aligns the disease ontology with the symptoms ontology, creating a core disease symptoms ontology.

With the merit of containing comprehensive disease-specific information in computer-understandable and human-readable format, DSOs may support other kinds of applications in addition to the ones researchers have explored (e.g., diagnostic modeling). This includes clinical question answering [40], query expansion [41], and therapy recommendation [42]. To support this argument, we provide detailed explanation below.

First of all, the disease-specific ontology may answer some disease-related questions. For some frequently asked questions [43], such as "what is the drug of choice for condition X?", "what is the cause of disease X?", and "what test is indicated in situation X?", ontologies could assist the clinicians to form well-built questions [40] by using the terms from the ontologies and could run the queries on the biomedical literature or electronic medical records to identify related articles and patient records to answer their questions. For consumers with little clinical background, disease-specific ontologies

could be used to expand or reformulate the original queries to the Google, PubMed, or MedlinePlus [41,44,45], which therefore may improve the effectiveness of the searches. For example, a person may be interested in the "treatment" of disease Y; however, they may not know what kind of treatments were available; by looking into the ontology, they may form a specific query with a comparison of two medications for the diseases. Third, disease-specific ontologies may facilitate the summarization of patient medical records. With the understanding about what information is relevant to the problem of interest, a system can extract disease-relevant information from a patient's long historical medical records and provide a summary to the clinicians. Moreover, disease-specific ontologies could be used for clinical researchers to identify a proper research cohort from an EHR. The phenotypes (e.g., signs, symptoms, diagnostic results) captured in the disease-specific ontologies can be useful for the development of cohort selection algorithms for finding target populations or subpopulations, which will further help clinical trial studies.

The development of disease DSOs is still at the beginning stage. As we reviewed all the ontologies stored in BioPortal and OBO foundry, less than 1% of the ontologies were built for specific diseases. Moreover, most of the disease-specific ontologies we found haven't covered a full range of the disease-specific medical knowledge. However, we foresee that the importance of the disease-specific ontologies will be increasingly recognized, and more and more DSOs will be developed.

2.2 Disease-Pertinent Knowledge Acquisition

Building disease-specific ontologies is labor-intensive. In the Medical Subject Headings (MeSH), there are over 2000 disease concepts in the disease categories. It is obvious that building ontologies for so many kinds of diseases could be a life-time job if

we do it manually. As we review the life cycle of ontology development [46], a crucial component is knowledge acquisition, which is a process of extracting, structuring, and organizing knowledge from a variety of knowledge sources. Therefore, it would be desirable to find or develop an automated or semi-automated knowledge acquisition method and extract the knowledge from existing knowledge sources. Most existing ontology development tools do not support an automated knowledge acquisition [8].

Working on the automated acquisition of disease-specific information, two important questions need to be addressed: what are the knowledge sources and how should we extract from them? To answer these two questions, we first review the prior endeavors of disease-specific knowledge acquisition in the biomedical domain.

Since the earlier 1990s, dozens of studies have investigated techniques for disease-*concept* association extraction from a variety of sources, where the *concept* could be associated genes [47], signs and symptoms [48], findings [49], medications [26,50], or lab tests [50]. The sources that have been mined broadly cover MEDLINE citations [26,51–56,28,57], Clinical records [49,26,50,58,59], NDFRT [60], DrugBank [61], FDA AERS [60], DailyMed [61], and AHFs Consumer Medication Information [61]. Among the sources, MEDLINE citations (title and abstract) and clinical records were the two most commonly used sources. Numerous knowledge acquisition techniques have been proposed to extract relational information from them, including co-occurrence-based statistics [49,26,50], natural language processing (NLP) [55,61], graph theory [47,62], conditional random fields [56], pattern learning [28,57], and others [56]. Among these techniques, co-occurrence-based statistics and natural language processing (NLP) are the two techniques mostly applied. Detailed review of the knowledge extraction techniques

can be found in Chapter 4 and 5.

Two barriers we identified led to the research work of this dissertation. First, most of the previous work focus on a large-scale extraction of disease-concept associations without a specific disease as a focus [26,50,28,57]. We are uncertain about their performance when applying them to the disease-specific level. It is important to develop and test approaches to support DSOs. Toward this end, we develop reference standards below in Chapter 3 and a disease-treatment extraction system in Chapter 4 with some comparison to previous works.

Second, existing automated extraction techniques not only identify the signal (i.e., relevant disease-concept associations) but also introduce noise (i.e., irrelevant disease-concept associations). The signal-to-noise ratio can be very low when focusing on high recall. The challenge remains when facing hundreds or thousands of concepts extracted for each disease in which the precision is low; how can we filter out the false positives? It is expensive to ask experts to manually determine the relevance of those extracted concepts to the disease of interest. In Chapter 5, we explore supervised machine learning techniques to overcome this barrier.

2.3 References

[1] Gruber TR. A translation approach to portable ontology specifications. Knowl Creat Diffus Util 1993;5:199–220.

[2] Gruber T. Toward principles for the design of ontologies used for knowledge sharing. Int J Hum Comput Stud 1995;43:907–28. doi:citeulike-article-id:230211.

[3] Noy NF, Mcguinness DL. Ontology development 101: a guide to creating your first ontology. Stanford Knowl Syst Lab Tech Rep KSL-01-05 Stanford Med Informatics Tech Rep SMI-2001-0880 2001.

[4] Genesereth MR, Fikes RE. Knowledge Interchange Format, Version 3.0 Reference Manual. Interchange 1992:1–68.

[5] Horrocks I, Patel-Schneider PF, van Harmelen F. From SHIQ and RDF to OWL: the making of a Web Ontology Language. Web Semant Sci Serv Agents World Wide Web 2003;1:7–26.

[6] Brickley D, Guha RV. RDF Schema 1.1 2014. https://www.w3.org/TR/rdf-schema/ (accessed October 1, 2017).

[7] Connolly D, Harmelen F Van, Horrocks I, Mcguinness DL, Patel-Schneider PF, Stein LA. DAML+OIL (March 2001) Reference Description 2001. https://www.w3.org/TR/daml+oil-reference (accessed October 1, 2017).

[8] Youn S, Mcleod D. Ontology development tools for ontology-based knowledge management. Encycl E-Commerce 2006.

[9] Gennari JH, Musen MA, Fergerson RW, Grosso WE, Crubézy M, Eriksson H, et al. The evolution of Protégé: An environment for knowledge-based systems development . International Journal of Human Computer Studies 2003; 58(1):89-123.

[10] Davis R, Shrobe H, Szolovits P. What is a knowledge representation? AI Magazine 1993;14:17–33.

[11] Jovic A, Prcela M, Gamberger D. Ontologies in medical knowledge representation. Proc. of Int. Conf. Information Technology Interfaces 2007:535–40.

[12] Horrocks I. Ontologies and the semantic web. Commun ACM 2008;51:58.

[13] Lambrix P, Habbouche M, Pérez M. Evaluation of ontology merging tools in bioinformatics. Bioinformatics 2003;19:1564–71.

[14] Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids

Res 2009;37:W170-3.

[15] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007;25:1251–5.

[16] Wang XH, Da Qing Zhang, Tao Gu, Pung HK. Ontology based context modeling and reasoning using OWL. IEEE Annu Conf Pervasive Comput Commun Work 2004 Proc Second 2004:18–22.

[17] Uschold M, Gruninger M. Ontologies: Principles, methods and applications. Knowl Eng Rev 1996;11:93–136.

[18] Castells P, Fernández M, Vallet D. An adaptation of the vector-space model for ontology-based information retrieval. IEEE Trans Knowl Data Eng 2007;19:261–72. doi:10.1109/TKDE.2007.22.

[19] Lopez V, Uren V, Motta E, Pasin M. AquaLog: An ontology-driven question answering system for organizational semantic intranets. Web Semant 2007;5:72–105.

[20] Ashburner M, Ball C a, Blake J a, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–9.

[21] Rosse C, Mejino Jr JLV. A reference ontology for biomedical informatics: The foundational model of anatomy. J Biomed Inform 2003;36:478–500.

[22] Stearns MQ, Price C, Spackman KA, Wang AY, Authority I, Cross A. SNOMED clinical terms: Overview of the development process and project status 2001:662–6.

[23] Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. IMIA Yearb Med Informatics 2008;47:67–79.

[24] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2010;25:1251.

[25] Rubin DL, Shah NH, Noy NF. Biomedical ontologies: a functional perspective. Brief Bioinform 2008;9:75–90.

[26] Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease-drug knowledge from biomedical and clinical documents: an initial study. J Am Med Informatics Assoc 2008;15:87–98.

[27] McCoy AB, Wright A, Laxmisan A, Ottosen MJ, McCoy J a, Butten D, et al.

Development and evaluation of a crowdsourcing methodology for knowledge base construction: identifying relationships between clinical problems and medications. J Am Med Inform Assoc 2012.

[28] Xu R, Wang Q. Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. BMC Bioinform 2013;14:181.

[29] Haug PJ, Ferraro JP, Holmen J, Wu X, Mynam K, Ebert M, et al. An ontology-driven, diagnostic modeling system. J Am Med Inform Assoc 2013;20:e102-10.

[30] Malhotra A, Younesi E, Gündel M, Müller B, Heneka MT, Hofmann-Apitius M. ADO: a disease ontology representing the domain knowledge specific to Alzheimer's disease. Alzheimer's Dement 2014;10:238–46.

[31] Younesi E, Malhotra A, Gündel M, Scordis P, Kodamullil AT, Page M, et al. PDON: Parkinson's disease ontology for representation and modeling of the Parkinson's disease knowledge domain. Theor Biol Med Model 2015;12:20.

[32] Malhotra A, Gündel M, Rajput AM, Mevissen HT, Saiz A, Pastor X, et al. Knowledge retrieval from pubmed abstracts and electronic medical records with the multiple sclerosis ontology. PLoS One 2015;10:e0116718.

[33] Chalortham N, Buranarach M, Supnithi T. Ontology development for type II diabetes mellitus clinical support system. Proc 4th Int Conf Knowl Inf Creat Support Syst 2009.

[34] Rahimi A, Liaw ST, Taggart J, Ray P, Yu H. Validating an ontology-based algorithm to identify patients with type 2 diabetes mellitus in electronic health records. Int J Med Inf 2014;83:768–78.

[35] El-Sappagh S, Farman A. DDO: a diabetes mellitus diagnosis ontology. Appl Informatics 2006;3.

[36] Hadzic M, Chang E. Ontology-based multi-agent systems support human disease study and control. Front Artif Intell Appl 2005;135:129.

[37] Bertaud-Gounot V, Duvauferrier R, Burgun A. Ontology and medical diagnosis. Inf Heal Soc Care 2012;37:51–61.

[38] Mendonça E a, Cimino JJ, Johnson SB, Seol YH. Accessing heterogeneous sources of evidence to answer clinical questions. J Biomed Inform 2001;34:85–98.

[39] Mohammed O, Benlamri R, Fong S. Building a diseases symptoms ontology for medical diagnosis: An integrative approach. First Int. Conf. Futur. Gener. Commun. Technol., 2012, p. 104–8.

[40] Demner-fushman D, Lin J. Answering clinical questions with knowledge-based

and statistical techniques. Computational Linguistics 2007; 33(1):63-103.

[41] Liu Z, Chu WW. Knowledge-based query expansion to support scenario-specific retrieval of medical free text. Inf Retr Boston 2007;10:173–202.

[42] Kazdin AE. Evidence-based treatment and practice: new opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. Am Psychol 2008;63:146–59.

[43] Ely JW, Osheroff JA, Gorman PN, Ebell MH, Chambliss ML, Pifer EA. A taxonomy of generic clinical questions : classification study. BMJ 2000;321:429–32.

[44] Bhogal J, Macfarlane A, Smith P. A review of ontology based query expansion. Inf Process Manag 2007;43:866–86.

[45] Plovnick R, Zeng Q. Reformulation of consumer health queries with professional terminology: A Pilot Study. J Med Internet Res 2004;6:e27.

[46] Stevens R, Goble C a, Bechhofer S. Ontology-based knowledge representation for bioinformatics. Brief Bioinform 2000;1:398–414.

[47] Özgür A, Vu T, Radev DR. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics 2008;24:i277–85.

[48] Wang X, Chused A, Elhadad N, Friedman C, Markatou M. Automated knowledge acquisition from clinical narrative reports. AMIA Annu Symp Proc, 2008, p. 783–7.

[49] Cao H, Markatou M, Melton GB, Chiang MF, Hripcsak G. Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. AMIA Annu Symp Proc, 2005, p. 106–10.

[50] Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. J Biomed Inform 2010;43:891–901.

[51] Cimino JJ, Barnett GO. Automatic knowledge acquisition from MEDLINE. Methods Inf Med 1993;32:120–30.

[52] Zeng Q, Cimino JJ. Automated knowledge extraction from the UMLS. AMIA Annu Symp Proc, 1998, p. 568–72.

[53] Mendonca EA, Cimino JJ. Automated knowledge extraction from MEDLINE citations. AMIA Annu Symp Proc, 2000, p. 575–9.

[54] Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pac Symp Biocomput

2000:517–28.

[55] Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing : interpreting hypernymic propositions in biomedical text 2003;36:462–77.

[56] Bundschus M, Dejori M, Stetter M, Tresp V, Kriegel H-P. Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinformatics 2008;9:207.

[57] Xu R, Li L, Wang Q. dRiskKB: a large-scale disease-disease risk relationship knowledge base constructed from biomedical text. BMC Bioinform 2014;15:105.

[58] Wang X, Chused A, Elhadad N, Friedman C, Markatou M. Automated knowledge acquisition from clinical narrative reports. AMIA Annu Symp Proc 2008:783–7.

[59] Islamaj Doğan R, Névéol A, Lu Z. A context-blocks model for identifying clinical relationships in patient records. BMC Bioinformatics 2011;12 Suppl 3:S3–S3.

[60] Wang X, Chase HS, Li J, Hripcsak G, Friedman C. Integrating heterogeneous knowledge sources to acquire executable drug-related knowledge. AMIA Annu Symp Proc, vol. 2010, 2010, p. 852–6.

[61] Névéol A, Lu Z. Automatic integration of drug indications from multiple health resources. Proc. 1st ACM Int. Heal. informatics Symp., 2010, p. 666–73.

[62] Zhang H, Fiszman M, Shin D, Miller CM, Rosemblat G, Rindflesch TC. Degree centrality for semantic abstraction summarization of therapeutic studies. J Biomed Inform 2011;In Press,:830–8.

[63] Bundschus M, Dejori M, Stetter M, Tresp V, Kriegel H-P. Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinformatics 2008;9:207.

CHAPTER 3

A METHOD FOR THE DEVELOPMENT OF DISEASE-SPECIFIC REFERENCE STANDARDS VOCABULARIES FROM TEXTUAL BIOMEDICAL LITERATURE RESOURCES

Reprinted with permission from Wang L, Bray BE, Shi J, Del Fiol G, Haug PJ. A method for the development of disease-specific reference standards vocabularies from textual biomedical literature resources. Artificial Intelligence in Medicine. 2016;68:47-57.

Artificial Intelligence in Medicine 68 (2016) 47–57

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

A method for the development of disease-specific reference standards vocabularies from textual biomedical literature resources

Liqin Wang[a,b,*], Bruce E. Bray[a,c], Jianlin Shi[a], Guilherme Del Fiol[a], Peter J. Haug[a,b]

[a] Department of Biomedical Informatics, University of Utah, 421 Wakara Way, Salt Lake City, UT 84108, USA
[b] Homer Warner Research Center, Intermountain Healthcare, 5121 South Cottonwood Street, Murray, UT 84107, USA
[c] Department of Internal Medicine, University of Utah, 30 North 1900 East, Salt Lake City, UT 84132, USA

ARTICLE INFO

Article history:
Received 5 June 2015
Received in revised form 22 February 2016
Accepted 25 February 2016

Keywords:
Knowledge extraction
Reference standards
Annotation
Saturation
Disease-specific ontology
Heart failure

ABSTRACT

Objective: Disease-specific vocabularies are fundamental to many knowledge-based intelligent systems and applications like text annotation, cohort selection, disease diagnostic modeling, and therapy recommendation. Reference standards are critical in the development and validation of automated methods for disease-specific vocabularies. The goal of the present study is to design and test a generalizable method for the development of vocabulary reference standards from expert-curated, disease-specific biomedical literature resources.

Methods: We formed disease-specific corpora from literature resources like textbooks, evidence-based synthesized online sources, clinical practice guidelines, and journal articles. Medical experts annotated and adjudicated disease-specific terms in four classes (i.e., *causes or risk factors*, *signs or symptoms*, *diagnostic tests or results*, and *treatment*). Annotations were mapped to UMLS concepts. We assessed source variation, the contribution of each source to build disease-specific vocabularies, the saturation of the vocabularies with respect to the number of used sources, and the generalizability of the method with different diseases.

Results: The study resulted in 2588 string-unique annotations for heart failure in four classes, and 193 and 425 respectively for pulmonary embolism and rheumatoid arthritis in *treatment* class. Approximately 80% of the annotations were mapped to UMLS concepts. The agreement among heart failure sources ranged between 0.28 and 0.46. The contribution of these sources to the final vocabulary ranged between 18% and 49%. With the sources explored, the heart failure vocabulary reached near saturation in all four classes with the inclusion of minimal six sources (or between four to seven sources if only counting terms occurred in two or more sources). It took fewer sources to reach near saturation for the other two diseases in terms of the treatment class.

Conclusions: We developed a method for the development of disease-specific reference vocabularies. Expert-curated biomedical literature resources are substantial for acquiring disease-specific medical knowledge. It is feasible to reach near saturation in a disease-specific vocabulary using a relatively small number of literature sources.

Published by Elsevier B.V.

1. Introduction

Disease-specific ontologies are knowledge bases intended to structure and represent disease-relevant information including disease etiology, diagnosis, treatment and prognosis. The availability of these ontologies could facilitate cross-disciplinary exchange and sharing of domain-specific knowledge. Disease-specific ontologies are also essential in supporting a variety of domain-specific computer applications, such as natural language processing, cohort selection, and clinical decision support [1,2]. For example, Haug et al. initiated a pneumonia-specific ontology, which supported the development of a clinical diagnostic modeling system [3]. Malhotra et al. constructed an Alzheimer's disease ontology and applied it to text mining on electronic health records [4].

However, the lack of comprehensive disease-specific ontologies hinders the development of such applications. BioPortal [5], an open repository of biomedical ontologies, currently hosts up to 467

* Corresponding author at: Department of Biomedical Informatics, University of Utah, 421 Wakara Way, Salt Lake City, UT 84108, USA.
E-mail address: liqin.wang@utah.edu (L. Wang).

http://dx.doi.org/10.1016/j.artmed.2016.02.003
0933-3657/Published by Elsevier B.V.

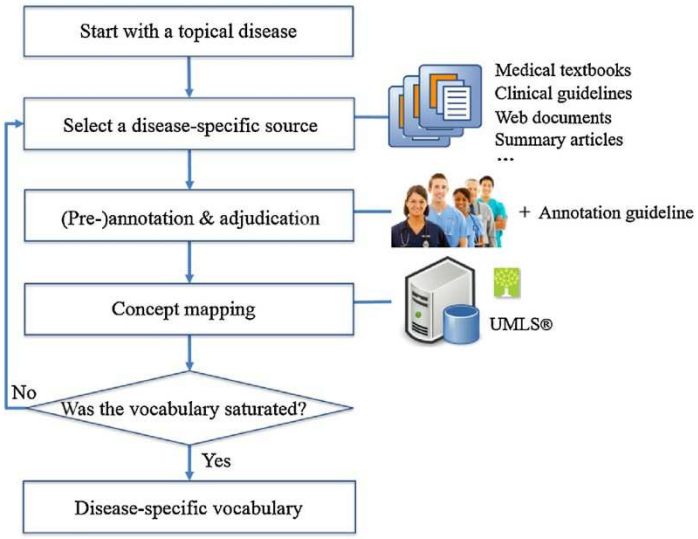

Fig. 1. Workflow for building near-saturated, disease-specific reference vocabularies from biomedical literature resources.

ontologies in various domains. However, among those ontologies less than 1% are disease-specific. Therefore, methods are needed to help develop disease-specific ontologies that can be made available to the community. A long-term goal of our research is to enable a platform that supports large-scale development of such ontologies.

Creating disease-specific ontologies is still a labor-intensive process. One of the main challenges is the knowledge acquisition, i.e., comprehensively ascertaining domain-specific concepts and relationships in the ontologies [6,7]. In knowledge engineering, domain experts are often used as the sources for acquiring medical knowledge. However, they are scarce and expensive. Another challenge is that while existing large terminologies, such as Disease Ontology [8] and Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [9], can be used as sources of concepts for disease ontologies, the relationships between the concepts are primarily hierarchical, with little non-hierarchical relations between diseases and their signs and symptoms, diagnostic procedures, and treatments. Therefore, it is not feasible to extract a comprehensive set of disease-related relationships from those terminologies.

A promising alternative to address disease-ontology development challenges is to learn ontologies from textual data [7,10]. The learning can be separated into multiple levels: learning terms, synonyms, concepts, relations, axioms and rules [7,10]. At the term level, for instance, Riloff proposed a corpus-based approach for building domain-specific semantic lexicons [11]. At relationship level, Sanchez and Moreno studied methods that learn non-taxonomic relationships from web documents [12]. Particularly for developing disease-specific ontologies, the learning is primarily focused on using narrative text sources, such as the biomedical literature, to automatically identify disease-relevant concepts and relations. The relationships include the taxonomy backbone (i.e., is-a relations) and non-hierarchical relations (e.g., treats, causes). Most hierarchical relations between biomedical concepts are well represented in large domain ontologies and terminologies, such as SNOMED CT and the Unified Medical Language System (UMLS). However, important gaps still exist in regards to non-hierarchical relations. Learning these relations is an active subject of research interest [13–17].

The goal of the present study is to design and test a generalizable method for the development of vocabulary reference standards from expert-curated, domain-specific documents, such as textbooks, and clinical guidelines. The vocabularies and analyses established will be used to help the development and testing of automated disease-specific knowledge acquisition algorithms.

In the process of developing reference vocabularies, the number and types of sources that are needed to maximize the number of concepts retrieved are unknown. One source is unlikely to provide all concepts and relations about a disease and it is not feasible to manually extract concepts from all literature sources available. Therefore, in present study, we investigate the number of sources that are needed to obtain saturation for a disease-specific vocabulary. We assessed the feasibility of acquiring disease-specific concepts and relationships in the classes of *causes and risk factors*, *sign and symptoms*, *diagnostic tests and results*, and *treatments* by manually annotating terms from a representative and diverse set of popular knowledge sources in cardiology. Last, we then tested the generalizability of our methods with two additional diseases in *treatment* class.

2. Methods

In the present study, a disease-specific vocabulary is understood to be a list of concepts that are semantically related to a disease or syndrome. We focused on gathering disease-related concepts into a collection rather than identifying their taxonomic structure [18]. The framework for acquiring disease-specific vocabulary is displayed in Fig. 1. This is an iterative process with the goal of reaching near saturation which in this study is defined as finding <5% new concepts with the introduction of a new resource. We first formed a corpus with a collection of textual biomedical literature documents on the topical disease. Then, we initiated the iterative process by selecting one source from the corpus. The following step is annotation and adjudication, where the documents were annotated using eHOST, an open source annotation tool [15], by medical experts based on an annotation guideline. Any conflicted annotations were adjudicated by consensus between experts. Annotations were then mapped to equivalent UMLS concepts or, if these were unavailable,

Table 1
Textual knowledge sources for extracting heart-failure-related concepts used to build a disease-specific vocabulary.

Sources	Types	Published/updated by	No. of citations	Included chapters/articles
Braunwald	Textbook	2011	4917	Chapter 26–34, excluded all the figures and references
Harrison's	Textbook	2011	9851	Chapter 234, exclude all the figures and references
UpToDate®	Evidence-based systematic online journal reviews	Dec 2013	N/A	The first three articles retrieved with the query of heart failure in adult [29–31]
DynaMed™	Evidence-based systematic online journal reviews	Feb 2014	N/A	The first document retrieved with the querying of "heart failure", include following sections: causes and risk factors, history and physical, diagnosis, treatment, and prevention and screening.
ACC guideline	Clinical practice guidelines	2009; 2013	1334; 535	For 2009 version, we include all the recommendations which are in bold, and all the tables. For 2013 version, all the content is included except the figures and references.
ESC guideline	Clinical practice guidelines	2012	1951	Included following chapters: 3–5, 7–10, 12–14.
ACC key data elements	Conclusive journal articles	2005	134	Section III. Heart failure clinical data standard elements and definitions

assigned to local codes. Subsequently, we annotated each additional source, assessing the saturation of the vocabulary after each new source was included. The iteration ended when the vocabulary reached near saturation. More details are provided in the following sections.

2.1. Selection and preparation of knowledge sources

There are a large number of textual resources available that provide disease-specific medical knowledge, such as textbooks, online evidence-based documents, journal articles, medical records, etc. When choosing source documents, we prefer those that are expert-curated, knowledge-dense, and evidence-based. As a starting point, the following types of knowledge sources were chosen: regularly updated textbooks, evidence-based synthesized online sources, clinical practice guidelines, and disease-specific journal articles. Medical textbooks provide a comprehensive and general overview of select diseases from their diagnosis to treatment. Two decades ago, Curley investigated physician's preferences for acquiring medical knowledge and found that among the many resources textbook and journals were most frequently used [19]. Nowadays, they are still among the conventional sources used by medical students to obtain medical knowledge. Another kind of source, clinical practice guidelines, have been made available for many medical domains and are used to "assist practitioner and patient decisions about appropriate health care for specific clinical circumstances." [20] They typically focus on therapeutic guidance. Due to the high demand for speed in answering daily clinical questions, online, point-of-care, evidence-based products have become available and have gained wide acceptance from healthcare professionals. Examples of these products include Clinical Evidence, DynaMed, InfoRetriever, PDxMD and UpToDate [21].

We consulted domain experts to choose one or two examples of each type of source mentioned above. For heart failure, a total of seven source documents were chosen (see Table 1): Braunwald's heart disease (Braunwald) [22], Harrison's principle of internal medicine (Harrison's) [23], UpToDate®, DynaMed™, American College of Cardiology Foundation (ACCF)/American Heart Association (AHA) guidelines for the management of heart failure (ACC guideline) [24,25], European Society of Cardiology (ESC) guidelines for the diagnosis and treatment of acute and chronic heart failure (ESC guideline) [26], and ACC/AHA key data elements and definitions (ACC key data elements) [27]. The publication date, popularity (number of citations), and the sections of each of the sources used in this study are further specified in Table 1. In the same way, we formed corpus for two other conditions. For pulmonary embolism, the initial corpus includes a clinical guideline [28], chapters from a textbook (Braunwald) [22], four articles from UpToDate®, and one document from DynaMed™. For rheumatoid arthritis, the corpus includes four articles from UpToDate®, one document from DynaMed™, and one chapter from Harrison's.

2.2. Annotation scheme

As we examined those expert-curated, disease-specific documents, we found that their content covers different aspects of a disease including etiology, diagnosis, treatment/prevention, and prognosis. To maintain feasibility for the annotation tasks, we provided a more granular classification, and restricted the annotations to four class types: *causes or risk factors*, *signs or symptoms*, *diagnostic tests or results*, and *treatment*; other classes were excluded from this annotation task, such as comorbidities and complications. Table 2 gives more detail about the definitions and examples for these annotation classes.

2.3. Annotation and adjudication

Annotation is a process to identify salient terms from a collection of narrative documents and assign them to a proper class, while adjudication is a process to resolve conflicting results between annotators. We used eHOST [32] for both processes. Besides annotation, eHOST supports dictionary export, pre-annotation, and measurement of inter-annotator agreement (IAA).

Initially, we developed an annotation guideline (available in the online supplement) with rules and classes. Before the actual annotation task, annotators went through a training process, and annotated a sample collection of documents. The degree of IAA between the two annotators was calculated using F-measure where we treated one annotator as the subject and the second annotator's results as if they were a gold standard [33] . Text annotation can be seen as an information retrieval task therefore normal Kappa statistics cannot be calculated without a

Table 2
Annotation scheme, definitions and examples.

Classes	Description	Examples
Causes or risk factors	Merges two overlapped subclasses: causes and risk factors, where causes refer to concepts or terms that can directly cause heart failure and risk factors are those factors associated with an increased risk of heart failure.	In industrialized countries, coronary artery disease (CAD) has become the predominant cause in men and women and is responsible for 60–75% of cause of heart failure. Hypertension contributes to the development of heart failure in 75% of patients.
Signs or symptoms	Groups medical signs and symptoms. In addition it includes the physical examination for which would usually result signs or symptoms.	The cardinal symptoms of heart failure are fatigue and shortness of breath. Nocturnal cough is a common manifestation of this process and a frequently overlooked symptom of heart failure.
Diagnostic tests or results	Includes phrases that describe procedures, panels, and measures that are done to a patient or a body fluid or sample in order to discover, rule out, or find more information about a medical problem.	A routine 12-lead ECG is recommended. A chest X-ray provides useful information about cardiac size and shape.
Treatment	Includes phrases that describe procedures, interventions, and substances given to a patient in an effort to resolve a medical problem.	Dietary restriction of sodium (2–3 g daily) is recommended in all patients with the clinical syndrome of heart failure and either preserved or depressed ejection fraction. Diuretics are the only pharmacologic agents that can adequately control fluid retention in advanced heart failure.

negative case count. Annotators iteratively annotated sample documents until the IAA score reached substantial agreement (IAA between 0.6 and 0.8) [34]. Then, annotators began working on the same set of documents from the corpora. Disagreements were resolved through a consensus process between two annotators. However, for the diseases in which the annotators reached almost perfect agreement (IAA greater than 0.8) on the first document, we proceeded to the subsequent documents with only one annotator per document.

We also used pre-annotation to improve the annotation quality and efficiency as suggested in previous studies [35,36]. Annotators were firstly assigned a small set of documents to annotate. After adjudication, we extracted a list of terms as a dictionary from these documents. Next, the subsequent documents were automatically pre-annotated with the dictionary compiled from previously adjudicated annotations. With these pre-annotated documents, annotators could modify or delete pre-annotations, or add missed occurrences of terms.

2.4. Mapping annotations to UMLS concepts

The annotation texts contain various lexical variations such as abbreviations/acronyms (e.g., "*EF*" for "*ejection fraction*"), synonyms (e.g., "*alcohol consumption*" and "*alcohol intake*"), compound terms (e.g., "*cardiac catheterization and revascularization*", "*coronary or peripheral vascular disease*"), and modifiers (e.g., "*daily serum electrolytes*"), which make it difficult to compare the annotations among the sources. To address this issue, we sought to map all the annotations from these source documents to standard terminologies, including SNOMED CT, Logical Observation Identifiers Names and Codes (LOINC), RxNorm and Medical Subject Headings (MeSH). This was expected to facilitate an analysis of the vocabulary sets obtained from different sources and used to form a final vocabulary. The terms that did not correspond to entries in standard terminologies may, in the future, be used to enhance existing terminologies. The UMLS Metathesaurus is the largest thesaurus in the biomedical domain. It has integrated hundreds of source terminologies, and provides cross-mapping to different source terminologies. We mapped the annotations to UMLS concepts while restricting the source terminologies to the four mentioned above. The 2014AB version was used in this study.

Concept mapping can be a subjective task. Lexical variations increase the complexity of the mapping. For some terms it may not even be possible to find mappings from the UMLS Metathesaurus. In order to reduce the subjectivity of the mapping and to make the mapping process more reliable and reproducible, we set up several mapping rules (see Table 3). For example, we restricted the semantic types for mapped concepts to the *treatment* and *diagnostic tests or results* classes. For terms with modifiers (e.g., daily serum electrolytes), we used a post-coordinated mapping approach that we removed the modifier (e.g., daily) and assigned the core term (e.g., serum electrolytes) with equivalent UMLS concept unique identifier (CUI). Compound terms (e.g., atrial and ventricular arrhythmias) were extended and mapped to multiple terms.

The mappings were automatically processed by MetaMap [37] and followed by manual verification and selection of concepts. Mapping rules were applied for some special cases (see examples in Table 3). The annotations that were not mappable to any targeted terminologies in any form were temporarily assigned local codes in order to support the analysis (e.g., atrial fibrillation surgery → atrial fibrillation surgery – Local Code: T0000001). The manual verification and selection were mainly done by one investigator (LW), with some assistance from a cardiologist (BEB). We also tested the agreement of the mappings by comparing the mapping results against mappings from another individual (JS) on 237 randomly sampled terms. The agreement of the mappings between these two individuals in terms of F-measure [33] was 0.84.

2.5. Saturation assessment

A natural process to build a vocabulary from a corpus (or knowledge sources) is to add all the acquired vocabulary from it at once. However, in order to answer our research question, we analyzed the accumulation process of the acquired vocabulary where we take the sources one by one, and determine whether, at a certain point, the vocabulary reaches some level of saturation. The accumulation rate of the vocabulary is calculated by the ratio of the number of new concepts from the last included source divided by the total number of concepts from all included sources (see Formula (1)). We considered a concept to be new when its code (either a UMLS CUI or temporary assigned code) did not appear in the vocabulary from previous documents. We chose 5% as an arbitrarily threshold

Table 3
Rules for mapping the annotations to UMLS concepts.

Mapping rules	Example
1. Map the terms to concepts that convey the term specific meaning within the context of the original sentence.	Sentence: "History of exposure to cardiotoxic substances through substance abuse: cocaine, amphetamine, ephedrine, other (specify)." (from ACC key data elements) *Cocaine* (Class: causes or risk factors) → Cocaine abuse (UMLS CUI: C0009171) Sentence: "Mechanical circulatory support in chronic heart failure evolved as a means of supporting patients awaiting transplantation, and this indication provided successful transition to heart transplantation and enhanced post-transplantation outcomes." (from Branwald) *Transplantation* (*Class: treatment*) → Heart transplantation (UMLS CUI: C0018823)
2. Separate terms into two or multiple terms when they contain "and" or "or" and the combined terms cannot be mapped to the target terminologies.	*Weight gain or loss* → weight gain (UMLS CUI: C0043094); weight loss (UMLS CUI: C0043096) *Mitral, aortic, tricuspid, and/or pulmonic valve surgical replacement* → Replacement of aortic valve (UMLS CUI: C0003506); Replacement of mitral valve (UMLS CUI: C0026268); Replacement of tricuspid valve (UMLS CUI: C0190119); Replacement of pulmonary valve (UMLS CUI: C0190129).
3. Restrict UMLS terminology sources to SNOMED CT, LOINC, RxNorm, and MeSH for concept mapping. For terms that can be mapped to multiple UMLS concepts, choose the concept that contains the target sources (example a). Terms that can be mapped to UMLS concepts, but not to one of the target sources should be left unmapped (example b).	For example, (a) *Abdominal fullness* was mapped to Abdominal bloating (UMLS CUI: C1291077) instead of Fullness abdominal (UMLS CUI: C0235318) because the C1291077 has source of SNOMED CT. (b) *Acute dyspnea* was left unmapped instead of mapping to acute dyspnea (UMLS CUI: C0743323) because C0743323 is not contained in the target sources.
4. Use semantic types to choose a proper mapping when terms can be mapped to multiple concepts.	For *diagnostic tests or results*, preferred semantic types are: Laboratory procedure, Laboratory or test result, e.g., *Blood urea nitrogen* → Blood urea nitrogen measurement (UMLS CUI: C0005845) *Atrial fibrillation* → ECG: atrial fibrillation (UMLS CUI: C0344434) For *treatment*, preferred semantic types are: pharmacologic substance, therapeutic or preventive procedure, e.g., *Digitalis* → Digitalis preparation (UMLS CUI: C0304520) *Yoga* → Yoga (UMLS CUI: C1883583)
5. Map term to the UMLS concept as close in meaning as possible.	*Chest radiograph* → Plain chest X-ray (UMLS CUI: C0039985) *Salt restriction* → Low sodium diet (UMLS CUI: C0012169)
6. Map terms with modifiers (e.g., daily, severe) by post-coordinating multiple UMLS concepts.	*History of Chagas disease* → Chagas disease (UMLS CUI: C0041234) + medical history (modifier) *Daily serum electrolytes* → Serum electrolytes measurement (UMLS CUI: C0587355) + daily (modifier)

to determine near saturation, i.e., a small number of new concepts are added by including new sources.

$$\text{Accumulation rate} = \frac{\text{the number of new concepts } (S_n)}{\text{the number of all concepts } \left(\sum_n^1 S_i\right)} \quad (1)$$

where S_i is the ith entered source.

The accumulation rate is the key factor used in this study to determine whether a vocabulary has reached near saturation or not. However, the rates can be affected by the order in which sources are included. Supposing that most of the concepts from one entered source have been presented in the existing vocabulary, the accumulation rate could drop significantly; however the rate could increase again if a subsequent source has substantially different vocabulary from others. To adjust for this situation, we determine the order of the sources in a given corpus by maximizing or minimizing the accumulation rate at each step. In another words, at each accumulation step the next incoming source is determined by selecting the source that can achieve the highest or lowest accumulation rate among the remaining sources. This method helps determine the lower bound and/or upper bound of the number of sources for reaching near saturation for a disease-specific vocabulary with the corpus explored. When the order of the sources changes, the number of sources for reaching near saturation falls in that range.

Although all the chosen sources are expert-curated and the annotation was done by medical experts, it does not guarantee all acquired relations are clinically-valid. For example, annotating conclusion sentences from a single clinical trial study could bring relations that are not clinically-valid into the vocabulary. We believe that terms that only occur in one source should be treated as less valid than those that occur in multiple sources. To address this issue, we include an analysis based on the concepts that appear in two or more sources which we called core concepts.

The saturation measurement is source-dependent. If the sources consistently overlap, then a smaller number of input sources are needed in order to reach a saturated status. Therefore, we use F-measure to measure the agreement of the chosen sources to indicate the degree of the source variety.

3. Results

For heart failure, two annotators reached substantial agreement on the first set of documents and therefore both were involved in the annotation and adjudication on the seven source documents. The annotation process resulted in 2588 string-unique annotations, which were mapped to 1648 concepts. The majority of these annotations (N = 2109, 81%) was mapped to 1232 UMLS concepts with the sources restricted to SNOMED CT, LOINC, RxNorm, and MeSH. The remainder were mapped to 416 local terms. For the other two conditions, two annotators reached perfect agreement in the first documents, and therefore, the subsequent documents were processed with only one annotator per document. For pulmonary embolism, we retrieved 193 string-unique annotations related to *treatment*, which were mapped to 96 concepts (83% were mapped to UMLS concepts). Similarly, for rheumatoid arthritis, we obtained 425 string-unique annotations, which were mapped to 279 concepts (83% were mapped to UMLS concepts). Independent of the diseases and class types, we obtained 3142 string-unique annotations, which mapped to 2049 concepts. On average, each concept had 1.5 corresponding terms or synonyms.

3.1. Concepts distributed by sources and classes

Fig. 2 shows the number of unique concepts per class across seven knowledge sources related to heart failure. Among the seven sources, Braunwald had the largest number of concepts in the

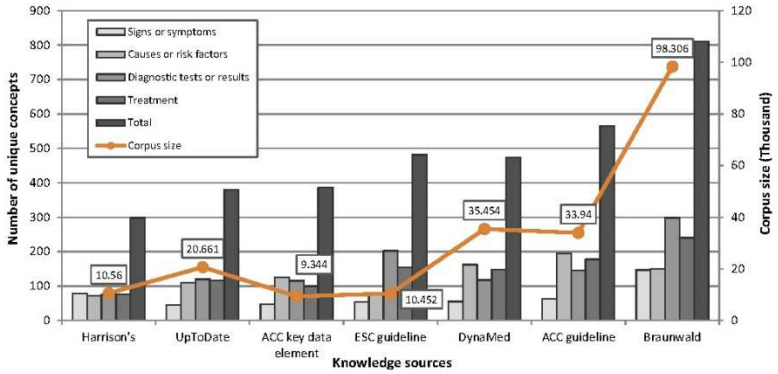

Fig. 2. Distribution of heart failure concepts extracted from different knowledge sources for signs or symptoms, causes or risk factors, diagnostic tests or results, and treatment. The numbers in the rectangles represent the corpus size (word count) of each source.

classes of *signs or symptoms*, *diagnostic tests or results*, and *treatment*, while the ACC guideline had the largest amount of concepts in the class of *causes or risk factors*. Fig. 2 also shows the number of acquired concepts along with the size of each document based on word count. Harrison's is the smallest corpus, containing 10,560 words, and from which we obtained 298 concepts, while Braunwald is the largest corpus (98,306 words) and from which we obtained 810 concepts. However, the size of the documents was not proportional to the number of obtained concepts. For example, UpToDate is almost twice as large as ACC key data elements and ESC guideline (20,661 vs. 9344 and 10,452 words); however, it provided a slightly smaller vocabulary of concepts (380 vs. 386 and 481 concepts).

Table 4 shows the contribution of each knowledge source to individual classes as well as the final heart failure vocabulary. The contribution of the seven knowledge sources ranged between 13–63% for the individual classes and between 18–49% for the final vocabulary. Among the seven sources, ACC guideline had the best contribution to the class of *causes or risk factors*, while Braunwald had best contribution to the other three classes. Some concepts were assigned to multiple classes, which explains why the sum total of the four classes is not equal to the total number of the vocabulary. For example, "hypertension" was assigned to the class of *causes or risk factors* and *diagnostic tests or results in a different context*.

3.2. Distribution of concept occurrence

Fig. 3 shows the log-log-scale distribution of the number of heart-failure concepts by concept frequencies. 747 concepts (45% of the total) were annotated only once in the entire corpus, and 247 concepts (15% of the total) were occurred twice, however, some concepts were annotated much frequently, such as one concept was annotated 471 times. Overall, the log-log-scale distribution appears approximately linear. Table 5 lists a small set of concepts that frequently occurred in the corpus, where the frequency of concept is measured by the occurrence of the corresponding annotations over the corpus.

3.3. Variety of the sources

Table 6 shows the agreement of concepts among the seven sources. The overall agreement between source pairs ranged from 0.28 to 0.46. From Table 6, ACC key data element consistently had lower agreement with other sources. Sources from the same category (e.g., textbook), such as Braunwald and Harrison's, did not show a stronger agreement than sources from different categories.

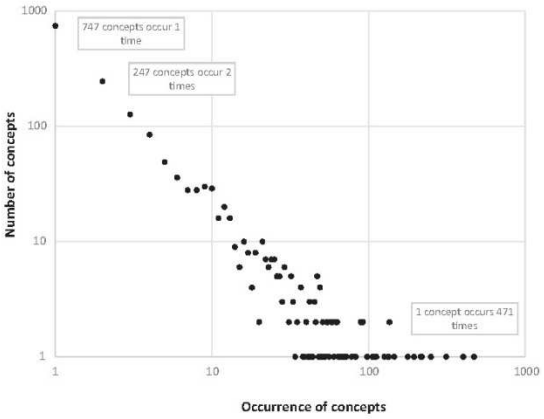

Fig. 3. Log–log scale plot of the distribution of the number of heart failure concepts by concept occurrence.

3.4. Concept accumulation by classes

With the explored corpus, Fig. 4A shows the minimum accumulation of heart failure concepts with the inclusion of additional sources. The accumulation curves of all four classes appear approximately linear. The number of concepts per class ranged from 170 concepts in *signs or symptoms* to 445 concepts in *diagnostic tests or results*. Fig. 4B shows the maximum accumulation of heart failure concepts. The accumulation curves of all four classes increase quickly with the inclusion of the first set of sources, but reach a plateau around the inclusion of the sixth source.

Fig. 4C and D show the accumulation rates with the inclusion of new sources (percent of identified concepts that are new) with two different orders of sources, where the 4C achieved the minimal accumulation rates and 4D achieved the maximum accumulation rates. The curves in Fig. 4C decline at the beginning, however, stay flatten around 20% accumulation rate since the inclusion of the fourth sources, and even increase slightly with the introduction of the sixth or seventh sources. We found that Braunwald is the last source in the minimum accumulation, which contained almost half amount of the concepts in the final vocabulary. The curves in Fig. 4D

Table 4
Contribution of each knowledge source to the four classes and the final heart-failure vocabulary.

Sources	Classes				
	Causes or risk factors (N = 435)	Signs or symptoms (N = 233)	Diagnostic tests or results (N = 590)	Treatment (N = 477)	Final vocabulary (N = 1648)
Harrison's	71 (16%)	79 (34%)	75 (13%)	77 (26%)	298 (18%)
UpToDate	110 (25%)	45 (19%)	120 (20%)	116 (24%)	380 (23%)
ACC key data element	126 (30%)	47 (20%)	115 (19%)	100 (23%)	386 (23%)
ESC guideline	81 (17%)	54 (23%)	205 (35%)	154 (32%)	484 (29%)
DynaMed	162 (37%)	55 (24%)	118 (20%)	148 (31%)	474 (29%)
ACC guideline	194 (45%)	63 (27%)	146 (25%)	178 (37%)	565 (34%)
Braunwald	150 (34%)	147 (63%)	298 (51%)	239 (50%)	810 (49%)

Table 5
Top 5 frequently occurring heart failure concepts in four classes.

Classes	UMLS CUI/code	Concept	Frequency
Signs or symptoms	C0013404	Dyspnea	136
	C0268000	Body fluid retention	55
	C0018810	Heart rate	48
	C0546817	Fluid overload	47
	C0015672	Fatigue	46
Causes or risk factors	C0020538	Hypertension	133
	C0027051	Myocardial infarction	133
	C0004238	Atrial fibrillation	132
	C0010054	Coronary arteriosclerosis	123
	C0011849	Diabetes mellitus	92
Diagnostic tests or results	C0428772	Left ventricular ejection fraction	219
	C0232174	Cardiac ejection fraction	195
	C1095989	Brain natriuretic peptide measurement	134
	C0022662	Kidney function tests	110
	C0013516	Echocardiography	82
Treatments	C0003015	Angiotensin-converting enzyme inhibitors	471
	C0001645	Adrenergic beta-antagonists	402
	C0012798	Diuretics	313
	C0521942	Angiotensin II receptor antagonist	250
	C1167956	Cardiac resynchronization therapy	215

Table 6
The agreement of the concepts among seven knowledge sources.

Agreement score	Braunwald	ACC guideline	DynaMed	ESC guideline	ACC key data element	UpToDate
Braunwald						
ACC guideline	0.45					
DynaMed	0.37	0.42				
ESC guideline	0.45	0.43	0.34			
ACC key data element	0.30	0.32	0.31	0.31		
UpToDate	0.43	0.46	0.38	0.40	0.30	
Harrison's	0.41	0.41	0.38	0.40	0.28	0.43

decline steeply with the inclusion of the second and third sources and are flatten after the inclusion of the fourth source. Although the number of concepts per class is different (from Fig. 4A), the accumulation rates are similar for all classes. After the sixth source, very few concepts are added to the vocabularies.

Fig. 5A and B show the minimum and maximum accumulation of the number of core concepts (i.e., concepts occur in two or more sources) respectively to the inclusion of additional sources. Fig. 5C and D are the corresponding analysis of the accumulation rates. Based on these two figures, the heart failure vocabulary reaches near saturation by using between four to seven sources regardless of the order the sources. Comparing the minimal accumulation rates between 4C and 5C, the accumulation rates in 5C decline faster with all four classes, and reach the 5% threshold after adding the seventh source, while in 4C, the accumulation rates never reach the 5% threshold. The accumulation rates of all four classes in 5D decline slightly faster than the rates in 4D, and reach the 5% threshold after adding the fourth source comparing sixth source in 4D. This confirms that concepts that are used more often can be gathered with fewer knowledge sources. Table 7 provides sample terms that occurred in one and multiple sources for each class.

3.5. Concept accumulation by conditions

A similar decline in the benefit of reviewing additional sources was seen in the other two diseases studied. Fig. 6A shows the maximum accumulation of treatment concepts along with increase of corpus size. For different diseases, the number of retrieved treatment concepts varies. From Fig. 6B, we found that the accumulation rates of heart failure decreased to the 5% threshold after adding the sixth source, while it only took three sources for rheumatoid arthritis to reach the threshold and four sources for pulmonary embolism.

Fig. 7A and B show the minimum and maximum accumulation rates of core *treatment* concepts per condition along with the addition of new sources. Based on these two figures, heart-failure *treatment* vocabulary of core concepts reached near saturation by

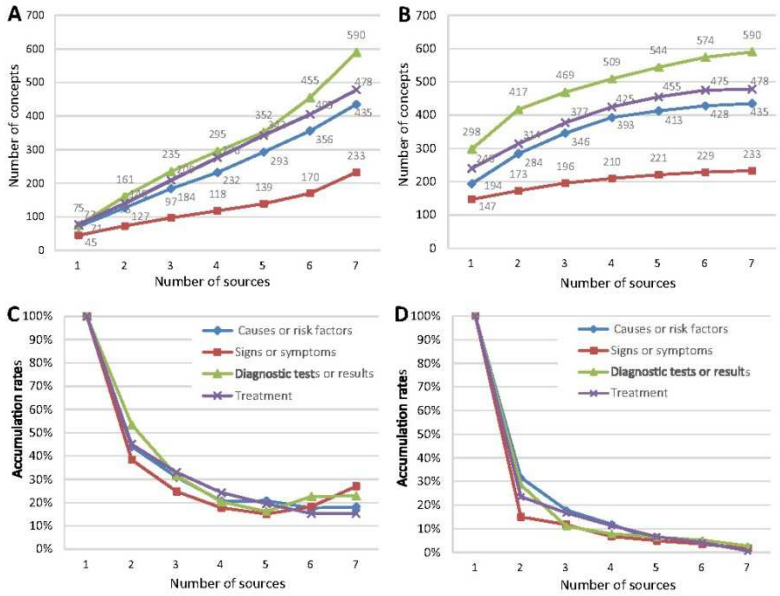

Fig. 4. (A) Number of heart failure concepts per class with the addition of new sources in minimum accumulation; (B) number of heart failure concepts per class with the addition of new sources in maximum accumulation; (C) minimum accumulation rates of heart failure concepts per class with the addition of new sources; (D) maximum accumulation rates of heart failure concepts per class with the addition of new sources.

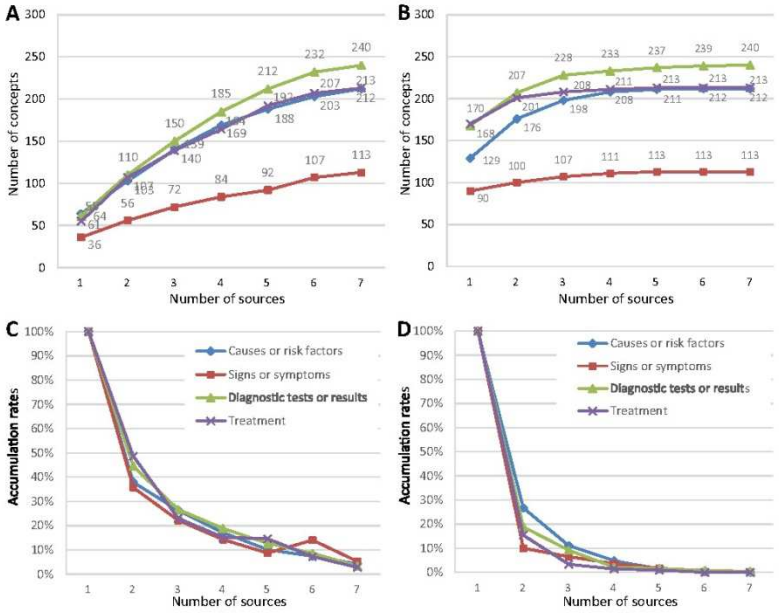

Fig. 5. (A) Number of heart-failure core concepts (occurred in two or more sources) with the addition of new sources in minimum accumulation; (B) number of heart-failure core concepts with the addition of new sources in maximum accumulation; (C) minimum accumulation rates of heart-failure core concepts per class with the addition of new sources; (D) maximum accumulation rates of heart-failure core concepts per class with the addition of new sources.

Table 7
Examples of heart-failure terms of four class types that occurred in one, two, three and seven sources.

Examples	Terms from 1 source	Terms from 2 sources	Terms from 3 sources	Terms from 7 sources
Signs or symptoms	Narrow pulse pressure; Jaw pain; Mottled skin; Presyncope; Purple-blue nail bed	Exercise intolerance; Abdomen distended; Abdominal pain; Apex beat displaced; Ventricular filling pressure increased	Reduced exercise tolerance; Early satiety; Respiratory distress; Dyspnea on exertion; Hypoxia	S3 – third heart sound; Orthopnea; Fatigue; Dyspnea; Angina pectoris
Causes or risk factors	Deep vein thrombosis; Dysglycemia; Collagen diseases; Cilostazol; Egg consumption	Hyperlipidemia; Chronic kidney failure; Pulmonary arterial hypertension; Viral myocarditis	Fabry disease; Angina, unstable; Familial cardiomyopathy; Beriberi; Rheumatic fever	Coronary arteriosclerosis; Myocardial ischemia; Myocardial infarction; Diabetes mellitus; Hypertension
Diagnostic tests or results	Measurement of liver enzyme; Blood cell count; Hemoglobin A1c measurement; Indirect bilirubin measurement; Oral glucose tolerance test	Blood pressure monitoring; Myocardial biopsy; ECG: left ventricular strain; Pharmacologic and exercise stress test; Prolonged QRS duration	Cardiovascular monitoring; Cardiopulmonary exercise test; ECG: atrial fibrillation; Serum calcium measurement; Blood pressure monitoring	Maximum oxygen uptake; Radionuclide ventriculography; Fluid overload; Magnetic resonance imaging; Left ventricular ejection fraction
Treatments	Alcohol deterrents; Sheng-Mai San; Dietary supplementation; Repair of pulmonary valve; Genetic counseling; Cell therapy	Ablation of atrioventricular node; Epoprostenol; Pericardiectomy; Weight reduction regimen; Hydralazine hydrochloride	Mitral valvuloplasty; Implantation of ventricular assist device; Fluid intake restriction; Implantation of CRT-D; Pneumococcal vaccination	Angiotensin II receptor antagonist; Inotropic agent; Heart transplantation; Diuretics; Angiotensin-converting enzyme inhibitors

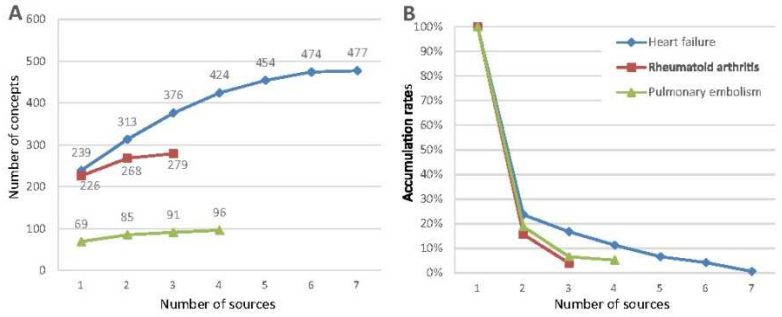

Fig. 6. (A) Number of *treatment* concepts per disease with the addition of new sources; (B) accumulation rates of *treatment* concepts per disease with the addition of new sources.

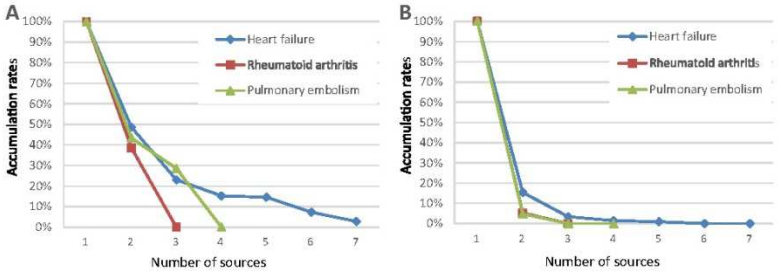

Fig. 7. (A) Minimum accumulation rates of core *treatment* concepts per disease with the addition of new sources; (B) maximum accumulation rates of core *treatment* concepts per disease with the addition of new sources.

using between three and seven sources; rheumatoid arthritis used between two and three sources; and pulmonary embolism used between two and four sources. All three conditions reached near saturation with a relatively small number of sources regardless of the order of sources.

4. Discussion

In this study, we assessed the feasibility of obtaining a near-saturated disease-specific vocabulary using a diverse set of expert-curated textual knowledge sources. We also estimated the

number of sources needed to reach near saturation with four disease concept classes (i.e., *causes or risk factors*, *sign or symptoms*, *diagnostic tests or results*, and *treatment*) with heart failure. We tested the generalizability of the method with two other conditions, i.e., pulmonary embolism, and rheumatoid arthritis.

From the study, we found that regardless of the difference of the total number of acquired concepts, the vocabularies of four concept classes reached near saturation at similar pace. For the four concept classes explored on the heart failure condition, the vocabularies took six sources to reach near saturation with the best order of sources. However, when order changes, they may require more sources to reach near saturation. When considering only the core concepts, three conditions all achieved near saturation with much fewer sources regardless of the order of sources. Overall, the results support the conclusion that it is feasible to obtain near-saturated reference standards for disease-specific vocabularies of core concepts using a relatively small number of knowledge sources. If choosing preferentially the sources with big contribution, the vocabularies of all concepts can also achieve near saturation with a relatively small number of sources.

The main contribution of the study lies in two aspects. First, the findings of this study are important for the development of disease-specific reference vocabularies. Developing reference standards usually involves substantial manual knowledge acquisition effort. The results of our study provide an estimate for the optimal number of text sources that can be used to find a balance between cost and saturation. Second, the method proposed in the present study provides an underlying approach for the development of disease-specific vocabularies. This includes the selection of sources, the guideline for annotation, the rules for mapping, and saturation assessment. Our approach is designed to be efficient as the method is able to determine a stopping point where a vocabulary has reached near saturation. Since the method had similar results in different conditions and concept classes, it is expected that the method and results will generalize to other conditions. We intend to use this approach to develop other disease-specific vocabularies and use them as reference standards for the development and testing of automated methods to generate disease-specific concept vocabularies. Xu et al. provides an example of such an automated method that could benefit from our reference standards [17]. Their method used known disease-drug pairs to learn patterns from biomedical text, enabling an automated, large-scale extraction of drug-disease treatment pairs. The reference standards can also be used to assess the performance of the automated systems by comparing the automated generated concepts to those manually extracted disease-specific vocabularies.

Seven authoritative sources that are frequently used by clinicians were chosen for heart failure. No previous study has investigated these sources regarding their contribution to a specific topic and the overlap among the sources. From our results, the agreement scores of the concepts among the seven sources ranged from 0.28 to 0.46, which indicates that these sources did not strongly overlap. The study results also suggest that specialized textbooks (e.g., Braunwald for heart failure) should be used as a starting point for building domain-specific vocabularies, as they appear to provide the broadest contribution to the vocabulary compared to other sources. However, the results discourage using textbooks as the only source, as even the lengthiest textbook (Braunwald) only captured half of the domain concepts represented in the final vocabulary. The use of multiple and diverse sources is critical to construct a comprehensive vocabulary. When applying our method to other conditions, we recommend that the most optimal approach is to start the annotation with comprehensive textbooks on topical disease, followed by relevant and most-updated clinical practice guidelines, and then topic summary articles from evidence-based synthesized online resources such as UpToDate.

The annotation process was sometimes complex and subjective. For example, for heart failure, it was difficult to discriminate between the treatment of its symptoms and treatments directed toward the causes and risk factors (e.g., hypertension, diabetes mellitus). A large portion of each source document was dedicated to discussing the treatment of the causes and risk factors of heart failure. Discriminating among these relationships is necessary to correctly associate diseases with their concepts.

Mapping annotations to standard terminologies or local terms is an essential step for saturation assessment. Without mapping, it may require to annotate a larger corpus to reach near saturation, however, which may only lead to more terms with all kind of varieties. After the mapping, we identify many terms ($N = 364$) that were not available in standard terminologies. These unmapped concepts could be used to enhance existing standard vocabularies. Another interesting finding is that almost half amount of concepts annotated only once in the corpus (see Fig. 3). Based on a manual review on the concepts with different occurrences, concepts that occurred only once or came from a single source (see Table 7) show less clinical relevance to the topic condition. In order to build a disease ontology with strong evidences, we may exclude those concepts. Besides, this distribution of concept occurrence over the corpus (see Fig. 3) almost follows a Zipf distribution. This could be possibly used for ranking strength of the relations of the concepts to the disease.

Our study has a few limitations. First, although the final vocabulary reached near saturation, we believe many disease-specific concepts are still missing. For example, "Angiotensin-converting-enzyme inhibitor" (ACE Inhibitor) – a class of drugs for treating heart failure, is present in the acquired vocabulary. But not all the drugs under this class were explicitly presented in the source documents. When experts mentioned the ACE inhibitor in those source documents, they were probably referring to the entire class of the drugs. These kinds of missing concepts could be inferred from relationships in existing standard terminologies, such as RxNorm and SNOMED CT. Second, the actual coverage of the final vocabulary is unknown. Assessing the actual coverage of a vocabulary is difficult because perfect reference vocabularies are not available. Third, the near saturation was reached with an optimal order of the sources. However, when the order of sources changed, the vocabulary may not be saturated. The upper bound of the number of sources for near saturation was not detected with a small number of sources explored in this study. However, for the vocabulary of core concepts is sufficient to reach near saturation with a relatively small number of sources regardless of the order of sources.

5. Conclusions

We provided an underlying approach for the development of disease-specific reference vocabularies focused on the concept classes of *causes and risk factors*, *signs and symptoms*, *diagnostic tests and results*, and *treatment*. Our findings show that expert-curated sources, such as textbooks, clinical guidelines, evidence-based summaries, and journal articles, are substantial sources for disease-specific medical knowledge. Their contribution to the vocabulary varies substantially for a specified condition. While the numbers of sources for reaching near saturation can vary modestly for different conditions, a relatively small number of text sources are sufficient to obtain a near-saturated vocabulary of sound disease-specific concepts. In the future we intend to develop automated techniques to extract disease-specific vocabularies from large corpora. The reference standards developed in the present study will

be used to assess the performance of the automated vocabulary extraction system.

Data availability

The extracted heart-failure-specific vocabulary, comprising 1648 concepts in the aspect of causes and risk factors, signs or symptoms, diagnostic tests or results, and treatment, is available at http://purl.bioontology.org/ontology/HFO. The rheumatoid arthritis vocabulary contains 279 concepts in the aspect of treatment and is available at: http://bioportal.bioontology.org/ontologies/RAO, while pulmonary embolism vocabulary with 96 concepts is available at: http://bioportal.bioontology.org/ontologies/PE. The hierarchical relationships in these vocabulary are obtained from the UMLS.

Acknowledgements

The authors thank Robert Hausam in the initial set up of annotation training, and Philip Brewster for the language editing. This work is supported in part by Grant LM010482 from the National Library of Medicine.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.artmed.2016.02.003.

References

[1] Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. IMIA Yearb Med Inform 2008;47:67–79.
[2] Rubin DL, Shah NH, Noy NF. Biomedical ontologies: a functional perspective. Brief Bioinform 2008;9:75–90, http://dx.doi.org/10.1093/bib/bbm059.
[3] Haug PJ, Ferraro JP, Holmen J, Wu X, Mynam K, Ebert M, et al. An ontology-driven, diagnostic modeling system. J Am Med Inform Assoc 2013;20:e102–10, http://dx.doi.org/10.1136/amiajnl-2012-001376.
[4] Malhotra A, Younesi E, Gündel M, Müller B, Heneka MT, Hofmann-Apitius M. ADO: a disease ontology representing the domain knowledge specific to Alzheimer's disease. Alzheimer's Dement 2014;10:238–46, http://dx.doi.org/10.1016/j.jalz.2013.02.009.
[5] Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res 2009;37:W170–3, http://dx.doi.org/10.1093/nar/gkp440.
[6] Noy NF, Mcguinness DL. Ontology development 101: a guide to creating your first ontology. Stanford Knowl Syst Lab Tech Rep KSL-01-05 Stanford Med Informatics Tech Rep SMI-2001-0880; 2001.
[7] Buitelaar P, Cimiano P, Magnini B. Ontology learning from text: methods, evaluation and applications. Amsterdam, The Netherlands: IOS Press; 2005, doi: 10.1.1.70.3041.
[8] Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, et al. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res 2012;40:D940–6, http://dx.doi.org/10.1093/nar/gkr972.
[9] Stearns MQ, Price C, Spackman KA, Wang AY, Authority I, Cross A. SNOMED Clinical Terms: overview of the development process and project status. AMIA Annu Symp Proc 2001:662–6.
[10] Cimiano P. Ontology learning and population from text: algorithms, evaluation and applications. Secaucus, NJ, USA: Springer-Verlag New York, Inc.; 2006, http://dx.doi.org/10.1007/978-0-387-39252-3.
[11] Riloff E, Shepherd J. A corpus-based approach for building semantic lexicons. Proc Second Conf Empir Methods Nat Lang Process 1997:117–24.
[12] Sanchez D, Moreno A. Learning non-taxonomic relationships from web documents for domain ontology construction. Data Knowl Eng 2008;64:600–23.
[13] Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease-drug knowledge from biomedical and clinical documents: an initial study. J Am Med Inform Assoc 2008;15:87–98, http://dx.doi.org/10.1197/jamia.M2401.
[14] Wang X, Chused A, Elhadad N, Friedman C, Markatou M. Automated knowledge acquisition from clinical narrative reports. AMIA Annu Symp Proc 2008:783–7.
[15] Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. J Biomed Inform 2010;43:891–901, http://dx.doi.org/10.1016/j.jbi.2010.09.009.
[16] Xu R, Li L, Wang Q. dRiskKB: a large-scale disease-disease risk relationship knowledge base constructed from biomedical text. BMC Bioinform 2014;15:105, http://dx.doi.org/10.1186/1471-2105-15-105.
[17] Xu R, Wang Q. Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. BMC Bioinform 2013;14:181, http://dx.doi.org/10.1186/1471-2105-14-181.
[18] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37:394–403.
[19] Curley SP, Connelly DP, Rich EC. Physicians' use of medical knowledge resources: preliminary theoretical framework and findings. Med Decis Mak 1990;10:231–41, http://dx.doi.org/10.1177/0272989X9001000401.
[20] Institute of Medicine (US) Committee to Advise the Public Health Service on Clinical Practice Guidelines. Clinical practice guidelines: directions for a new program. Washington, DC: National Academies Press (US); 1990.
[21] Alper BS. Practical evidence-based Internet resources. Fam Pract Manage 2003;10:49–52.
[22] Bonow RO, Mann DL, Zipes DP, Libby P. Braunwald's heart disease: a textbook of cardiovascular medicine. 9th ed. Philadelphia, PA: W.B. Saunders Company; 2011.
[23] Longo D, Fauci A, Kasper D, Hauser S, Jameson J, Loscalzo J. Harrison's principles of internal medicine. 18th ed. New York: McGraw-Hill; 2011.
[24] Hunt SA, Abraham WT, Chin MH, Feldman AM, Francis GS, Ganiats TG, et al. 2009 focused update incorporated into the ACC/AHA 2005 guidelines for the diagnosis and management of heart failure in adults: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Circulation 2009;119:e391–479, http://dx.doi.org/10.1161/circulationaha.109.192065.
[25] Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE, Drazner MH, 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 2013;62:e147–239, http://dx.doi.org/10.1016/j.jacc.2013.05.019.
[26] McMurray JJV, Adamopoulos S, Anker SD, Auricchio A, Böhm M, Dickstein K, et al. ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: The Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2012 of the European Society of Cardiology. Eur Heart J 2012;33:1787–847, http://dx.doi.org/10.1093/eurheartj/ehs104.
[27] Radford MJ, Arnold JMO, Bennett SJ, Cinquegrani MP, Cleland JGF, Havranek EP, et al. ACC/AHA key data elements and definitions for measuring the clinical management and outcomes of patients with chronic heart failure: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Data Standards (Writing Committee to Develop Heart Failure Clinical Data Standards). Circulation 2005;112:1888–916, http://dx.doi.org/10.1161/CIRCULATIONAHA.105.170073.
[28] Konstantinides SV, Torbicki A, Agnelli G, Danchin N, Fitzmaurice D, Galiè N, et al. 2014 ESC guidelines on the diagnosis and management of acute pulmonary embolism. Eur Heart J 2014;35:3033–69, http://dx.doi.org/10.1093/eurheartj/ehu283, 3069a–3069k.
[29] Colucci WS. Evaluation of the patient with heart failure or cardiomyopathy. UpToDate; 2013.
[30] Colucci WS. Treatment of acute decompensated heart failure: components of therapy. UpToDate; 2013.
[31] Colucci WS. Overview of the therapy of heart failure due to systolic dysfunction. UpToDate; 2013.
[32] South BR, Shen S, Leng J, Forbush TB, Duvall SL, Chapman WW. A prototype tool set to support machine-assisted annotation. In: Proc. 2012 Work. Biomed. Nat. Lang. Process. 2012. p. 130–9.
[33] Hripcsak G, Rothschild AS. Agreement, the F-measure, and reliability in information retrieval. J Am Med Inform Assoc 2005;12:296–8, http://dx.doi.org/10.1197/jamia.M1733.
[34] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
[35] Fort K, Sagot B. Influence of pre-annotation on POS-tagged corpus development. In: Proc. Fourth ACL Linguist. Annot. Work. 2010. p. 56–63.
[36] Lingren T, Deleger L, Molnar K, Zhai H, Meinzen-Derr J, Kaiser M, et al. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. J Am Med Inform Assoc 2014;21:406–13, http://dx.doi.org/10.1136/amiajnl-2013-001837.
[37] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: AMIA Symp. 2001. p. 17–21, doi:D010001275 [pii].

CHAPTER 4

GENERATING DISEASE-PERTINENT TREATMENT VOCABULARIES FROM MEDLINE CITATIONS

Reprinted with permission from Wang L, Del Fiol G, Bray BE, Haug PJ. Generating disease-pertinent treatment vocabularies from MEDLINE citations. Journal of Biomedical Informatics. 2017;65:46-57.

Journal of Biomedical Informatics 65 (2017) 46–57

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Generating disease-pertinent treatment vocabularies from MEDLINE citations

Liqin Wang [a,b,*], Guilherme Del Fiol [a], Bruce E. Bray [a,c], Peter J. Haug [a,b]

[a] Department of Biomedical Informatics, University of Utah, 421 Wakara Way, Salt Lake City, UT 84108, USA
[b] Homer Warner Research Center, Intermountain Healthcare, 5121 South Cottonwood Street, Murray, UT 84107, USA
[c] Department of Internal Medicine, University of Utah, 30 North 1900 East, Salt Lake City, UT 84132, USA

ARTICLE INFO

Article history:
Received 12 July 2016
Revised 4 October 2016
Accepted 15 November 2016
Available online 16 November 2016

Keywords:
Information extraction
Data mining
Ontology
MEDLINE citations
SemMedDB
Treatment

ABSTRACT

Objective: Healthcare communities have identified a significant need for disease-specific information. Disease-specific ontologies are useful in assisting the retrieval of disease-relevant information from various sources. However, building these ontologies is labor intensive. Our goal is to develop a system for an automated generation of disease-pertinent concepts from a popular knowledge resource for the building of disease-specific ontologies.

Methods: A pipeline system was developed with an initial focus of generating disease-specific treatment vocabularies. It was comprised of the components of disease-specific citation retrieval, predication extraction, treatment predication extraction, treatment concept extraction, and relevance ranking. A semantic schema was developed to support the extraction of treatment predications and concepts. Four ranking approaches (*i.e.*, occurrence, interest, degree centrality, and weighted degree centrality) were proposed to measure the relevance of treatment concepts to the disease of interest. We measured the performance of four ranks in terms of the mean precision at the top 100 concepts with five diseases, as well as the precision-recall curves against two reference vocabularies. The performance of the system was also compared to two baseline approaches.

Results: The pipeline system achieved a mean precision of 0.80 for the top 100 concepts with the ranking by *interest*. There were no significant different among the four ranks ($p = 0.53$). However, the pipeline-based system had significantly better performance than the two baselines.

Conclusions: The pipeline system can be useful for an automated generation of disease-relevant treatment concepts from the biomedical literature.

1. Introduction

Disease-specific ontologies are knowledge bases intended to structure and represent disease-relevant information including disease etiology, diagnostic characteristics, treatments and prognosis [1]. By providing rich domain knowledge, they can be very useful in assisting the retrieval of disease-relevant information from sources like clinical data repositories, biomedical literature, and online health resources, which therefore can better meet the information needs of various healthcare communities.

Building and maintaining such ontologies is labor-intensive. One major challenge is to identify disease-specific vocabularies that form the core of disease-specific ontologies. For example, in a previous study [1] we asked medical experts to manually develop reference vocabularies for three diseases from selected biomedical literature sources. The annotation of selected documents took around 100 man-hours, not counting the document preparation, guideline development, experts training, adjudication, and concept mapping. From the same study, we also found that existing literature sources were sufficient to provide disease-specific vocabulary. Therefore, there is an opportunity for the development of algorithms that can automatically extract vocabulary components from these sources.

In the present study, we address the challenge described above by developing a set of knowledge extraction techniques that automatically generate disease-pertinent vocabulary from existing sources. We chose the MEDLINE database as our knowledge source because it contains a large collection of published journal citations and covers a variety of diseases. We have focused on treatment concepts associated with the disease of interest, including direct

* Corresponding author at: Department of Biomedical Informatics, University of Utah, 421 Wakara Way, Salt Lake City, UT 84108, USA.
E-mail address: liqin.wang@utah.edu (L. Wang).

http://dx.doi.org/10.1016/j.jbi.2016.11.004
1532-0464/

treatment and prevention of the problem or complications caused by the problem; however, the method can also be adapted to other disease domains (*e.g.*, signs, symptoms, diagnostic tests).

2. Background

2.1. Disease-specific information needs and barriers

Disease-specific information is frequently sought by people in the healthcare communities, including clinicians, healthcare consumers, clinical researchers and medical knowledge engineers. The types of information that have been sought include medical knowledge (information that is understood to be generalizable to the care of all patients), patient data (information about a specific person), and population statistics (aggregated data about groups or populations of patients) [2]. For example, a variety of published studies investigated physicians' information needs by analyzing their clinical questions raised in the course of patient care [2–4]. A large proportion of the questions were related to disease-specific medical knowledge, such as "what is the drug of choice for condition x?", "what test is indicated in situation x?", and "how should I treat condition x?" [5]. Clinicians also frequently seek disease-specific patient information (*e.g.*, medical history, physical exam) from clinical data repositories. With the wide adoption of electronic medical records, available patient data has been shown a marked increase. Thus, it is important to be able to distil and filter medical records to show the patient information that is relevant to a specific problem of interest.

Healthcare consumers also frequently seek health information online to better understand and manage their own health [6]. Research shows that the top two major health topics searched online are related to the personal medical problems and the treatment for these problems [7]. Clinical researchers and medical knowledge engineers also demand disease-specific information in order to understand, model, and analyze clinical data. For example, when conducting a retrospective clinical study, clinical researchers need to understand the details of the clinical problem (*e.g.*, disease-specific signs and symptoms, diagnostic tests, comorbidities) in order to properly identify "research subjects" from an EHR.

2.2. Disease-specific ontologies

Ontologies are explicit and formal representations of domain knowledge, which enable the management, sharing, and reuse of domain knowledge [8,9]. Disease-specific ontologies intend to integrate vocabularies of different aspects of the disease, such as signs and symptoms, medications, therapeutic procedures, diagnostic procedures, and laboratory tests and imaging. To minimize information overload, it is crucial to develop effective information retrieval systems capable of retrieving relevant information to meet different information needs. For healthcare consumers, who are likely to have low health literacy [10], it is important to assist them forming optimal queries to retrieve relevant information from online health sources [11]. We anticipate that these ontologies will facilitate the retrieval of specific information from a variety of sources, such as websites [12], biomedical literature [13], and clinical data repositories [14–16]. Disease-specific ontologies can support information retrieval systems by providing domain-specific concepts and relations necessary to direct the formulation or expansion of initially simple queries tied to clinical concepts. In addition, the medical knowledge contained in disease-specific ontologies could be used by clinical researchers and medical knowledge engineers to understand the diseases, and the vocabularies in the ontologies may further assist their research or engineering work (*e.g.*, cohort selection, text annotations).

2.3. Relation extraction in biomedical domain

Domain experts can develop disease-specific ontologies, but individuals with the required expertise are scarce and expensive. A long-term goal of our research is to create a platform to facilitate large-scale development of these ontologies. One of the critical tasks in building disease-specific ontologies is to acquire medical knowledge like concepts and relationships related to the disease of interest [8,17]. This kind of medical knowledge has been substantially documented in sources like the biomedical literature, web documents, and clinical data repositories, although most of it is represented in unstructured and narrative format. We therefore hope to take advantage of these sources and investigate automatic techniques to extract disease-specific medical knowledge from them.

Automatic extraction of relational medical knowledge from the biomedical literature is an active subject of research interest [18–22]. Researchers have attempted to extract disease-specific medical knowledge from the biomedical literature ever since the 1990s [23,24]. In the earliest stage, the methods merely relied on co-occurrence-based statistics. For examples, Zeng and Cimino used MeSH co-occurrence information from the UMLS to obtain disease-chemical associations [24]. Chen et al. used co-occurrence statistics to extract disease-drugs relations from MEDLINE abstracts [18].

Along with the advanced development of NLP techniques, a variety of rule-based and machine-learning-based methods have been used for relation extraction. A typical example of rule-based system is SemRep [25,26] which is built upon UMLS and MetaMap. It interprets the biomedical knowledge presented in a given sentence from the scientific literature in the form of predications {subject PREDICATE object}, where the subject and object are biomedical concepts from the UMLS Metathesaurus and the PREDICATE is a semantic relation from the UMLS Semantic Network [25–27]. For example, from the sentence "this paper will review the earlier and present studies in the development of rasagiline for treatment of PD and discuss its pharmacology and applicable mechanism of action", SemRep extracts the predication {Rasagiline TREATS Parkinson's disease}. Based on a preliminary evaluation, the precision and recall of SemRep are 78% and 49% respectively [26]. More recently, Xu and Wang applied a pattern-based approach to extract disease-drug and disease-disease risk relationships from MEDLINE citations [21,22].

Machine learning techniques have also been successfully applied to relation extraction. From one standpoint, relation extraction is a classification problem which is to predict semantic relations held between two identified entities in a given sentence [28]. Researchers have employed different classification models using diverse lexical, syntactic and semantic features derived from the text to make predication on the relations. For example, Rosario and Hearst compared graphical models and neural network using lexical, syntactic, and semantic features to distinguish among seven relation types that can occur between the entities "treatment" and "disease" in bioscience texts [29]. Zeng et al. exploited a convolutional deep neural network to extract lexical and sentence level features which were fed into a softmax classifier to predict the relationships between two marked nouns [30]. From another standpoint, relation extraction is a sequence labeling problem, for which researchers have applied kernel-based approaches to label the relationships between two entities. For example, Bundschus et al. used conditional random field technologies to extract disease-treatment associations from PubMed abstracts [31]. Giuliano et al. investigated a kernel-based approach based on shallow linguistic processing for extracting relations between entities from biomedical literature [32].

In the present work, we intend to develop an automated approach to extract treatment vocabularies from the biomedical literature for a given disease of interest. Unlike previous studies which worked on semantic interpretation of the relationships from the biomedical literature, we focused on filtering and ranking disease-specific concepts for a given disease of interest. In addition, our work builds on previous tools and methods, in particular SemRep.

2.4. SemMedDB

SemRep is routinely used to process the entire set of MEDLINE citations (*i.e.*, the titles and abstracts) to extract structured predications, which are then stored in a repository called SemMedDB [33]. There are currently over 83 million semantic predications in this database version June 30, 2015, approximately 93% of which are associative (or, non- "IS-A" predication). Although SemMedDB provides structured predications that could facilitate the acquisition of medical knowledge from the biomedical literature, further inference is needed to filter noisy data and to retrieve information that is most useful for a disease-specific ontology. For example, a query in SemMedDB for a collection of predications that include congestive heart failure retrieves thousands of predications. Within these predications, concepts may range widely from pharmaceutical substances to signs and symptoms and related genes. Therefore, many retrieved concepts and predications could be outside scope for a disease-specific ontology. In addition, concepts that are irrelevant to the main search topic may be retrieved due to errors in the underlying SemRep NLP process and inaccurate or outdated information presented in MEDLINE abstracts. We addressed these issues in the development of our automatic knowledge extraction system from SemMedDB.

3. Materials and methods

The study method is comprised of two parts: (1) the development of a pipeline-based process to extract disease-specific, treatment-related information from biomedical literature; and (2) an experiment to compare the pipeline-based process to extract disease-specific treatment vocabulary with two baseline approaches in terms of precision-recall curves and mean average precision.

3.1. Pipeline-based process

The pipeline-based process developed in the present study consists of the following steps (see Fig. 1): (1) retrieval of therapeutic citations from MEDLINE for the disease of interest using a search strategy that aims to retrieve scientifically sound studies; (2) retrieval of all predications and their corresponding sentences from SemMedDB for the citations retrieved in Step 1; (3) development of a semantic schema from the UMLS and existing disease-specific ontologies to identify treatment-related predications from this list; (4) retrieval of treatment-related predications from the predications in Step 2 using the semantic schema from Step 3; (5) extraction of treatment concepts from the treatment predications extracted in Step 4 from the list generated in Step 3; (6) ranking of the treatment concepts extracted in Step 5 using four ranking algorithms.

3.1.1. Step 1: Retrieval of disease-pertinent MEDLINE citations

The first step retrieves biomedical citations from MEDLINE database regarding the therapy of a given disease. We built a search strategy based on the PubMed Clinical Queries, which is a set of filters that are tuned to retrieve scientifically sound clinical studies in topics such as treatment, diagnosis, and prognosis [34–36]. The Clinical Query filters provide two modes: *broad* and *narrow*. The *broad treatment* filter has shown a sensitivity of 99% and a specificity of 70%, while the *narrow treatment* has shown a sensitivity of 93% and specificity of 97%. In the present study, we focused on sensitivity and used the *broad* filter.

Although Clinical Query filters perform well in retrieving clinical trial studies, the query does not cover other types of study design, such as systematic reviews, which would also be useful for retrieving disease-specific medical knowledge. Hence, we extended the Clinical Query treatment filter to retrieve systematic review articles (see Fig. 2). In addition, we added the following restrictions: *English language*, *abstract available*, *human subjects*, and *core clinical journals*. We obtained the list of clinical journals by combining the PubMed core clinical journals (http://www.nlm.nih.gov/bsd/aim.html) with a list of journals categorized under "clinical medicine" in Web of Science (http://ip-science.thomsonreuters.com/mjl/). For each disease of interest, we added a MeSH term for the disease as a major topic. The modified Clinical Query filter can also be extended to retrieve articles for other disease-associated concepts, such as etiology, diagnosis, and prognosis.

3.1.2. Step 2: Predication extraction with SemRep

In this step, the input is all the PMIDs that were assigned to those MEDLINE citations retrieved from step 1. The output is the predications generated by the SemRep from those MEDLINE citations as well as the sentences where the predications came from. More specifically, we took all the PMIDs to form SQL scripts to query the SemMedDB [33] to retrieve all the predications and sentences. The version of SemMedDB we used was updated with citations published through June 30, 2015. Citations published after this date were not yet available in SemMedDB, therefore we excluded those citations from the study.

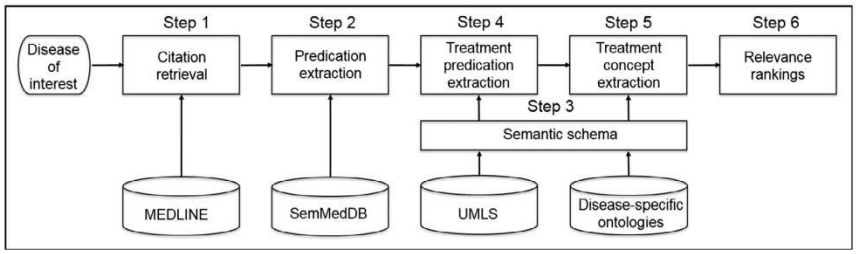

Fig. 1. Flowchart of automatically extracting disease-specific, treatment vocabulary from the biomedical literature and the ranking of treatment concepts.

```
(Therapy/Broad[filter] OR (systematic[sb] AND ("systematic review"[ti] OR "meta-analysis"[ti]
OR "Cochrane Database Syst Rev"[journal]))) AND "QUERY_TERM"[Majr] AND
"humans"[MeSH Terms] AND "english"[language] AND (hasabstract[text]) AND
(JOURNALLIST)
```

Fig. 2. Modified Clinical Query for retrieving treatment-related citations for the disease of interest from MEDLINE. In the query, "QUERY TERM" is the MeSH term for the disease of interest. "JOURNALLIST" is a list of clinical journals, *e.g.*, "CA-CANCER J CLIN", "NEW ENGL J MED".

Table 1
The semantic types and groups of treatment concepts.

Semantic groups	Included semantic types
Procedures	Educational Activity, Health Care Activity, Therapeutic or Preventive Procedure
Chemicals & Drugs	All[a]
Activities & Behaviors	All[a]
Devices	Medical Device

[a] Refer to http://semanticnetwork.nlm.nih.gov/download/SemGroups.txt for detailed semantic types included by a specific semantic group.

3.1.3. Step 3: Development of semantic schema

The semantic schema consisted of a set of metapredications whose arguments are defined based on high-level domains based on UMLS semantic groups [37]; for example, {*Procedures* TREATS *Disorders*}. The development of a semantic schema is a one-time process that supports knowledge extraction of treatment concepts for any disease of interest. The development of the semantic schema was performed in two steps: (1) selection of relevant semantic groups to filter treatment concepts (Step 4), and (2) definition of relevant metapredications to filter treatment predications (Step 5).

To select relevant semantic groups, we analyzed the semantic groups and types that were present in the heart failure reference vocabulary (http://bioportal.bioontology.org/ontologies/HFO) that had been manually created in a previous study [1]. The rationale for this approach is the assumption that the majority of semantic groups and types covered in disease treatment vocabularies would also be covered in the heart failure reference vocabulary.

A total of 413 treatment concepts were retrieved, from 38 semantic types and 9 semantic groups (*i.e.*, *Chemicals & Drugs*, *Procedures*, *Physiology*, *Devices*, *Activities & Behaviors*, *Concepts & Ideas*, *Objects*, *Disorders*, and *Organizations*). The majority of the heart failure treatment concepts belonged to two semantic groups: *Chemicals & Drugs* and *Procedures*. We manually reviewed the other seven semantic groups and, based on domain knowledge, decided to include only four semantic groups: *Chemicals & Drugs*, *Procedures*, *Devices*, and *Activities & Behaviors* (Table 1). We also excluded a subset of the semantic types from the *Procedures* and *Devices* semantic groups. For example, from *Procedures*, we excluded *Diagnostic Procedure*, *Laboratory Procedure*, *Molecular Biology Research Technique*, and *Research Activity*.

We followed a similar process for metapredications, also using the heart failure vocabulary.

We retrieved a total of 54,991 predications from SemMedDB from 15,994 citations. Forty percent (N = 22,019) of the predications contained treatment concepts from the heart failure vocabulary. We then generated 205 unique metapredications based on the retrieved predications, such as {Chemicals & Drugs, ADMINISTERED_TO, Living Beings}. Next, we removed the metapredications that did not contain any of the four semantic groups selected in the previous step. In addition, we excluded metapredications whose predicate was not treatment-related predicates, such as DIAGNOSES, CAUSES, STIMULATES, PRODUCES, PREDISPOSES, as well as negation predications. The remaining metapredications were grouped into four categories (Table 2). For each category, we identified the predication arguments that were most relevant for extracting treatment concepts. However, we noted some exceptions. For example, in category 3, for metapredications where the arguments are *Chemical & Drugs* and *Devices*, their corresponded predications are usually about the comparison or co-occurrence of a treatment (*Chemical & Drugs*) with a "placebo" (*Devices*), therefore, only the concepts from the position of *Chemical & Drugs* will be retrieved.

3.1.4. Step 4: Extraction of relevant treatment predications

Many predications retrieved in Step 2 could be not related to the treatment (*e.g.*, a predication {congestive heart failure CAUSES cardiomyopathy, dilated}), or were generic and of little interest (*e.g.*, {pharmaceutical preparations TREATS pneumonia}). To filter out generic predications, we adopted the *novelty* approach proposed by Fiszman et al. [38]. A predication is considered as generic when it has a generic concepts which is determined by whether the hierarchical depth in the Metathesaurus is less than an empirical distance. Each concept of the predications has the attribute of *novelty* in the SemMedDB. We exclude predications that contain non-novel concepts.

We then used the semantic schema to separate the treatment predications from irrelevant predications. To do so, we excluded predications that did not match one of the metapredications. For example, the predication {Adrenergic beta-Antagonists PREVENTS heart failure} matches the metapredication {Chemicals & Drugs PREVENTS Disorders}, while predication {congestive heart failure CAUSES cardiomyopathy, dilated} does not match any metapredications in the semantic schema.

Table 2
Semantic schema for classifying treatment predications. The predication arguments in underline are the ones from which treatment concepts are extracted.

Category	Subject	Relation	Object
1	Chemicals & Drugs/Procedures/Devices/ Activities & Behaviors	ADMINISTERED TO/AUGMENTS/AFFECTS/ASSOCIATED WITH/ DISRUPTS/INHIBITS/TREATS/PREVENTS	ANY semantic groups
2	ANY semantic groups	USES	Chemical & Drugs/Procedures/Devices/ Activities & Behaviors
3[a]	Chemicals & Drugs/Procedures/Devices/ Activities & Behaviors	COEXISTS WITH/compared with/same as/INTERACTS WITH/ METHOD OF/lower than/higher than	Activities & Behaviors/Drugs/ Procedures/Devices
4	Chemicals & Drugs/Procedures/Devices/ Activities & Behaviors	ISA	Activities & Behaviors/Drugs/ Procedures/Devices

[a] For metapredications where the subject is *Chemical & Drugs* and the object is *Devices*, and vice-versa, only *Chemical & Drugs* concepts are extracted.

3.1.5. Step 5: Extraction of disease-specific treatment concepts

After obtaining treatment predications, we extracted the concepts in the subject or object according to the semantic schema in Table 2. However, these extracted concepts could still be too general for the disease of interest. To exclude general concepts, we used an approach based on the assumption that concepts associated with a large number of diseases (*i.e.*, common concepts) are likely to be general.

In order to identify common concepts, we took all MeSH terms (from UMLS Version 2014AB) with the semantic type of *disorders* (N = 5109), and repeated Steps 1, 2, and 4 above to generate disease-treatment pairs. A subset of 2683 MeSH terms were associated with disease-treatment pairs. Then, we analyzed the retrieved treatment concepts and the number of associated disorders for each treatment concept. If a treatment concept was associated with more than an arbitrary threshold of 20% of disease MeSH terms (N = 536), the concept was considered to be a common concept. Applying this criterion, we generated a set of 69 common concepts. Table 3 shows examples of common concepts.

3.1.6. Step 6: Concept ranking

Ranking concepts has three purposes. First, the ranking might convey the information of the strength of the association. As we know, some treatment concepts might have stronger association with the disease of interest. For example, both "carvedilol" and "fish oil" are retrieved as treatment of heart failure, however, "carvedilol" is mentioned much more frequently in the literature than fish oil as a treatment of heart failure. Second, ranking concepts could make the true relevant concepts appear earlier in the result list than the noise. Although the semantic schema are able to filter some treatment-irrelevant information, noisy information can still be introduced because the semantic schema was focused on sensitivity. For example, given a disease of interest (*i.e.*, heart failure), we extracted a treatment predication {Trastuzumab TREATS Breast cancer metastatic}, where the concept "Trastuzumab" was discussed as a cause of heart failure rather a treatment. Last but not least, a ranked list could speed up the review of automatically extracted concepts. The knowledge authors could prioritize their work with the ranked output.

We explored four approaches to rank the concepts: *occurrence*, *interest*, *degree centrality*, and *weighted degree centrality*.

(1) **Occurrence**: the frequency of the occurrence of a treatment concept in the retrieved treatment predications for a given disease of interest (Formula (1)). The assumption is that the more often a concept is mentioned in the context of disease-specific treatment predications, the stronger the confidence that it is as a treatment for the disease of interest.

$$Occurrence(t_i, d) = a_i \tag{1}$$

Table 3
Sampled common concepts.

CUI	UMLS concept	# of co-occurred diseases
C0040808	Treatment Protocols	1445
C1273870	Management procedure	1418
C1273869	Intervention regimes	1361
C0011900	Diagnosis	1326
C1533685	Injection procedure	1265
C0543467	Operative Surgical Procedures	1248
C0184661	Procedures	1201
C0032042	Placebos	1193
C0001617	Adrenal Cortex Hormones	1172
C0728940	Excision	1091
C1522577	Follow-up	1083
C0185125	Application procedure	1064
C0023977	Long-term care	1041
C0220908	Screening procedure	989

where a_i is the frequency of the occurrence of a concept t_i in the treatment predications.

(2) **Interest**: A treatment concept may have a high *occurrence* score among the other extracted treatment concepts simply because it frequently occurs in the entire database. However, the relation between the concept and the disease of interest can still be weak. *Interest* is a measure that attempts to correct this weakness of *occurrence*, the idea of which is very similar to the TF-IDF (term frequency inverse document frequency) – a statistic that is intended to reflect how important a word is to a document in a collection of corpus [39]. We define the *interest* is the ratio of the *occurrence* of a treatment concept to the sum of the *occurrence* of all treatment concepts retrieved for a given disease of interest divided by logarithm of the ratio of the occurrence of a treatment of interest to all treatment concepts in the database (see Formula (2)). The denominator is a simple way of measuring the commonality of a concept.

$$Interest(t_i, d) = \frac{a_i / \sum_i^M a_i}{\log\left(Ai / \sum_i^M Ai\right)} \tag{2}$$

where a_i is the frequency of the occurrence of a concept t_i in the treatment predications, A_i is the total frequency of the occurrence of the concept t_i in the entire database, while M is the total number of retrieved treatment concepts.

(3) **Degree centrality**: Occurrence-based statistics ignore the linkage between concepts. Since the treatment predications extracted in step 4 can form a graph, we analyzed the formed network and use the centrality to identify important vertices (*i.e.*, treatment concepts) within the graph. Degree centrality is the simplest of many centrality approaches, which measures the significance of the concepts in the graph by counting their connectivity to other concepts. We do not look at whether a concept is directly connected to the disease of interest or not; rather, we assess whether concepts are in the center of the graph. The following formula was used to calculate the degree centrality of a given concept in the graph:

$$C_D(i) = \sum_j^N x_{ij} \tag{3}$$

where i is the focal node, j represents all other nodes, N is the total number of nodes, and x is the adjacency matrix, in which the cell x_{ij} is defined as 1 if node *i* is connected to node *j*, and 0 otherwise. Zhang et al. have used degree centrality for semantic abstraction summarization of therapeutic studies, in which degree centrality was used to select important nodes from a graph [37]. Özgür *et al.* also used degree centrality for mining gene-disease association from biomedical literature [40].

(4) **Weighted degree centrality**: Weighted degree centrality is a harmonization between the frequency of occurrence and degree centrality [41].

$$\begin{aligned} C_D^{w\alpha}(i) &= k_i \times \left(\frac{s_i}{k_i}\right)^{\alpha} = k_i^{1-\alpha} \times s_i^{\alpha} \\ k_i &= C_D(i) \\ S_i &= C_D^{w}(i) = \sum_j^N w_{ij} \end{aligned} \tag{4}$$

where k_i is the degree centrality score of node i, or $C_D(i)$ as described in Formula (3). S_i is the sum of weighted adjacency matrix in which w_{ij} is the value that represents the weight of the edge (*i.e.*, the occurrence of a predication) between node *i*

and node j. α is a positive tuning parameter that can be set according to the research setting and data. We used $\alpha = 0.5$ in this study to harmonize the occurrence and the degree centrality in one ranking.

3.2. Experiment

We conducted an experiment to test the following *null* hypotheses: there is no difference in precision at top 100 extracted concepts among the rankings produced by the four ranking approaches in the pipeline-based algorithms (H1); and there is no difference in precision at top 100 extracted concepts among the rankings produced by the pipeline, predication, and MeSH-based extraction methods (H2). In addition, we also evaluated the performance of the system against the manually extracted treatment vocabulary with precision-recall curves.

3.2.1. Baseline approaches

We compared our approach with two baselines in terms of extracting disease-specific treatment concepts from MEDLINE citations.

Baseline 1: The Medical Subject Headings (MeSH) vocabulary is used to index and catalog articles in MEDLINE. MeSH *qualifier terms*, in conjunction with the MeSH main headings, offer a convenience to group citations together when they are related to a particular aspect of a subject. For example, *Platelet Aggregation Inhibitors/therapeutic use* indicates that the citation is about the use of the drug class *platelet aggregation inhibitors* in the treatment of a disease. After reviewing the qualifiers defined in the MeSH Topical Qualifiers [42] and examples in the MEDLINE database of how those qualifiers were used with the MeSH headings, we selected the following qualifiers: "methods", "instrumentation", "therapeutic use", "pharmacology", and/or "administration & dosage". For example, the qualifier "administration & dosage" is defined as "used with drugs for dosage forms, routes of administration, frequency and duration of administration, quantity of medication, and the effects of these factors.", a drug MeSH term could be possibly assigned with the qualifier "administration & dosage". Based in their definition, the qualifiers "methods" and "instrumentation" were used with procedures and techniques, including diagnostic procedures and therapeutic procedures. The qualifiers "therapeutic use", "pharmacology", and/or "administration & dosage" were used with drugs or chemical substances.

From the articles retrieved by Step 1, we were able to extract a collection of MeSH terms associated with the therapeutic qualifiers of interest. We then obtained the UMLS concepts for these MeSH terms using the mappings established in the UMLS Metathesaurus. Next, the resulting UMLS concepts were restricted using the same semantic types and groups described in Table 1 in order to avoid the inclusion of concepts not related to treatment. The remaining concepts were ranked based on their frequency of occurrence.

Baseline 2: This baseline approach simply used the predications to obtain disease-specific treatment concepts. We first extracted the predications with the pattern of {Subject TREATS/PREVENTS Object}, where the object is the disease of interest. We then extracted all the concepts in the subject position. Thereafter, we ranked the concepts based on their frequency of the occurrence in the retrieved predications.

3.2.2. Validation of extracted concepts

We selected five diseases cases for hypothesis testing. Two diseases, pulmonary embolism (PE) and rheumatoid arthritis (RA), were chosen from a previous study, for which we have developed reference treatment vocabularies with 80 and 232 concepts respectively. The reference vocabularies are available in BioPortal as rheumatoid arthritis ontology (https://bioportal.bioontology.org/ontologies/RAO) and pulmonary embolism ontology (https://bioportal.bioontology.org/ontologies/PE). The other three diseases (diabetes mellitus, asthma, and schizophrenia) were chosen from a previous publication on knowledge extraction from existing knowledge resources [18].

In order to measure the performance of different knowledge extraction approaches, we validated the extracted concepts for the selected diseases. This was done by comparing to reference standards (for the two diseases with reference standards) and manual review.

For automated comparison to reference standards, we used exact matching and one-way hierarchical matching where any extracted concepts that were children of reference concepts were considered as positive. The hierarchical relationships were obtained from the UMLS Metathesaurus MRREL and MRHIER tables.

For manual review, the goal was to verify if false-positive concepts according to the reference standard were indeed true-positives or just gaps in the reference standard. For example, "tumor necrosis factor-alpha inhibitor" (a drug class used to treat rheumatoid arthritis) was extracted by our system as a treatment for rheumatoid arthritis. However, this drug class was not present in the reference standard. Upon review one of the source sentences: "Tumour necrosis factor-alpha (TNFalpha) inhibitors are effective agents in treating RA; however, their cost effectiveness as first-line agents has not been investigated", we confirmed that "tumor necrosis factor-alpha inhibitor" is indeed a treatment for rheumatoid arthritis. This review was done by one of the authors (LW) with additional clinician review if such judgement could not be made directly based on the source sentences.

3.2.3. Outcome measures

The primary outcome for the two hypotheses was precision at K and secondary outcomes were the overall precision and recall. Precision at K was the ratio of the number of "true positive" concepts among the top K ranked concepts divided by K. We calculated the precision at K for five testing diseases for different rankings and algorithms. We choose the parameter $K = 100$, believing that as knowledge engineers, it is a fair amount of concepts that they would go through. When calculating the precision at K, for diseases having reference standards, we not only validated the extracted concepts with the reference standards, but also manually verified false positive concepts in case they were in fact correct concepts, but missing in the reference standard. For three diseases without reference standards, the top 100 concepts of each disease were manually validated.

To evaluate ranked results, interpolated precision-recall curves were plotted to visualize the trade-off between precision and recall, where the precision and recall were calculated based on the reference standards. The precision-recall curves also provided a visual comparison among the ranks in the pipeline-based approach and between the pipeline-based approach and the baselines. We plotted the interpolated precision-recall curves only for the two diseases with reference vocabularies. An error analysis were also conducted based on manual inspection of false-positive and false-negative concepts.

3.2.4. Statistical analysis

To test the difference among the different rankings in the pipeline-based system (H1), we first measured the top 100 precision obtained by four different rankings for five diseases. We then calculated the mean precision for each ranking. We used analysis of variance (ANOVA) to test the significance of the difference. For pairwise comparisons, we used the Tukey honest significant difference (HSD) post-hoc test.

Table 4
The numbers of retrieved citations, predications, treatment predication, and treatment concepts for five testing diseases.

Test cases	Citations	Predications	Treatment predications	Candidate treatment concepts
Rheumatoid arthritis	11,263	53,039	26,914	1984
Pulmonary embolism	3031	12,820	5101	706
Diabetes mellitus	32,552	166,140	72,730	3873
Asthma	17,286	94,001	39,189	2385
Schizophrenia	6910	25,086	14,701	1018

Table 5
Exampled output for rheumatoid arthritis with ranking scores and sample source sentences.

CUI	Concept	Semantic type	Occurrence	Interest	DC	WDC	Source sentences
C0025677	Methotrexate	Phsu	4102	0.3095	394	1271.29	CONCLUSIONS: This study confirms previous observations from a dose-ranging study showing that anakinra, in combination with MTX, is an effective and safe treatment for patients with RA who have inadequate responses to MTX alone
C0666743	Infliximab	Phsu	1974	0.1724	212	646.91	Infliximab therapy was also associated with improvements in health-related quality of life in patients with Crohn's disease or rheumatoid arthritis
C0717758	Etanercept	Aapp	1313	0.1263	148	440.82	CONCLUSION: Etanercept as monotherapy was safe and was superior to MTX in reducing disease activity, arresting structural damage, and decreasing disability over 2 years in patients with early, aggressive RA
C0242708	Antirheumatic Drugs, Disease-Modifying	Phsu	940	0.1002	158	385.38	Early diagnosis and treatment with disease-modifying antirheumatic drugs (DMARDs) are necessary to reduce early joint damage, functional loss, and mortality
C0393022	Rituximab	Aapp	1004	0.0844	125	354.26	CONCLUSIONS: Evidence from RCTs suggests that RTX and ABT are more effective than supportive care

DC = degree centrality; WDC = weighted degree centrality; phsu = pharmaceutical substance; aapp = amino acid, peptide, or protein; MTX = methotrexate; RA = rheumatoid arthritis; RTX = rituximab; ABT = abatacept; RCTs = randomised controlled trials.

Table 6
Top 100 precision for treatment concepts extracted for five diseases.

Diseases	Top 100 precision					
	Pipeline-based				B1	B2
	Occurrence	Interest	DC	WDC		
Rheumatoid arthritis	0.87	0.89	0.82	0.84	0.56	0.76
Pulmonary embolism	0.63	0.66	0.65	0.63	0.31	0.38
Diabetes mellitus	0.78	0.82	0.66	0.75	0.46	0.54
Asthma	0.8	0.81	0.76	0.81	0.54	0.66
Schizophrenia	0.81	0.83	0.74	0.77	0.31	0.62
Mean precision	0.78	0.80	0.73	0.76	0.44	0.59
Std. deviation	0.089	0.085	0.071	0.081	0.121	0.142
95% Confidence interval	(0.67, 0.89)	(0.70, 0.91)	(0.638, 0.814)	(0.66, 0.86)	(0.29, 0.59)	(0.41, 0.77)

DC = degree centrality; WDC = weighted degree centrality; B1 = MeSH-based baseline; B2 = predication-based baseline.

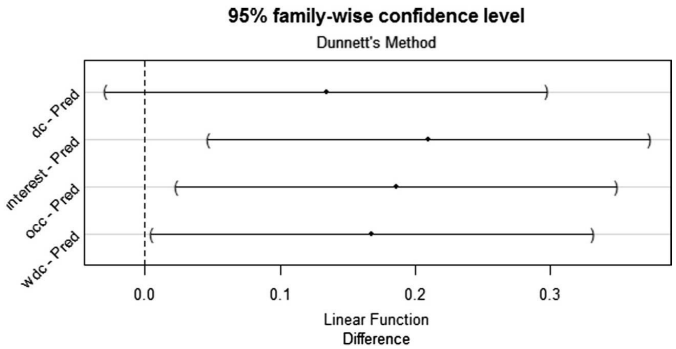

Fig. 3. 95% family-wise confidence level for the difference of the precision of top 100 concepts between the pipeline-based system and the Predication-based system.

To test the difference between the pipeline-based system vs. predication-based system and the pipeline-based system vs. the MeSH-based approach (H2), we calculated the mean top 100 precision for the two baselines across the same five diseases. We used ANOVA to test the significance of the differences between pipeline-based system and predication-based approach, followed

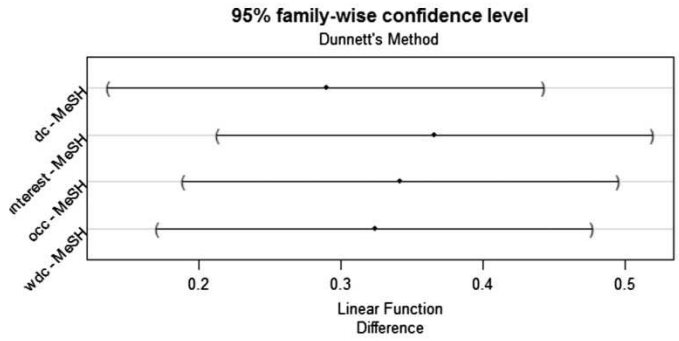

Fig. 4. 95% family-wise confidence level for the difference of the precision of top 100 concepts between the pipeline-based system and the MeSH-based system.

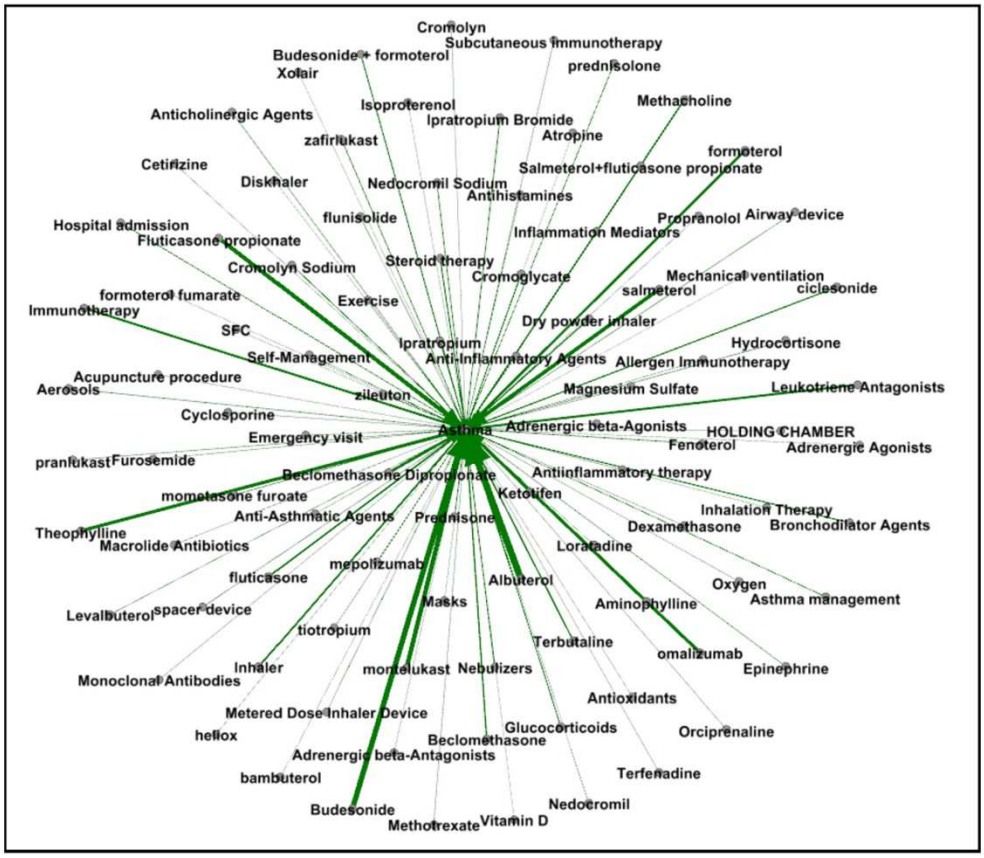

Fig. 5. Weighted graph of exampled treatment concepts for asthma.

by the Dunnett post-hoc test for comparisons between the four ranks in the pipeline-based system with the control (or the baseline). In the same way, we tested the significance of difference between the pipeline-based system and the MeSH-based approach. All statistical analyses were based on a significance level of 0.05 and were performed with R version 3.2.5.

4. Results

4.1. System outputs on five diseases

Table 4 shows the number of citations, predications, treatment predications, and treatment concepts retrieved from each step for

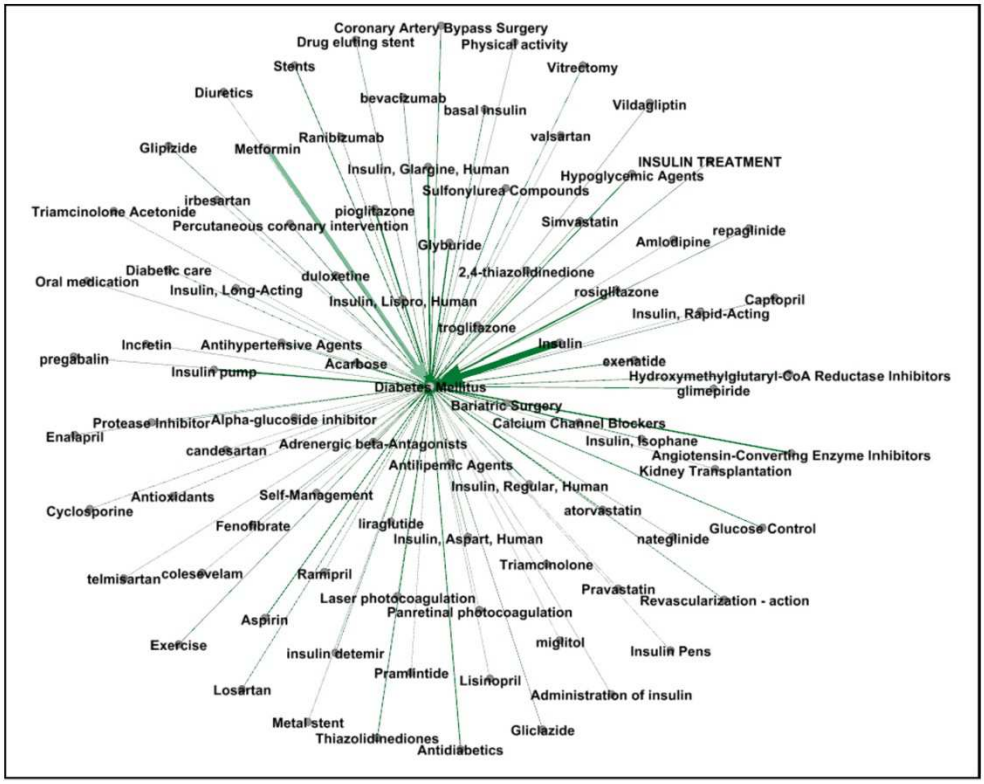

Fig. 6. Weighted graph of exampled treatment concepts for diabetes mellitus.

the five test diseases. The number of retrieved citations varied by disease. On average, each citation was able to generate 4–5 predications, and less than half of those predications were treatment predications. The number of candidate treatment concepts also varied based on the disease of interest.

Table 5 shows sample output from the pipeline-based system for rheumatoid arthritis. The output consists of the following attributes: UMLS CUI, concept name, semantic type, four ranking scores (*occurrence*, *interest*, *degree centrality*, and *weighted degree centrality*), and sentences extracted from the abstract and titles of the published articles.

4.2. Performance of pipeline-based algorithms versus baselines

Table 6 shows the precision of the top 100 treatment concepts extracted by the pipeline system and baselines on five diseases: rheumatoid arthritis, pulmonary embolism, diabetes mellitus, Alzheimer's disease, and asthma.

In the pipeline-based approaches, the difference among *occurrence*, *interest*, *degree centrality*, and *weighted degree centrality* was not significant (mean top 100 precision = 0.78 vs. 0.80 vs. 0.73 vs. 0.76; $p = 0.53$).

According to the ANOVA test, there was a significant difference in mean precision at top 100 among the pipeline-based and predication-based approaches (*occurrence* 0.78 vs. *interest* 0.80 vs. *degree centrality* 0.73 vs. *weighted degree centrality* 0.76 vs. predication-based 0.59; $p = 0.022$). With the HSD post-hoc test, three ranks (*i.e.*, *interest*, *occurrence*, and *weighted degree centrality*) in the pipeline-based system significantly outperformed the predication-based baseline (see Fig. 3), while no significant difference was found between the degree centrality and the predication-based baseline. According to the ANOVA test, there was a significant difference in mean precision at top 100 among the pipeline-based and the MeSH-based baseline (*occurrence* 0.78 vs. *interest* 0.80 vs. *degree centrality* 0.73 vs. *weighted degree centrality* 0.76vs. MeSH-based 0.44; $p < 0.0001$). With the HSD post-hoc test, the pipeline-based approach with all four ranks significantly outperformed the MeSH-based approach (see Fig. 4).

Figs. 5 and 6 provide a visualization of the treatment vocabularies generated by the pipeline-based system for asthma and diabetes.

4.3. Precision-recall curves

The precision-recall curves compared the performance of the different approaches against the manually developed reference vocabularies. Fig. 7 shows the interpolated precision-recall curves on *rheumatoid arthritis* and *pulmonary embolism*. By including all extracted concepts, the recall of rheumatoid arthritis was 0.59, and the recall of pulmonary embolism was 0.66. Recall for the pipeline based approach was less than 1 for both diseases, indicating that the automated system captured only a subset of

the concepts in the gold standard. The predication-based baseline approach reached a recall of 0.58 for rheumatoid arthritis and 0.56 for pulmonary embolism while, the MeSH-based baseline reached a recall of 0.34 for both pulmonary embolism and rheumatoid arthritis.

4.4. Error analysis

We identified 143 false negative concepts for rheumatoid arthritis, and 43 false negative concepts for pulmonary embolism. All these false negative concepts were included in the error analysis. We identified over two thousand false positive concepts for these two diseases and analyzed the false positive concepts among the top 100 ranked concepts of each disease retrieved by any of the ranks, which resulted in 47 false positive concepts for rheumatoid arthritis and 76 for pulmonary embolism.

Three main reasons could be attributed to false negative concepts or lowered recall: (1) about one third of the reference concepts were not present in the extracted sentences and predications (*e.g.*, "fluindione" and "lanoteplase" for pulmonary embolism). A few false negative concepts were missed because their semantic types were not included in the semantic schema of the automated system, such as 'systemic' and 'nutritional'. (2) One third of the reference concepts existed in the extracted citations and sentences, however were missed because they were not captured by SemRep. For example, in "Tai Chi and yoga are complementary therapies which have, during the last few decades, emerged as popular treatments for rheumatologic and musculoskeletal diseases" two predications were extracted: {Complementary therapies TREATS Rheumatologist} and {Complementary therapies TREATS Musculoskeletal Diseases}; however, none of the predications included the relevant concepts "Tai Chi" and "yoga". (3) One third of reference concepts were missed because equivalent annotations were mapped to UMLS CUIs with different granularity in the reference vocabulary. For example, 'resistance training' was mapped to C0872279 (Resistance Training) in the reference standard, but was mapped to C0814409 (Resistance education) in SemMedDB. The reference was more likely to include the entire annotation as a concept while SemRep mapped more granular fragments to UMLS concepts. For example, from the sentence "in this systematic review, outcomes for total wrist fusion were comparable and possibly better than those for total wrist arthroplasty in rheumatoid patients", SemRep extracted the predication {Arthroplasty TREATS Patients}, while in the reference the "total wrist arthroplasty" was mapped to C0408314 (total wrist arthroplasty).

Several reasons were attributed to false positive concepts or lowered precision. (1) Among the analyzed false positive concepts, 40% were correct disease-specific treatments that were missing in the reference vocabularies. Examples include "methotrexate treatment", "tumor necrosis factor therapy", and "Hip Replacement, Total" for rheumatoid arthritis; and "Prescription of prophylactic

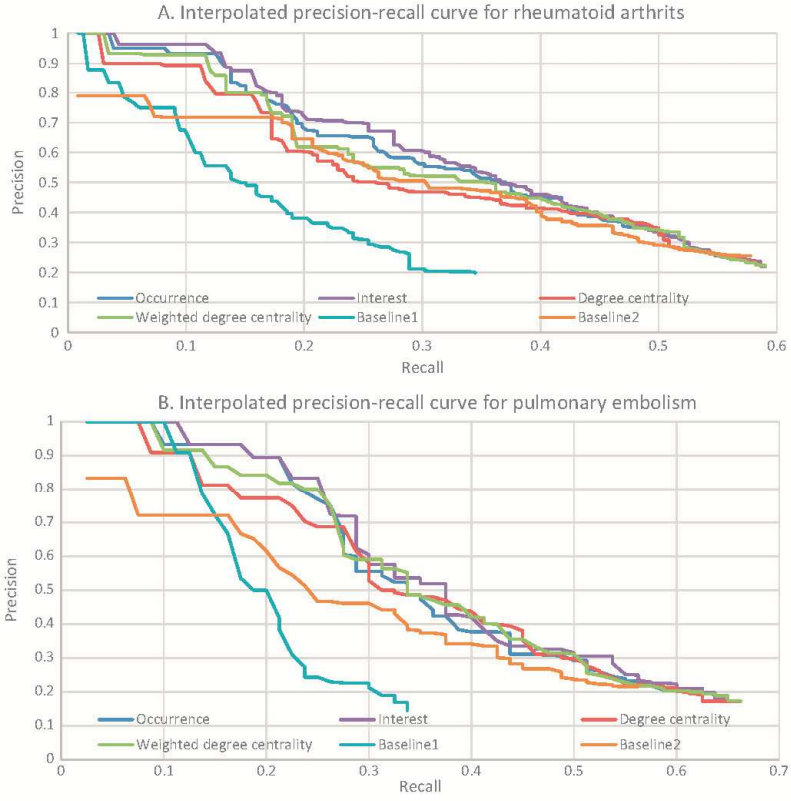

Fig. 7. (A) Interpolated precision-recall curves for rheumatoid arthritis; (B) Interpolated precision-recall curves for pulmonary embolism.

anticoagulant", "Prescription of prophylactic anticoagulant", "Compression Stockings", and "Angioplasty, Balloon" for pulmonary embolism. (2) Many false positive concepts were biomarkers of tests and assessments for treatment monitoring, usually with the semantic type of "amino acid, peptide, or protein". Examples include "neurohormonal factor", "N-terminal pro-B-type natriuretic peptide". (3) The false positive concepts could be studied as adverse events or risk factors for the disease of interest. Especially for pulmonary embolism, many false positive concepts were related to complications of certain procedures or medications that increase the risk of pulmonary embolism, such as "Arthroplasty", "Repair of hip", "Splenectomy". (4) False positive concepts were also caused by errors introduced by NLP tools. For example, from the sentence "this indicates that the MHAQ and RA-HAQ generally fail to identify appropriately the extent of functional loss in RA", the predication {Ametantrone TREATS Rheumatoid Arthritis} was extracted, where "HAQ" (Health Assessment Questionnaire) was incorrectly mapped to "ametantrone".

5. Discussion

In this study, we developed a pipeline-based knowledge extraction system to automatically generate disease-specific treatment vocabularies from the biomedical literature. The system is designed to retrieve disease-specific treatment-related articles, predications, and a ranked list of concepts. Comparing to a MeSH-based and a predication-based concept extraction approaches, our system had significantly higher precision for extracting the top 100 concepts. We also compared different algorithms ranking the extracted concepts; there was no significant difference among four ranks. Our system achieved an average precision of 0.8 for the top 100 concepts. We conclude that this pipeline-based system could be useful in generating disease-specific treatment vocabulary from the biomedical literature for building disease-specific ontologies. Besides, manual review of the system output would be necessary in order to generate a high-quality treatment vocabulary from these automated generated concepts. As an individual without much clinical background, we estimated the time for judging the relevance of the treatment concepts to the disease of interest by reading the origin sentences and citations, which is about one minute per concept. Comparing to manually acquisition, this could be much more efficient.

We reported that the pipeline system has achieved an average precision of 0.80 ranked by *interest* based on five test diseases. However, as the results show, for well-studied diseases (*e.g.*, rheumatoid arthritis) with many associated biomedical articles, the system would have higher precision, while for those with less articles (*e.g.*, pulmonary embolism), their precision is relatively lower. Therefore, the reported performance would not reflect the system's performance on diseases that have not been extensively investigated, such as new or rare diseases.

Our system has achieved a relatively low recall based on two test diseases (*i.e.*, pulmonary embolism and rheumatoid arthritis). Based on the error analysis in Section 4.4, approximately two thirds of the false negative concepts were probably attributed to the relation extraction tool we have used. However, there exist many other approaches aimed at extracting semantic relations from the biomedical literature or web documents, and some of them were also used UMLS and/or MetaMap [43]. Therefore, our system may gain further recall by incorporating the output of other relation extraction approaches or tools as secondary knowledge sources in addition to the SemMedDB to our proposed pipeline process.

Although the automated generated vocabulary was not able to identify 100% of the concepts in our manually generated reference vocabularies, the automated approach was able to extract some relevant treatment concepts that were missing in these reference vocabulary. This included cases of concepts with finer granularity or new information that was not included in the guidelines, textbooks, or online documents used to build the reference vocabularies. What's more, rather than starting from scratch, we build our system upon publically available resources, such as PubMed Clinical Queries, MEDLINE citations, and SemMedDB. In addition, we developed semantic schemas for treatment from an existing disease-specific treatment vocabulary to filter treatment predications rather simply relying on predicates such as "TREATS" or "PREVENTS". In this way, more information could be captured, for example, the evidence about the comparison between two medications can also be identified.

The main contribution of our study lies in three areas; the tuned selection of articles, the filtering of predications from millions of predications in the SemMedDB, and the ranking of concepts specific to the disease of interest. As Fig. 7 shows, predication-based approach has lower precision comparing to the pipeline system, which indicates that purely using SemRep predications would require much more review effort. In addition, the MeSH-based approach have lower recall comparing to the pipeline system, which indicates that using MeSH heading in the MEDLINE citations would not result as good coverage of the treatment vocabulary as using the pipeline system.

Our approach is innovative in two ways. First, compared to previous studies [18,22], we not only retrieve disease-specific pharmaceutical substances, but also other types of treatment, such as procedures, devices, and activities. In terms of disease-drug pairs, it is interesting to compare the results with previous studies [18,22]. However, we found such comparisons to be difficult since there were substantial differences in study goals, evaluation methods, and reference standards. In a simple comparison to the work of Chen *et al.* [18], our study found a greater number of disease-relevant citations and disease-drug pairs. Comparing to Xu's work [22], we have achieved a similar recall at a precision of 0.80, with the caveat that the reference standards used in both studies were different. Second, we were able to collect the source sentences and PubMed citations related to the disease-specific treatments. This could be useful for anyone who are interested in expanding their knowledge on a specific treatment. The extracted concepts also provide an index for over thousands of disease-specific treatment-related citations and sentences from MEDLINE. Researchers or clinicians can use this index to trace the evidence in the biomedical literature of a specific treatment for the disease of interest.

Our proposed approach was designed to be generalizable to other disease domains, such as diagnostic tests, signs, and symptoms. Yet, some adaptation is necessary including developing specific semantic schemas and defining common concepts for other disease domains. The same approach used to develop the semantic schema and define common concepts in the present study can be followed to adapt the algorithms to other disease domains.

The study has several limitations. First, the semantic schema for extracting treatment predications and concepts were developed based on a reference vocabulary of one disease (*i.e.*, heart failure), and might not be generalize to some types of disease. Second, we defined a list of common concepts to be filtered from extracted treatment concepts in Section 3.1.4. The selection of common concepts is based on an arbitrary cut-off threshold. Third, as the algorithm evaluation demonstrated, our reference standards had gaps in coverage and therefore were not perfect. Last the approach to judging the correctness of extracted concepts for diseases without a reference vocabulary was not as rigorous as the approach used to develop the reference vocabularies.

6. Conclusions

We investigated a pipeline-based approach to extract disease-specific treatment concepts from the biomedical literature to assist the development of disease-specific vocabularies. The pipeline-based approach obtained a mean precision of 0.8 for the top 100 retrieved concepts, which was significantly higher than two baselines. The performance of four ranking strategies (*e.g.*, *occurrence*, *degree centrality*, *weighted degree centrality*, and *interest*) was not statistically significant different. In the future, we intend to extend the system to extract concepts on other disease aspects, including signs, symptoms, and diagnostic tests.

Acknowledgements

The authors thank Thomas Rindflesch and Marcelo Fiszman for providing access to Semantic MEDLINE and useful input related to the database. The authors also thank Olivier Bodenreider for inputs on using the UMLS and MeSH. This work was supported in part by Grants LM010482 and 1R01LM011416 from the National Library of Medicine.

References

[1] L. Wang, B.E. Bray, J. Shi, G. Del Fiol, P.J. Haug, A method for the development of disease-specific reference standards vocabularies from textual biomedical literature resources, Artif. Intell. Med. (2016), http://dx.doi.org/10.1016/j.artmed.2016.02.003.

[2] P.N. Gorman, Information needs of physicians, J. Am. Soc. Inform. Sci. 46 (1995) 729–736, http://dx.doi.org/10.1002/(SICI)1097-4571(199512)46:10<729::AID-ASI3>3.0.CO;2-2.

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

[4] G. Del Fiol, T.E. Workman, P.N. Gorman, Clinical questions raised by clinicians at the point of care, JAMA Int. Med. 174 (2014) 710, http://dx.doi.org/10.1001/jamainternmed.2014.368.

[5] J.W. Ely, J.A. Osheroff, P.N. Gorman, M.H. Ebell, M.L. Chambliss, E.A. Pifer, A taxonomy of generic clinical questions: classification study, BMJ 321 (2000) 429–432.

[6] R.J. Cline, K.M. Haynes, Consumer health information seeking on the Internet: the state of the art, Health Educ. Res. 16 (2001) 671–692, http://dx.doi.org/10.1093/her/16.6.671.

[7] S. Fox, D. Fallows, Internet Health Resources, PewResearchCenter, 2003. <http://www.pewinternet.org/2003/07/16/internet-health-resources/>.

[8] N.F. Noy, D.L. Mcguinness, Ontology development 101: a guide to creating your first ontology, Stanford Knowl Syst Lab Tech Rep KSL-01-05 Stanford Med Informatics Tech Rep SMI-2001-0880, 2001.

[9] O. Bodenreider, Biomedical ontologies in action: role in knowledge management, data integration and decision support, IMIA Yearb Med. Inform. 47 (2008) 67–79.

[10] G. Eysenbach, A.R. Jadad, Evidence-based patient choice and consumer health informatics in the internet age, J. Med. Internet Res. 3 (2001) e19.

[11] Q.T. Zeng, J. Crowell, R.M. Plovnick, E. Kim, L. Ngo, E. Dibble, Assisting consumer health information retrieval with query recommendations, J. Am. Med. Infor. Assoc. 13 (2006) 80–90, http://dx.doi.org/10.1197/jamia.M1820.

[12] G. Fu, C.B. Jones, A.I. Abdelmoty, Ontology based spatial query expansion in information retrieval, Lect. Notes Comput. Sci. – ODBASE2005 3761 (2005) 11466–11482.

[13] H.J. Lowe, G.O. Barnett, Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches, JAMA 271 (1994) 1103–1108, http://dx.doi.org/10.1001/jama.271.14.1103.

[14] P.J. Haug, J.P. Ferraro, J. Holmen, X. Wu, K. Mynam, M. Ebert, et al., An ontology-driven, diagnostic modeling system, J. Am. Med. Inform. Assoc. 20 (2013) e102–e110, http://dx.doi.org/10.1136/amiajnl-2012-001376.

[15] A. Malhotra, E. Younesi, M. Gündel, B. Müller, M.T. Heneka, M. Hofmann-Apitius, ADO: a disease ontology representing the domain knowledge specific to Alzheimer's disease, Alzheimer's Dement 10 (2014) 238–246, http://dx.doi.org/10.1016/j.jalz.2013.02.009.

[16] N. Chalortham, M. Buranarach, T. Supnithi, Ontology development for type II diabetes mellitus clinical support system, in: Proc 4th Int. Conf. Knowl. Inform. Creat Support Syst., 2009.

[17] P. Buitelaar, P. Cimiano, B. Magnini, Ontology Learning From Text: Methods, Evaluation and Applications, IOS Press, Amsterdam, The Netherlands, 2005. doi: 10.1.1.70.3041.

[18] E.S. Chen, G. Hripcsak, H. Xu, M. Markatou, C. Friedman, Automated acquisition of disease-drug knowledge from biomedical and clinical documents: an initial study, J. Am. Med. Infor. Assoc. 15 (2008) 87–98, http://dx.doi.org/10.1197/jamia.M2401.

[19] X. Wang, A. Chused, N. Elhadad, C. Friedman, M. Markatou, Automated knowledge acquisition from clinical narrative reports, AMIA Annu. Sympos. Proc. (2008) 783–787.

[20] A. Wright, E.S. Chen, F.L. Maloney, An automated technique for identifying associations between medications, laboratory results and problems, J. Biomed. Inform. 43 (2010) 891–901, http://dx.doi.org/10.1016/j.jbi.2010.09.009.

[21] R. Xu, L. Li, Q. Wang, DRiskKB: a large-scale disease-disease risk relationship knowledge base constructed from biomedical text, BMC Bioinform. 15 (2014) 105, http://dx.doi.org/10.1186/1471-2105-15-105.

[22] R. Xu, Q. Wang, Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing, BMC Bioinform. 14 (2013) 181, http://dx.doi.org/10.1186/1471-2105-14-181.

[23] J.J. Cimino, G.O. Barnett, Automatic knowledge acquisition from MEDLINE, Methods Inf. Med. 32 (1993) 120–130.

[24] Q. Zeng, J.J. Cimino, Automated knowledge extraction from the UMLS, AMIA Annu. Sympos. Proc. (1998) 568–572.

[25] T.C. Rindflesch, M. Fiszman, The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text, J. Biomed. Inform. 36 (2003) 462–477, http://dx.doi.org/10.1016/j.jbi.2003.11.003.

[26] T.C. Rindflesch, M. Fiszman, B. Libbus, Semantic interpretation for the biomedical research literature, in: S. Fuller, W. Hersh, C. Friedman, H. Chen (Eds.), Med. Infor. Knowl. Manage. Data Min. Biomed., Springer, 2005, pp. 399–422.

[27] T.C. Rindflesch, L. Tanabe, J.N. Weinstein, L. Hunter, EDGAR: extraction of drugs, genes and relations from the biomedical literature, in: Pac. Sympos. Biocomput., 2000, pp. 517–528, http://dx.doi.org/10.1016/j.bbi.2008.05.010.

[28] I. Hendrickx, S.N. Kim, Z. Kozareva, P Nakov, L. Romano, S. Szpakowicz. SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals. 5th Int. Work. Semant. Eval. ACL 2010, 2010, p. 33–8.

[29] B. Rosario, M.a. Hearst, Classifying semantic relations in bioscience texts, in: Proc. 42nd Annu. Meet. Assoc. Comput. Linguist, 2004, p. 430, http://dx.doi.org/10.3115/1218955.1219010.

[30] D. Zeng, K. Liu, S. Lai, G. Zhou, J. Zhao, Relation classification via convolutional deep neural network, COLING (2014) 2335–2344.

[31] M. Bundschus, M. Dejori, M. Stetter, V. Tresp, H.-P. Kriegel, Extraction of semantic biomedical relations from text using conditional random fields, BMC Bioinfor. 9 (2008) 207, http://dx.doi.org/10.1186/1471-2105-9-207.

[32] C. Giuliano, A. Lavelli, L. Romano, V. Sommarive, Exploiting shallow linguistic information for relation extraction from biomedical literature, EACL (2006).

[33] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, T.C. Rindflesch, SemMedDB: a PubMed-scale repository of biomedical semantic predications, Bioinformatics 28 (2012) 3158–3160, http://dx.doi.org/10.1093/bioinformatics/bts591.

[34] R.B. Haynes, N.L. Wilczynski, K.A. McKibbon, J.C. Walker, C.J. Sinclair, Developing optimal search strategies for detecting clinically sound studies in MEDLINE, J. Am. Med. Infor. Assoc. 1 (1994) 447–458.

[35] R.B. Haynes, N.L. Wilczynski, Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey, BMJ 328 (2004) 1040, http://dx.doi.org/10.1136/bmj.38068.557998.EE.

[36] 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 (2005) 1179, http://dx.doi.org/10.1136/bmj.38446.498542.8F.

[37] H. Zhang, M. Fiszman, D. Shin, C.M. Miller, G. Rosemblat, T.C. Rindflesch, Degree centrality for semantic abstraction summarization of therapeutic studies, J. Biomed. Inform. (2011) 830–838, http://dx.doi.org/10.1016/j.jbi.2011.05.00 (in press).

[38] M. Fiszman, T.C. Rindflesch, H. Kilicoglu, Abstraction summarization for managing the biomedical research literature, in: Proc. HLT-NAACL Work. Comput. Lex. Semant., 2003, pp. 76–83.

[39] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Inform. Process. Manage. 24 (1988) 513–523.

[40] A. Ozgür, T. Vu, G. Erkan, D.R. Radev, Identifying gene-disease associations using centrality on a literature mined gene-interaction network, Bioinformatics 24 (2008) i277–i285, http://dx.doi.org/10.1093/bioinformatics/btn182.

[41] T. Opsahl, F. Agneessens, J. Skvoretz, Node centrality in weighted networks: generalizing degree and shortest paths, Soc. Netw. 32 (2010) 245–251, http://dx.doi.org/10.1016/j.socnet.2010.03.006.

[42] https://www.nlm.nih.gov/mesh/topsubscope.html.

[43] V. Nebot, R. Berlanga, Exploiting semantic annotations for open information extraction: an experience in the biomedical domain, Knowl. Inf. Syst. 38 (2014) 365–389, http://dx.doi.org/10.1007/s10115-012-0590-x.

CHAPTER 5

USING CLASSIFICATION MODELS FOR THE GENERATION OF DISEASE-SPECIFIC MEDICATIONS FROM BIOMEDICAL LITERATURE AND CLINICAL DATA REPOSITORY

Reprinted with permission from Wang L, Haug PJ, Del Fiol G. Using classification models for the generation of disease-specific medications from biomedical literature and clinical data repository. Journal of Biomedical Informatics. 2017;69:259-266.

5.1 Abstract

Mining disease-specific associations from existing knowledge resources can be useful for building disease-specific ontologies and supporting knowledge-based applications. Many association mining techniques have been exploited. However, the challenge remains when those extracted associations contained much noise. It is unreliable to determine the relevance of the association by simply setting up arbitrary cut-off points on multiple scores of relevance; and it would be expensive to ask human experts to manually review a large number of associations. We propose that machine-learning-based classification can be used to separate the signal from the noise, and to provide a feasible approach to create and maintain disease-specific vocabularies.

We initially focused on disease-medication associations for the purpose of

simplicity. For a disease of interest, we extracted potentially treatment-related drug concepts from biomedical literature citations and from a local clinical data repository. Each concept was associated with multiple measures of relevance (i.e., features) such as frequency of occurrence. For the machine purpose of learning, we formed nine datasets for three diseases with each disease having two single-source datasets and one from the combination of previous two datasets. All the datasets were labeled using existing reference standards. Thereafter, we conducted two experiments: 1) to test if adding features from the clinical data repository would improve the performance of classification achieved using features from the biomedical literature only, and 2) to determine if classifier(s) trained with known medication-disease data sets would be generalizable to new disease(s).

Simple logistic regression and LogitBoost were two classifiers identified as the preferred models separately for the biomedical-literature datasets and combined datasets. The performance of the classification using combined features provided significant improvement beyond that using biomedical-literature features alone ($p\text{-value}<0.001$). The performance of the classifier built from known diseases to predict associated concepts for new diseases showed no significant difference from the performance of the classifier built and tested using the new disease's dataset.

It is feasible to use classification approaches to automatically predict the relevance of a concept to a disease of interest. It is useful to combine features from disparate sources for the task of classification. Classifiers built from known diseases were generalizable to new diseases.

5.2 Introduction

The biomedical literature and electronic medical records offer great opportunities for acquiring disease-specific medical knowledge. Automated extraction of disease-*medication* associations from these knowledge sources can speed the process of building disease-specific concept vocabularies which could be further used for various clinical applications, such as automated annotation of biomedical text [1,2], identification of diseased cohorts [3], and development of diagnostic models [4]. In the present study, we propose an approach for automated extraction of disease-*concept* associations from the biomedical literature and a clinical data repository (CDR). The approach uses machine learning classification models to predict the relevance of concepts to the disease of interest. The approach is developed based on former studies [5–8] and it overcomes a common challenge faced in these studies, which is to use the metrics of relevance of the disease-*concept* associations to effectively decrease the manual efforts necessary to review noisy collections of associations in order to build disease-specific concept vocabularies. To build classification models, we evaluated the proposition that combining features derived from a clinical data repository with those from the biomedical literature would result in better performance than using features from a single source. We also conducted an exploratory assessment of the model's generalizability in predicting the disease-*concept* associations extracted for other diseases.

5.3 Background and Significance

Dozens of studies have investigated techniques for extracting disease-*concept* associations from the biomedical literature and electronic medical records. The *concepts* studied have included associated genes [9], signs and symptoms [10], findings [11],

medications [7,8], and lab tests [7]. Numerous knowledge acquisition techniques have been proposed to extract relational information, including co-occurrence-based statistics [7,8,11], natural language processing (NLP) [12,13], graph theory [9,14], and others [15,16]. Zeng and Cimino retrieved disease-chemical relationships from the UMLS co-occurrence table (MRCOC) simply based on the co-occurrence of MeSH terms assigned to published articles[17]. Cao *et al.* used NLP and co-occurrence statistics (i.e., chi-square statistics and the proportion confidence interval) to extract disease-finding associations [11]. Chen *et al.* applied similar techniques to extract disease-drug pairs from PubMed® citations and clinical documents [8]. In those studies, NLP techniques have been used mainly for named entity recognition when the sources of the data were in "free-text" form. In addition, Rindflesch *et al.* developed a rule-based system called SemRep that extracts the semantic relations between the concepts identified in a particular sentence in the biomedical literature [12,18]. For example, given the sentence "a randomized trial of etanercept as monotherapy for psoriasis", a semantic predication was generated: *etanercept TREATS psoriasis*. Bundschus *et al.* explored using conditional random fields to identify the semantic relations between disease and medications and between disease and genes in biomedical text [15]. Xu and Wang used a pattern-learning approach to extract disease-drug and disease-disease risk pairs from biomedical abstracts [16,19]. In addition, the authors of the present study have developed a pipeline-based system which combines multiple techniques (i.e., document retrieval, SemRep, UMLS semantic network, and co-occurrence-based statistics) to extract disease-specific treatments (including medications, surgical procedures, medical devices, and activities) from biomedical titles and abstracts [6]. More details about this work can be

found in section 3.1.

Existing statistically-based automated extraction techniques score the disease-*concept* candidate set allowing some reduction in noise, but leaving behind a large number of "bad" concept-disease pairs. The precision can be very low when focusing on high recall. For example, in a previous study, when counting all retrieved treatment concepts, we achieved a precision of less than 0.3 on two test diseases when comparing to manually-created reference vocabularies [6]. The challenge escalates when facing hundreds or thousands of concepts extracted for each disease in light of low precision. Ultimately, filtering out false-positives requires manual expert review, which is costly and time-intensive.

Disease-*concept* associations extracted by automated techniques have been assigned statistical scores, such as frequency of occurrence, which may provide some sort of indication for the strength of the relationship between the disease of interest and extracted *concepts*. Researchers previously investigated potential approaches to set proper thresholds based upon those statistical scores to identify a subset of important associations for further investigation. For example, Cao *et al.* explored using the volume test of Diaconis and Efron to identify thresholds using the chi-square score [20]. However, choosing cut-off points on these statistical scores is either empirical or arbitrary, and it would not generally apply well to a situation where extracted concepts are assigned multiple scores.

To determine the relevance of extracted concepts to the disease of interest is a binary classification issue. To address the above challenge, machine-learning-based classification techniques can possibly be used to predict the relevance of extracted

disease-*concept* associations based upon the multiple statistical scores. This would eliminate a significant number of irrelevant concepts and keep a subset of "interesting" concepts for further investigation.

To develop an appropriate classification model, we considered two important questions: (1) what features should be used to build the model; and (2) how generalizable is the model?

Disease-specific associations could be extracted from different sources by multiple techniques, which generate different kinds of measures of relevance (i.e., features). For example, in a prior study, we used four scoring strategies (i.e., frequency of occurrence, interest, degree centrality, and weighted degree centrality) to extract disease-treatment associations from the biomedical literature [6]. Wright *et al.* applied five co-occurrence-based statistics (i.e., support, confidence, chi square, interest, and conviction) to extract disease-medication and disease-lab test associations from the electronic medical records [7]. Studies have shown that combining the results of extraction by different techniques/queries from a single source led to progressively improving retrieval performance [21–23]. Other studies also show that the results of extraction from the different sources are somewhat complementary [5,8]. With these findings in mind, we assumed that by combining the measures of relevance generated by different techniques from different sources (i.e., the biomedical literature and a CDR) as features within a classification system, the performance of the classifiers may be improved compared to using a single feature or features only from a single source.

The generalizability of the classification model is important because it is difficult and expensive to build a classifier for each disease. However for different diseases, the

range and distribution of the value of the relevance measures may be different. This could affect the performance of a classifier when trained and tested on different disease datasets. We measure the generalizability of the classifier by determining if a classifier trained and tested on different disease's datasets achieved as good performance as the classifier trained and tested on the same disease's dataset.

The ultimate goal of this study is to develop machine learning classifiers that could reduce the manual effort necessary to review noisy collections of disease-specific concepts. To achieve this goal, in the present study, we initially focused on disease-medication associations, and searched for classification models appropriate to predict the relevance of groups of medications to a specific disease. The models were designed to incorporate multiple statistical scores. We assessed two research questions: (1) Would adding the features from the CDR improve the performance of models that used features from biomedical literature only; (2) Would models built from known disease-medication associations be effective in predicting disease-medication associations for new diseases?

5.4 Materials and Methods

The study methods consisted of the following steps (see Figure 5.1): (1) extraction of disease-specific medications from the biomedical literature; (2) extraction of disease-specific medications from a local CDR; (3) preparation of datasets for classification, including merging the datasets from the disparate sources and validating disease-medication associations using reference standards; (4) searching for preferred classifiers for different datasets; and (5) statistical analysis. The reference standards in Figure 5.1 are the reference vocabularies we built in a prior study for three diseases (i.e., heart failure, pulmonary embolism, and rheumatoid arthritis) [5].

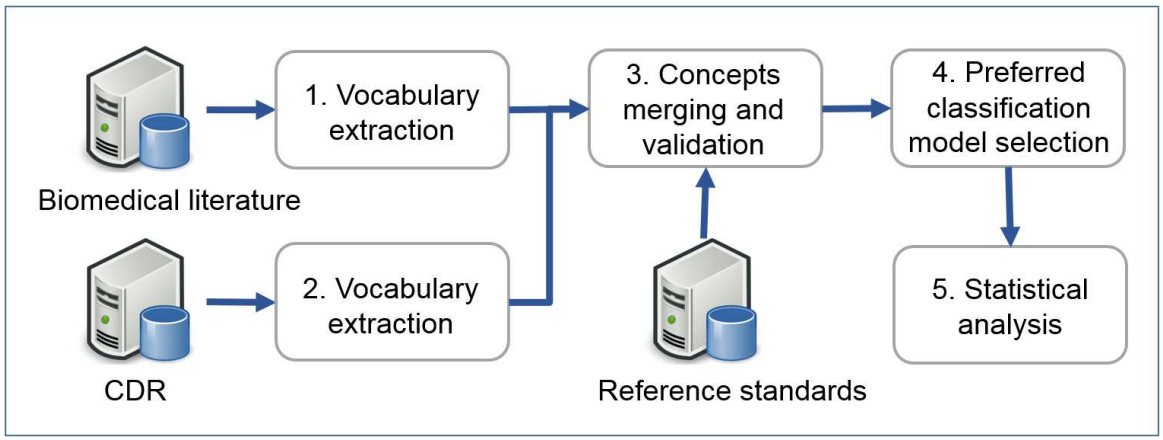

Figure 5.1. Workflow for testing supervised learning of classification models to generate disease-specific reference vocabularies from the biomedical literature and the CDR.

5.4.1 Association Extraction from the Biomedical Literature

In an earlier study, we developed a pipeline-based system to extract disease-specific treatment concepts from MEDLINE citations [6]. That pipeline system consists of several sequential steps, including citation retrieval, predication extraction, treatment predication and concept extraction based on a semantic schema, and relevance ranking. In the citation retrieval step, we developed our PubMed queries based on PubMed Clinical Queries to retrieve disease-pertinent, treatment-related citations from the entire MEDLINE database. The sentences in those retrieved citations are then parsed into predications, which are semantic relations in the form of a triple *subject predicate object* (e.g., *Chronic Obstructive Airway Disease AFFECTS Left Ventricular Function*). Treatment predications are semantic relations that match a predefined semantic treatment schema, which is a set of rules that define which predications are valid treatment predications. For example, the predication *Captopril TREATS Congestive heart failure* matches the semantic schema *Pharmaceutical Substance TREATS Disease or Disorder*.

In the next step, the treatment concepts (e.g., Captopril) are extracted from the retrieved treatment predications.

In the last step, the extracted concepts are assigned four different scores: (1) *occurrence*, which is the frequency of the concept in the treatment predications; (2) *interest*, which is the ratio of the *occurrence* of a treatment concept to the sum of the *occurrences* of all treatment concepts retrieved for a given disease of interest divided by the ratio of the occurrence of a treatment of interest to all treatment concepts in the database; (3) *degree centrality*, which measures the significance of the concepts in the graph by counting their connectivity to other concepts; and (4) *weighted degree centrality*, which is a harmonization between the frequency of occurrence and the degree centrality. These four scores, when used independently, were not significantly different in terms of the mean precision of the top 100 ranked concepts.

The treatment concepts returned from the pipeline-based system broadly covered the semantic groups of *procedures*, *chemicals & drugs*, *activities & behaviors*, and *devices*. In the present study, as we focused on medications, we further limited the returned concepts to those whose source terminology was "RXNORM" and whose term type was "IN" (ingredients). The four relevance scores described above were generated for each concept.

5.4.2 Association Extraction from CDR

For mining disease-medication associations from a local CDR, we adapted the approach proposed in [7] which resulted in 89.2% accuracy for the top 500 disease-drug associations. The approach first uses frequent item set mining to locate commonly co-occurring items in a database, and then uses association rule mining to identify the

direction of the associations. Thereafter, co-occurrence-based statistics are applied to sort the disease-lab test and disease-medication associations from electronic health records.

For mining frequent item sets, the Apriori algorithm was used [24]. Apriori was designed to operate on databases of transactions. For example, consider a transaction in the grocery store as a set of items purchased by the customer during one visit to the store; a large collection of these transactions could be used to identify associations between purchased items. In the clinical setting, a transaction would be a set of diagnoses, medications, procedures, and lab tests associated with a patient in a clinical encounter. From a database with millions of transactions, we simply count the number of transactions in which the disease of interest and the medications co-occurred. With each two-item set, a two-by-two table can be constructed (Table 5.1), where *a* is the number of transactions in which X and Y co-occurred, *b* is the number of transactions that contain X but not Y, *c* is the number of transactions that contains Y but not X, and *d* is the number of transactions that contain neither X nor Y.

From Table 5.1, the following statistical measures were calculated:

1. Support (X, Y) is simply the number of transactions in which item X and Y co-occur. It is based on the underlying assumption that two associated concepts are more likely to appear together than pairs of unrelated concepts.

$$\text{Support}(X, Y) = a \tag{1}$$

2. Confidence (X, Y) is the proportion of all transactions that contain X that also contain Y. It provides a way to account for the directionality of associations. Take an example in [7], in *confidence (insulin, diabetes)*, the proportion of patients who have been prescribed insulin and have diabetes would be

Table 5.1
Two-by-two contingency table for the frequent items X and Y. X is the disease of interest and Y is the medication co-occurred with X.

	Y	Y'
X	a	b
X'	c	d

different from *confidence (diabetes, insulin)*, which is the proportion of patients with diabetes who have been prescribed insulin.

$$\text{Confidence} = a/(a+b) \tag{2}$$

3. Interest (X, Y) is the proportion of confidence (X→Y) divided by the proportion of all transactions that contain Y. It accounts for the weakness of confidence when Y is highly occurrent in the database.

$$\text{Interest} = [a/(a+b)] / [(a+b)/(a+b+c+d)] = a*(a+b+c+d)/(a+b)^2 \tag{3}$$

4. Chi-square (X, Y) is a statistical test that measures the significance of the association between X and Y.

$$\text{Chisq} = (a*d-b*c)^2 * (a+b+c+d)/[(a+b)*(c+d)*(b+d)*(a+c)] \tag{4}$$

We used clinical data from Intermountain Healthcare, a regional US healthcare provider which consists of 22 hospitals and over 150 clinics. The patient data collected from these hospitals and clinics is stored in an enterprise data warehouse (EDW). Intermountain also maintains a database called the analytic health repository (AHR), which is a subset of the EDW that contains commonly accessed classes of medical data (e.g., patient, diagnosis, prescription, lab tests, procedures) expressed using standard medical terminologies such as ICD-9, LOINC, and SNOMED CT [4]. In the present

study, we used diagnoses and medication data at the encounter level from the AHR and restricted the query timeframe to between 01/01/2008 and 12/31/2010.

Our approach differed slightly from the one proposed in [7]. We built a database table of transactions that combined the information of medications and diagnosed problems at the encounter level, while Wright *et al*. created the transactions at the patient level. We assumed that the drugs prescribed to a patient were more specific to the problems diagnosed and managed in that particular encounter. In the transactions table, each transaction corresponded to a patient encounter, and included all prescribed medications and diagnostic codes that happened as a part of that patient encounter. The AHR drug prescription table does not have any encounter information, but we linked the medication prescriptions to a specific encounter by checking whether the prescription time falls within the specific time frame of an encounter for the targeted patient.

In order to extract treatment concepts that match the concepts extracted from the biomedical literature, we converted the original codes of medication to UMLS CUIs. The prescribed drugs were coded in First Data Bank codes which were also mapped to the RxNorm codes. Besides, the drugs were recorded at the clinical drug level (SCD) {ingredient+strength+dose form} (e.g., Lisinopril 5 mg oral tablet) of RxNorm. We convert the concepts to the ingredient level through the "*ingredient_of*" RxNorm relationship. For example, 'Lisinopril' is the *ingredient_of* 'Lisinopril 5 mg oral tablet'. Multi-ingredient drugs were decomposed into their individual ingredients. All the ingredients were mapped to UMLS concepts through querying the UMLS MRCONSO table. The diagnoses were coded using ICD-9-CM codes; we kept the original codes in the transactions table. In final, we built a table containing over 10 million transactions to

support frequent item sets mining and calculate statistical scores for extracted disease-medication associations.

We focused on extracting medications associated with a specific disease of interest. However, sometimes children concepts (e.g., ICD9 codes 428.0, 428.1, 428.2) as opposed to the exact disease of interest (e.g., heart failure) are present in the database of transactions. To address these cases, we expanded the disease of interest to include its children concepts. As a result, the cell *a* in Table 5.1 was calculated as the sum of the number of transactions containing X and Y or the children concepts of X and Y.

5.4.3 Dataset Preparation

Supervised learning requires a labeled dataset from which to build classifiers. In a previous study, we manually created reference vocabularies for heart failure (https://bioportal.bioontology.org/ontologies/HFO), rheumatoid arthritis (https://bioportal.bioontology.org/ontologies/RAO), and pulmonary embolism (https://bioportal.bioontology.org/ontologies/PE) [5]. We chose these three diseases for our study. All three reference vocabularies contain a near-saturated set of disease-associated treatments (e.g., medications, surgical therapy), where the near saturation is defined as finding <5% new concepts with the introduction of a new knowledge source (e.g., textbook) [5]. In the present study, since we focused on disease-medication associations, we formed a subset of each vocabulary containing only medication treatments (see supplements for an example of heart failure-related medication concepts). The majority of concepts were represented as UMLS concept unique identifiers (CUIs); concepts that could not be mapped to a UMLS CUIs were excluded from the study. These reference vocabulary subsets were used to create labeled datasets.

For the dataset features, we first extracted two lists of disease-specific medications separately from the biomedical literature and the CDR through the approaches introduced in sections 5.4.1 and 5.4.2. Second, we created another list of medications for each disease by merging the medications extracted from two sources, while keeping their original features: four features from the biomedical literature (i.e., occurrence, interest, degree centrality, weighted degree centrality) and four features from the CDR (i.e., support, confidence, chi-square, and interest). If a concept only had feature values from one source, the values of the features from the other source were marked as missing values. Then, we labeled each concept by comparing the concept to the the target subset of reference vocabulary with exact mapping. A concept that was in a disease's reference standards was labeled as reference positive ("RefPos"), and a concept that was not found in the disease's reference standards was labeled as reference negative ("RefNeg"). More details about these datasets can be found in section 5.5.1. Thereafter, the data was organized into a standard format, called Weka attribute-relation file format (ARFF) (http://weka.wikispaces.com/arff) and fed to Weka for further analysis.

In total, we created nine datasets based upon these three diseases where each disease had three datasets: one from the biomedical literature, one from a local CDR, and the third one from the combination of these two datasets.

5.4.4 Searching for Preferred Classification Models

Machine learning environments provide a variety of classification algorithms that can be used to build predictive models for disease-medicine associations. In this study, we have three kinds of datasets. Each dataset has slightly different feature sets, which may favor different models. We used the three heart failure datasets to search for

effective classification models. We used the rheumatoid arthritis and pulmonary embolism datasets for testing the chosen classification models.

To identify useful classifiers, we applied Weka, a general purpose, open-source, data mining toolkit, which includes over 50 classifiers in version 3.7, as well as a variety of data transformation and feature selection algorithms [25]. In this study, we did not explore all the classifiers available; instead, we focused on a subset of commonly-used classifiers.

First of all, we included three ensemble-based classifiers: adaptive boosting M1 (ADB) [26], LogitBoost (LGB) [27]), and bagging (BAG) [28]. Ensemble methods are learning algorithms that construct a set of classifiers (such as neural networks or decision trees) then classify new data points by taking a weighted vote of their predictions [29]. Previous research has shown that an ensemble is often more accurate than any of the single classifiers in the ensemble [29]. Ensemble approaches generally refer to two kinds of learning techniques (i.e., boosting and bagging). Two of the three classifiers we chose use the boosting approach and one uses bagging. Each ensemble-based classifier requires a specific base classifier. In our study, we chose the classic "decision stump" classifier (a machine learning model consisting of a one-level decision tree) for two boosting approaches, and a decision tree classifier for the bagging approach.

Second, we included a typical lazy and memory-based learning approach called locally weighted learning (LWL). LWL was introduced by Atkeson *et al.* in 1997 and is based on a locally weighted linear regression [30]. The "decision stump" classifier was used as the base classifier and the "brute force" search algorithm was used for nearest neighbor search. In addition, four single classifiers were chosen, including: Bayes

Network (BYN) [31], multilayer perceptron (MLP), simple logistic regression (SLR) [32], and random forest (RDF) [33]. This approach provided experience with a variety of classification algorithms. Weka's default parameters were used for all of those classifiers.

To compare among the classification models, we used the area under the receiver operating characteristic (ROC) curve (AUC) as a single measure of a classifier's performance for evaluating which model is better on average. AUC is a general measure of effectiveness often preferred over other measures (e.g., accuracy) in comparing classifiers [34,35]. We calculated the mean AUC for each model in three heart failure datasets using 10 repetitions of 10-fold cross-validation [36]. In addition, we calculated the 95% confidence interval for the mean AUC. Here, we report the top five classifiers based on their mean AUC with their 95% confidence intervals.

5.4.5 Statistical Analysis

After found the best classification model for each kind of dataset, we conducted exploratory experiments to assess our two research questions.

5.4.5.1 Comparison Between Using Single-Source and Combined Datasets

The first hypothesis is that adding the features and instances from the CDR would improve the performance of a classifier developed using the dataset from the biomedical literature only. The expectation is, no matter what classifier has been used, the best performance obtained using the combined dataset should outperform the best performance obtained using the single-source dataset. As indicated above, we tested if there is significant difference in the performance using the combined dataset versus the dataset from biomedical literature as measured by the AUC. In this experiment, we used

heart failure biomedical literature and combined datasets to select preferred classification models. The same datasets were used to train the classifiers. We used the remainder of the disease datasets (rheumatoid arthritis and pulmonary embolism) for testing the classifiers. The AUCs were generated separately for the biomedical literature dataset and combined dataset. To test the significance of the differences between two AUCs, we used the nonparametric DeLong test [37] which was implemented using an R package called pROC [38]. This is consistent with the recommendations made in [39,40].

5.4.5.2 Generalizability Assessment

The second experiment tests the hypothesis that a classifier built from known diseases' datasets will accurately predict the relevance of the medications extracted for other diseases. More specifically, how effective is a classifier trained using a labeled dataset from one or more diseases in accurately determining the relevance of disease-medication associations in a dataset for a new disease; how will this classifier comparing to the performance of a classifier which was both trained and tested on the new disease's dataset? Although the training datasets are different, if the classifier achieves similar performance, then we would infer that a classifier trained on diseases' labeled dataset would be generalizable to new disease(s).

In this study, we used only combined datasets for this experiment. We formed an experiment group with classifiers trained with each combination of two diseases' datasets and with their performance tested with the third disease. We also formed a control group with classifiers trained and tested with the dataset of the third disease using 10 repetitions of 10-fold cross-validation. We then compare the performance between the two groups.

We used a 95% confidence interval to assess the statistical significance of the difference of the mean AUCs.

5.5 Results

5.5.1 Datasets from the Biomedical Literature and Clinical Database

Table 5.2 reports the number of instances for each dataset. For example, for heart failure, we extracted 465 candidate medication concepts from the biomedical literature, 1144 from the CDR, and 1340 after merging them together.

Table 5.2
Summary description of the number of instances and features of the classification datasets created from the biomedical literature and clinical data repository.

Dataset (Disease, Source)	No. of instances	No. of RefPos
HF1 (HF, CDR)	1144	88
HF2 (HF, Biomedical literature)	465	100
HF3 (HF, Combined)	1340	107
RA1 (RA, CDR)	1011	62
RA2 (RA, Biomedical literature)	425	77
RA3 (RA, Combined)	1226	82
PE1 (PE, CDR)	930	18
PE2 (PE, Biomedical literature)	141	35
PE3 (PE, Combined)	998	36

HF: Heart failure; RA: Rheumatoid arthritis; PE: Pulmonary embolism

5.5.2 "Preferred" Model Selection

We calculated the mean AUC of all tested classification models from Weka separately for the heart failure datasets (i.e., HF1, HF2, and HF3) with 10 repetitions of 10-fold cross-validation. Table 5.3 shows the top 5 classifiers in terms of mean AUC as well as the 95% confidence interval. Based on the ranking in Table 5.3, we chose SLR as the preferred classification model for the biomedical literature datasets, MLP as the preferred classification model for the CDR datasets, and LGB as the preferred classification model for the combined datasets.

5.5.3 Comparison of Performance between Different Datasets

After the preferred models were selected, we trained the models with the heart failure datasets, and tested these trained models on the datasets from rheumatoid arthritis and pulmonary embolism. The AUC for SLR tested on the two biomedical literature

Table 5.3
Mean AUC of top 5 classifiers on two kinds of feature sets as well as 95% confidence interval.

Rank	Classifier: Mean AUC (95% Confidence Interval)		
	Biomed	**CDR**	**Combined**
1	SLR: 0.872 (0.870 – 0.874)	MLP: 0.795 (0.791 – 0.799)	LGB: 0.931 (0.927 – 0.935)
2	LWL: 0.872 (0.869 – 0.874)	BYN: 0.794 (0.789– 0.799)	LWL: 0.926 (0.924 – 0.927)
3	MLP: 0.871 (0.869 – 0.874)	BAG: 0.782 (0.776 – 0.788)	BYN: 0.918 (0.914 – 0.921)
4	LGR: 0.870 (0.867 – 0.873)	LGB: 0.781 (0.767 – 0.795)	ADB: 0.918 (0.912 – 0.923)
5	BAG: 0.863 (0.859 – 0.867)	SLR: 0.780 (0.774 – 0.786)	MLP: 0.899 (0.893 – 0.905)

datasets (RA2 and PE2) was 0.893, and the AUC of LGB on combined datasets (RA3 and PE3) was 0.947. Comparison of the AUC of the best detectors built from the biomedical literature dataset and the combined dataset yielded a significant difference among the two values (p-value 0.0077).

5.5.4 Generalizability Assessment

In Table 5.4, we reported the AUC of the LGB classifier which was trained with each combination of two disease datasets and tested on the third disease while comparing to the performance of the LGB classifier trained and tested on the same dataset of the third disease. From this table, we see that in all three cases, using a classifier trained with any two disease datasets to make a prediction on the third disease dataset can achieve excellent performance (AUC > 0.9). In addition, in two of three cases, the AUCs were above the upper bound of the 95% confidence interval of the internally trained group.

5.6 Discussion

In the present study, we tested using machine-learning classification models as a secondary filter to reduce the noise when extracting disease-specific medications from the biomedical literature and clinical data repository. Two research questions were answered: (1) would the performance of classification on extracted associations from the biomedical literature be improved by adding the features and drug instances extracted from the CDR; (2) would a classifier built from labeled datasets of some diseases be generalizable to new diseases. In this study, we choose SLR as the “preferred” classifier for the biomedical literature datasets and LGB as the “preferred” classifier for the datasets created from the features and instances of both the biomedical literature and

Table 5.4
The AUC and 95% confidence interval of seven classifiers with different combination of training and testing datasets.

Case		**Control**	
Two-disease datasets for training and one-disease datasets for testing	**AUC**	**Cross-validation within one disease dataset**	**Mean AUC (95% confidence interval)**
HF+PE→**RA**	0.933	**RA**	0.929 (0.925, 0.932)
HF+RA→**PE**	0.982	**PE**	0.979 (0.977, 0.981)
RA+PE→**HF**	0.932	**HF**	0.931 (0.927, 0.935)

clinical data repository. For the first research question, we found that the classification performance (i.e., AUC) on the biomedical literature datasets significantly improves from 0.893 to 0.947 after adding the features from the CDR. We did not test adding biomedical literature to the CDR, because the performance of CDR alone is worse than biomedical literature alone (see Table 5.3). In addition, it is easier to access the biomedical literature. There are additional challenges in trying to use clinical data. Therefore, it is important to know whether the CDR data adds value. For the second research question, we found that the classifiers built from the datasets of two diseases can be used effectively to predict the relevance of associations extracted in a third disease dataset; the performance may surpass that of a classifier trained with the dataset of the third disease itself. These findings support the conclusion that combining features from the CDR and biomedical literature significantly improves performance in terms of AUC compared with using features from those datasets alone. In addition, the classifiers built from one or two diseases generalized well to new diseases.

In Table 5.2, we also observed that the three biomedical literature datasets have a higher proportion of "refPos" instances comparing to the three corresponding CDR datasets. For example, for heart failure datasets, the HF2 has 21.51% (100 out of 465) of "refPos" instances while HF1 has 7.69% (88 out of 1144) "refPos" instances. We note that the CDR contributed a few "refPos" concepts that were missing from the biomedical literature. For example, for heart failure, 100 "refPos" were extracted from the biomedical literature, and 7 new "refPos" concepts were contributed from the CDR for a total of 107 "refPos" concepts in the combined datasets.

When choosing the "preferred" classifier for each kind of feature set, there were often no significant differences among the classifiers based on the mean AUC and 95% confidence interval. For the purpose of this study, we picked the ones that had a relatively higher mean AUC and smaller standard deviations. However, the key finding is that a variety of classification models provided similar results. In addressing a particular problem, other performance measures (e.g., precision, recall, f-measure) besides the AUC may contribute to the selection of an optimal classifier.

The main contribution of this study lies in three areas. First, we built classification models based on multiple numerical measures of relevance to filter irrelevant associations from the many associations extracted from the biomedical literature and CDR. If employed in a process for expert review, these classifiers may reduce the human effort spent in manual review of those extracted associations. The level of effort reduction would depend on the thresholds set by the users to focus on higher precision or recall for the classifiers. Second, we also tested and found that combining the features and instances from different sources would improve the overall performance of the

classification. This is particularly helpful when the performance of classification on the datasets from individual sources was low. Third, the classifiers built from a small subset of diseases can be generalizable to classify in other diseases. In this study, we have three diseases with labeled datasets; we can use these datasets to train a classifier that can effectively detect the relevant associations for other diseases.

There are also several limitations of this study. First, the accuracy of machine-learning-based classifiers is affected by a variety of factors including hyper-parameter settings, feature selection, and discretization. Altering those factors can be expected to affect the performance of the classifiers. In the present study, we chose to survey a broad range of classifiers in the Weka. However, rather than searching across a full range of hyper-parameters, feature selection algorithms, and discretization's, we chose to use Weka's default parameter settings. Second, in terms of the generalizability, from the present study, we found that classifiers built from one or two diseases can generalize well to new common diseases. However, it is not known if the performance will also generalize to less common diseases, with a small number of published articles or patient records. Third, when preparing the training set, we mapped reference concepts from previous studies to the extracted concepts. Therefore, we may have incorrectly labeled some concepts as false positives if they were not present in the reference standards. We anticipate a future analysis to explore a mechanism for identifying refinements to these algorithms that will yield the best classifications. Finally, the models appear generalizable to new diseases only when the datasets were from the same knowledge sources (the biomedical literature and a clinical data repository); it is uncertain whether the models will perform similarly when incorporating new knowledge sources.

Specifically, we only used the clinical data from one site, Intermountain Healthcare; the performance of the classifiers tested in this study may change when using the data from other sites or from multiple sites.

5.7 Conclusion

Machine-learning classification models can be used to identify relevant medications used for treating a disease of interest by taking advantage of the numerical scores generated from prior studies when extracting disease-associated concepts from the biomedical literature and clinical data repository. Combining the datasets generated from the biomedical literature and CDR improves the classification performance obtained with single-source datasets. Classifiers built from one or two diseases appear to generalize well to new diseases. In the future, we intend to integrate some of tested classification models into our pipeline system to automate generation of disease-specific medications with much reduced noise associations. We will also explore the application of those classifiers in generating other kinds of disease-specific concept vocabularies (e.g., diagnosis tests, signs, or symptoms).

5.8 Acknowledgements

The authors acknowledge Susan Rea, Bart Dodds, and Philips Jackson to provide help in understanding and querying the AHR database at Intermountain Healthcare. The authors also thank Jeffrey Ferraro for very helpful comments on earlier versions of this manuscript. This work was supported in part by Grants LM010482 from the National Library of Medicine.

5.9 References

[1] Malhotra A, Younesi E, Gündel M, Müller B, Heneka MT, Hofmann-Apitius M. ADO: a disease ontology representing the domain knowledge specific to Alzheimer's disease. Alzheimer's Dement 2014;10:238–46.

[2] Younesi E, Malhotra A, Gündel M, Scordis P, Kodamullil AT, Page M, et al. PDON: Parkinson's disease ontology for representation and modeling of the Parkinson's disease knowledge domain. Theor Biol Med Model 2015;12:20.

[3] Rahimi A, Liaw ST, Taggart J, Ray P, Yu H. Validating an ontology-based algorithm to identify patients with type 2 diabetes mellitus in electronic health records. Int J Med Inf 2014;83:768–78.

[4] Haug PJ, Ferraro JP, Holmen J, Wu X, Mynam K, Ebert M, et al. An ontology-driven, diagnostic modeling system. J Am Med Inform Assoc 2013;20:e102-10.

[5] Wang L, Bray BE, Shi J, Del Fiol G, Haug PJ. A method for the development of disease-specific reference standards vocabularies from textual biomedical literature resources. Artif Intell Med 2016;68:47-57.

[6] Wang L, Del Fiol G, Bray BE, Haug PJ. Generating disease-pertinent treatment vocabularies from MEDLINE citations. J Biomed Inform 2016;65:46-57.

[7] Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. J Biomed Inform 2010;43:891–901.

[8] Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease-drug knowledge from biomedical and clinical documents: an initial study. J Am Med Informatics Assoc 2008;15:87–98.

[9] Özgür A, Vu T, Radev DR. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics 2008;24:i277–85.

[10] Wang X, Chused A, Elhadad N, Friedman C, Markatou M. Automated knowledge acquisition from clinical narrative reports. AMIA Annu Symp Proc, 2008, p. 783–7.

[11] Cao H, Markatou M, Melton GB, Chiang MF, Hripcsak G. Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. AMIA Annu Symp Proc, 2005, p.106–10.

[12] Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform 2003;36 (3):462–77.

[13] Névéol A, Lu Z. Automatic integration of drug indications from multiple health

resources. Proc. 1st ACM Int. Heal. informatics Symp., 2010, p. 666–73.

[14] Zhang H, Fiszman M, Shin D, Miller CM, Rosemblat G, Rindflesch TC. Degree centrality for semantic abstraction summarization of therapeutic studies. J Biomed Inform 2011;44(5):830–8.

[15] Bundschus M, Dejori M, Stetter M, Tresp V, Kriegel H-P. Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinform 2008;9:207.

[16] Xu R, Wang Q. Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. BMC Bioinform 2013;14:181.

[17] Zeng Q, Cimino JJ. Automated knowledge extraction from the UMLS. AMIA Annu Symp Proc, 1998, p. 568–72.

[18] Rindflesch TC, Fiszman M, Libbus B. Semantic interpretation for the biomedical research literature. In: Chen H, Fuller S, Hersh W, Friedman C, editors. Med. informatics Knowl. Manag. data Min. Biomed., Springer; 2005, p. 399–422.

[19] Xu R, Li L, Wang Q. dRiskKB: a large-scale disease-disease risk relationship knowledge base constructed from biomedical text. BMC Bioinform 2014;15:105.

[20] Cao H, Hripcsak G, Markatou M. A statistical methodology for analyzing co-occurrence data from a large sample. J Biomed Inform 2007;40:343–52.

[21] Aslam JA, Montague M. Models for metasearch. Sigir-01 2001:285–93.

[22] Belkin NJ, Kantor P, Fox EA, Shaw JA. Combining the evidence of multiple query representations for information retrieval. Inf Process Manag 1995;31:431–48.

[23] Lee JH. Analyses of multiple evidence combination. ACM SIGIR Forum 1997;31:267–76.

[24] Agrawal R, Srikant R. Fast algorithms for mining association rules. Proc 20th Int Conf Very Large Data Bases VLDB, 1994, p. 487–499.

[25] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD Explor Newsl 2009;11:10–8.

[26] Freund Y, Schapire RE. Experiments with a new boosting algorithm. Proc. Thirteen. Int. Conf. Mach. Learn., 1996, p. 148–56.

[27] Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting. Ann Stat 2000;28:337–407.

[28] Breiman L. Bagging predictors. Mach Learn 1996;24:123–40.

[29] Dietterich TG. Ensemble methods in machine learning. MCS ’00 Proc. First Int. Work. Mult. Classif. Syst., 2000, p. 1–15.

[30] Atkeson CG, Moorey AW, Schaalz S, Moore AW, Schaal S. Locally weighted learning. Artif Intell 1997;11:11–73.

[31] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Mach Learn 1997;29:131–63.

[32] Landwehr N, Hall M, Frank E. Logistic model trees. Mach Learn 2005;59:161–205.

[33] Breiman L. Random forests. Mach Learn 2001;45:5–32.

[34] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 1997;30:1145–59.

[35] Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 2005;17:299–310.

[36] Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 1998;10:1895–923.

[37] D.M. DeLong DLC-PERD. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45.

[38] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinform 2011;12:77.

[39] Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 2006;7:1–30.

[40] García S, Fernández A, Luengo J, Herrera F. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power. Inf Sci (Ny) 2010;180:2044–64.

CHAPTER 6

DISCUSSION

6.1 Summary

To facilitate the large-scale building of disease-specific ontologies, in this dissertation, we have explored both manual and automated acquisition of disease-specific, assertional knowledge from three kinds of knowledge resources: expert-curated articles, biomedical literature, and clinical data repositories. The principal findings of each phase of the research are discussed as follows.

In the first study, we answered the question of whether it is feasible to use only a small number of expert-curated textual knowledge sources to acquire a disease-specific vocabulary that reaches a saturated coverage. For this, we manually extracted disease-specific vocabulary from a collection of documents which include clinical guidelines, textbooks, UpToDate, and DynaMed [1]. For one disease case (i.e., heart failure), the vocabulary reached near saturation in four disease aspects (i.e., *treatment*, *diagnostic tests or results*, *signs or symptoms*, and *causes or risk factors*) with the inclusion of a minimum of six sources, or between four to seven sources if only counting terms occurred in two or more sources. It took fewer sources to reach near saturation for the other two diseases regarding the *treatment* class. The principal finding of this phase's research is that it is feasible to use a limited number of expert-curated articles to obtain a

disease-specific, near-saturated vocabulary. This finding from this study is meaningful to the development of disease-specific reference vocabularies.

The second study answered a research question of whether it is feasible to automatically acquire disease-specific vocabulary from the biomedical literature. For this, we developed a pipeline-based system to automatically extract disease-specific treatment concepts from the biomedical literature. The system achieved a mean precision of 0.8 for the top 100 retrieved concepts based on three diseases case (i.e., heart failure, pulmonary embolism, and rheumatoid arthritis). When comparing four ranking strategies (i.e., occurrence, interest, degree centrality, and weighted degree centrality), although interest has a slightly better mean precision at the top 100, there is no significant difference among the four rankers. When comparing the automated results to the manual results, the pipeline-based system was able to capture over half of the concepts in the reference vocabularies. With further error analysis, we found that the overall recall can be higher. In addition, the system also captured many relevant concepts that did not exist in the reference vocabularies. A prerequisite of achieving higher recall is the improvement of the natural language processing tool used to process the biomedical literature. An improvement of the semantic schema may further improve the precision because many concepts were not treatment related. From this study, we concluded that the pipeline-based system we developed is a promising tool for an automated extraction of disease-specific treatment vocabulary for any disease of interest.

The third study investigated whether classifiers, generated using machine learning techniques, can be used to reduce the manual effort necessary to review noisy collections of disease-specific concepts extracted from both the biomedical literature and the clinical

data repository. For three types of datasets compiled from different sources (i.e., biomedical literature, clinical data repository, and combination of the two sources), the results favor different "preferred" classification models. These were simple logistic regression, multilayer perceptions, and LogitBoost, respectively. The study results show that the classifiers developed with the combined datasets significantly outperforms the classifiers developed with either the biomedical literature dataset or the clinical data repository dataset. The results also show that the performance of a classifier on a specific disease dataset shows no significant difference when trained on the same or on another diseases' dataset. Therefore, we concluded that it is a promising approach to use classification techniques based upon different measures of relevance to reduce the noisy collections of disease-specific concepts extracted from the biomedical literature and clinical data bases. Combining the features from disparate sources improved the performance of classification. The classifiers trained with the dataset of known diseases could be generalized to new diseases.

When comparing the three kinds of knowledge resources we have explored, we note different merits and limitations. For expert-curated documents, the main advantage is that they intensely contain disease-specific information, so that with only a few documents, the extracted vocabulary can reach near-saturation. However, the extraction of medical knowledge from them remains in a manual way. One reason for that is the variety of the representation of the knowledge in these documents, such as tables and figures, which brought new challenges to the natural language processing tools especially in understanding the content of the documents. Therefore, a manual acquisition with maximal assistance from the computer (such as pre-annotation with a dictionary) could

be currently desirable for extracting knowledge from this kind of documents. For MEDLINE citations, it is a popular knowledge resource used for knowledge extraction in the biomedical domain. A key advantage of using MEDLINE citations is that MEDLINE broadly covers a variety of diseases and aspects (e.g., treatment) that human beings have explored. However, some disadvantages are that as we learned from the study in Chapter 4: (1) MEDLINE citations may contain out-of-date information which is hard to discern; (2) an NLP tool is necessary to handle free-text data. However, preprocessing with NLP could also introduce noise. A common issue when dealing with both expert-curated documents and MEDLINE citations is that the concept granularity for human writing can contain vague or inconclusive statements. For the clinical data repository, the overall precision of extracted vocabularies was relatively lower than the precision of vocabularies generated from the biomedical literature (Table 5.2). However, we found in the study in Chapter 5 that it was still very useful for the knowledge extraction; it improved the overall recall after combining the results of extraction from both the biomedical literature and clinical data repository and the performance of classification when discerning true-relevant concepts.

There are three potential merits of automated acquisition comparing to manual acquisition. First, automated techniques offer almost an instant retrieval of possible disease-related concepts from the biomedical literature and clinical data repositories. Although the initial results of automated extraction contained noisy information which required substantial manual investigation, applying classifiers could possibly reduce a collection of noisy information and therefore may reduce the effort of manual review. The manual acquisition from expert-curated documents involves multiple steps including

preparation of documents, annotation and adjudication, and concept mapping, none of which is trivial. Based on the experience with the three studies in this dissertation, the automated acquisition would be more efficient than the manual approach; however, a further evaluation would be necessary to prove this. Second, mining huge amounts of data with automated techniques could provide new information that was missed from the expert-curated documents. Third, comparing to the concepts extracted from expert-curated documents, the concepts extracted from the MEDLINE citations preserve the links to the origin evidence (or individual clinical trial studies). This could be useful for some information seekers.

6.2 Significance of Contributions

Knowledge acquisition is one of the core topics of clinical informatics [2]. The dissertation adds contributions to the body of literature in disease-specific knowledge acquisition from existing biomedical knowledge resources in three aspects. First, we provided a mechanism to build disease-specific reference vocabularies and verified the amount of sources required in order to build vocabularies achieving near saturation. Second, we developed a novel pipeline-based approach to mine MEDLINE citations for disease-specific treatment concepts and relations. Third, we used classification to incorporate disparate sources for an automaticaaly generated disease-specific vocabulary with a control of the signal-to-noise ratio. The technologies we explored in this research lay a foundation of a clinical knowledge authoring and sharing service (cKASS) [3] which would assist people to build disease-specific ontologies.

The dissertation adds contributions to the development of knowledge-based clinical applications through providing disease-specific computerized medical knowledge.

Those applicable domains in general cover but are not limited to information retrieval, clinical decision support systems, and data analytics in healthcare.

6.3 Limitations

The research described in this dissertation has mainly three limitations. First, the disease-specific information focused on by the three studies were narrowed down from four classes (i.e., *causes or risk factors*, *signs or symptoms*, *diagnostic tests and results*, and *treatment*) in Chapter 3, to one class (i.e., *treatment*) in Chapter 4, and to a subclass of *treatment* (i.e., medication) in Chapter 5. Although we argued that the automated techniques could be extensible to extract other diseases classes, at this point, the performance of the pipeline-based system in Chapter 4 and classifiers in Chapter 5 is unknown in the unstudied disease classes. Second, the reference vocabularies we developed in the study in Chapter 3 and used for the other two studies were not perfect and had gaps; therefore, the performance reported in the studies in Chapter 4 and 5 (e.g., precision and recall) may not reflect the true performance. Third, we mainly evaluate each phase of the study with three testing diseases: heart failure, pulmonary embolism, and rheumatoid arthritis. With a small sample of testing diseases, the results and conclusion generated from them may not be representative of the entire disease pool.

6.4 Generalizability of the Results

We discuss the generalizability of the results from two aspects: whether the results were generalizable to other diseases, disease classes, and beyond disease-specific information.

The principal finding of the first research study in Chapter 3 is that using a limited

number of expert-curated articles is able to produce a disease-specific, near-saturated vocabulary. We believe that this finding applies to the majority of diseases. We observe that as the complexity of the diseases increase, the number of documents used for achieving a near-saturated vocabulary would also be slightly increased. Heart failure, one of the three diseases we have explored, is among one of the most complex diseases. Therefore, diseases with less complexity probably require an equal or smaller number of documents comparing to heart failure. Besides, the finding applies to four classes of disease-specific concepts, including causes or risk factors, signs or symptoms, diagnostic tests or results, and treatments. Although one disease was tested, we find that the four classes achieved near saturation at the similar speed.

In the study of Chapter 4, we developed a pipeline-based system which is able to extract disease-specific treatment vocabularies. It achieves a mean precision of 0.80 on the top 100 concepts. As we have tested on five diseases, the precision would vary slightly among diseases (Table 4.6). We argue that the main framework of the pipeline-based system could be reused and extended to extract concepts of other disease classes, such as causes and risk factors. However, the performance in other disease classes is currently unknown.

In the study of Chapter 5, the principle findings are that combining features from disparate sources would improve the performance of classification, and the classifiers built from some diseases could be generalizable to new diseases. We believe that the results of this study would be generalizable to the majority of diseases, although there might be an exception for some rare diseases which have a small number of entries in MEDLINE and few records in the clinical data repositories. We assume the finding

would be generalizable to other disease classes; however, further study is required.

6.5 Future Directions

While this dissertation has demonstrated the potential of automated extraction of disease-specific treatment information from expert-curated documents, biomedical literature, and a clinical data repository, many opportunities for extending the scope of this dissertation remain. This section presents two of these directions.

6.5.1 Extend the Scope of this Dissertation

We will extend the scope of this dissertation in three aspects: the techniques, the data (or knowledge sources), and the types of disease-specific information.

In terms of disease-concept associations, the automated techniques developed in this dissertation mainly focused on extracting disease-specific treatment information. In the future, we will fully extend the techniques to three other classes, including disease-specific causes and risk factors, diagnosis tests, and signs and symptoms. For the pipeline-based system built in Chapter 4, we expect to add new semantic schemas in order to capture the information in those expanded classes. In addition, we will expand to capture other information about the disease-specific vocabularies, such as synonyms, definitions, and PubMed IDs in order to make the ontologies meet different needs. For example, synonyms would be useful for the annotation of the biomedical literature and electronic medical records.

In terms of the data, the knowledge resources we have exploited are attributed to a very small portion of the big data available in the world. The unstructured data is intact in this study. Some disease-*concept* associations like signs or symptoms, or causes and risk

factors may not be easily captured from structured data sources. We have to expand our work to use unstructured data (e.g., clinical notes) in order to capture different kinds of disease classes.

To unveil those disease-specific associations from large and unstructured data sets, we will exploit other techniques besides the ones developed in this dissertation. With recent advanced work in the artificial intelligence field (e.g., neural networks), we found some interesting techniques that could be potentially used for mining semantic association from unstructured data. For example, neural word embedding (e.g., word2vec), a technique used for computing continuous vector representations of words, is able to capture a large number of precise syntactic and semantic word relationships from very large data sets [4,5], which might be also useful for our purpose.

6.5.2 Knowledge Authoring Tool

In order to obtain practical disease-specific ontologies, it is necessary to develop a tool that allows human experts to manipulate and validate automated machine generated results [3]. We intend to integrate the techniques exploited in this dissertation as well as some techniques investigated by other researches into the tool to provide a single platform to generate disease-specific vocabularies as well as a user interface for the interaction with domain experts for validation of the relevance of the extracted concepts. We would implement the machine-learning-based classifications into the tool for pre-selection of relevant concepts from the noise. After the tool was developed, further evaluation on the system's performance and user satisfaction would be needed.

6.6 References

[1] Wang L, Bray BE, Shi J, Del Fiol G, Haug PJ. A method for the development of disease-specific reference standards vocabularies from textual biomedical literature resources. Artif Intell Med 2016;68:47-57.

[2] Gardner RM, Overhage JM, Steen EB, Munger BS, Holmes JH, Williamson JJ, et al. Core content for the subspecialty of clinical informatics. J Am Med Informatics Assoc 2009;16:153–7.

[3] Wang L, Zhang M, Conway M, Haug P, Chapman W. Using cKASS to facilitate knowledge authoring and sharing for syndromic surveillance. Emerg Health Threats J 2011:11147.

[4] Mikolov T, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. NIPS, 2013.

[5] Mikolov T, Corrado G, Chen K, Dean J. Efficient estimation of word representations in vector space. ICLR Work., 2013.