A CONCEPT-BASED CLUSTERING APPROACH TO CLINICAL INFORMATION RETRIEVAL

A thesis submitted for the degree of
Doctor of Philosophy

by

Vong Wan Tze

Faculty of Engineering, Computing and Science
Swinburne University of Technology
Sarawak, Malaysia

2016

ABSTRACT

Current medical question-answering (MedQA) systems assume that users have a clear understanding of their search targets and are aware of their knowledge deficit when formulating a clinical question. Less emphasis has been placed on strategies to assist users in clarifying and recognizing their information needs during the information search process. The PICO, an acronym for population/problem, intervention, comparison and outcome, is a question framework for formulating well-defined and answerable clinical questions. In this thesis, the question framework was used to extract key medical concepts from a collection of documents. A concept similarity clustering approach was then applied to organize and visualize the collection into a hierarchy of relevant concepts for browsing, exploring and searching purposes. CliniCluster is a semi-automated clinical question answering engine designed with the capability to support and assist users in narrowing down and better understanding their search intent, and in finding documents that best match their search request.

The studies described in this thesis can be divided into four main parts. The first part details the text processing and knowledge extraction methods employed to mine PICO elements from documents resulting from a set of test questions. Besides, a series of statistical separation tests were conducted to determine the most effective combination of weighting scheme and similarity/distance metric for concept-based similarity measurement between documents. In the second part, using both well-formulated and poorly-formulated questions, a comparative study was performed to determine the most effective agglomerative hierarchical clustering algorithm and the most appropriate hierarchical structure for clustering and visualization of a collection of documents. The third part evaluates the performance of CliniCluster compared to three existing search engines in retrieving highly relevant documents using known-item search method. The last part is a pilot questionnaire survey conducted among a group of health care providers to investigate the usability and user satisfaction with the support and assistance provided by CliniCluster.

The main contributions of this thesis fall into four categories. First, separation tests revealed that the “titles and abstracts" contain the most salient PICO elements to represent each of the retrieved documents, and the combination of “binary” weighting scheme and “Yule”/“Yule2” similarity metric is the most effective method to measure the concept-based similarity between documents. Second, cluster structure analysis

showed that the clustering algorithm, Ward-Link", produces the most appropriate hierarchical structure to organize and visualize a collection of documents in a hierarchical manner. Besides, an exhaustive search of documents can be avoided by cutting a hierarchy at a certain level and by labelling each cluster in the hierarchy with the most representative therapy topics. Third, using known-item search method, CliniCluster was found superior to CQA-1.0, Google and Google Scholar in ranking highly relevant and evidence-based documents at higher positions in search results. Lastly, the pilot survey conducted among health care providers revealed that the majority of the respondents agreed that CliniCluster assisted them in narrowing down search results and in quickly identifying relevant documents, and they were satisfied with the ease of completing a search task using CliniCluster. The overall results showed that the proposed concept similarity clustering approach can be used to organize and visualize a collection of documents to support the search of relevant documents. Besides, CliniCluster was found to have the capability to support and assist users in finding and recognizing highly-relevant and evidence-based documents for clinical question answering.

DECLARATION

I, Vong Wan Tze, hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material that to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due reference is made in the text.

Wan-Tze Vong

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my principal supervisor, Associate Professor Patrick Then, whose expertise, understanding and patience added considerably to my graduate experience. I appreciate his vast knowledge and skill in many areas, and his assistance in writing papers and this thesis. I would also like to express my sincere gratitude to my co-supervisor, Dr. Alan Fong, for his help and advice on all aspects of the project.

I must also acknowledge Associate Professor Manas Kumar Haldar, Sim Kwan Yong and Sim Kwan Hua for their valuable suggestion and constructive criticism during the weekly group meeting. I would like to thank Tiong Lee Len for helping me recruit doctors and pharmacists during the validation phase of the project. A very special thanks goes out to Associate Professor Enn Ong for her motivation and encouragement throughout the project work.

I would also like to thank my family for the support they provided me through my entire life. I extend many thanks to my past and current research group members Lesley Lu, Chee-Ping Lai, Brian Loh, Yakub Sebastian, Bismita Choudhury, Caleb Lai and Isaac Goh for their encouragement and support.

In conclusion, I recognize that this research would not have been possible without the financial assistance of Swinburne University of Technology and the cooperation of CRC SGH (Clinical Research Centre at Sarawak General hospital). I express my gratitude to those agencies.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Acronym	Definition	Acronym	Definition
AB	Abstract	[P]	Problem/Population
AL	Average-Link	$\overline{P11}$	11-Point Interpolated Precision
ANOVA	Analysis of Variance	PICO	Problem/Population, Intervention,
AP	Average Precision		Comparison and Outcome
ARP	Average R-Precision	*PG*	Percentage Gain
AUC	Area Under the Curve	PMID	PubMed Unique Identifier
BO	Binary Occurrence	POEs	Patient-Oriented Evidence
[C]	Comparison	POEM	Patient-Oriented Evidence that
CL	Complete-Link		Matters
CUI	Concept Unique Identifier	QA	Question Answering
Doc	No. of Documents	*R*	*Recall*
DOEs	Disease-Oriented Evidence	RN	Chemicals
EBM	Evidence-based Medicine	*RP*	R-precision
EEP	Essential Evidence Plus	S_{Date}	Score of Recency
F	F-measure	$S_{Journal}$	Score of Journal
F_{max}	Maximum F-measure	S_{MD}	Score of Mean Difference
FPR	False Positive Rate	S_{SOE}	Score of Strength-of-Evidence
HLvl	Hierarchy Level	S_{Study}	Score of Study Design
[I]	Intervention	SOR	Strength-of-Recommendation
LoE	Level of Evidence	SORT	Strength-of-Recommendation
MAP	Mean Average Precision		Taxonomy
MedQA	Medical Question Answering	TF	Term Frequency
MeSH	Medical Subject Heading	TFIDF	Term Frequency-Inverse Document
MH	MeSH Terns		Frequency
MMTx	MetaMap Transfer	TI	Title
MRR	Mean Reciprocal Rank	TO	Term Occurrence
N_{Rel}	No. of Relevant Documents not	ToX	Toronto XML
	Retrieved	*TPR*	True Positive Rate
NLM	National Library of Medicine	ROC	Relative Operating Characteristic
N_{Irrel}	No. of Irrelevant Documents	UMLS	Unified Medical Language System
	Retrieved	WL	Ward-Link
M_{Rel}	No. of Relevant Documents not	$\%Rel$	Percentage of Relevant Documents
	Retrieved		
[O]	Outcome		
P	Precision		
P@k	Precision at *k*		

1. CHAPTER I: Introduction

Physicians seek information to answer patient-care questions. Besides, they must cope with the rapidly growing body of medical information in order to stay current with the latest medical development. Not surprisingly therefore, information retrieval systems have become an indispensable tool for searching and gathering of medical information. However, there are numerous barriers to the uptake of clinical evidence for patient care. The main barriers include lack of time, limited literature searching skills and limited ability to identify the best available evidence. Therefore, medical question-answering systems have been developed on recent years to assist physicians in quickly locating high quality and truly useful clinical information

1.1. BACKGROUND

The purpose of this section is to give some background of the information needs of physicians and their information seeking behavior at the point of care.

1.1.1. Why do physicians search for clinical information?

The use of the Internet has shifted the role of patients from passive recipients to active consumers of healthcare. Patients are becoming more knowledgeable about their health conditions and are empowered to get more involved in information sharing and decision making (Saca-Hazboun, 2007; Schardt et al., 2007; Gordon, 2011). Despite the change of patient-physician relationship, health care providers remain the most influential source of information among "internet informed" patients for medical decisions (Couper et al., 2010). The "internet informed" patients, however, have been shown to increase the pressure of health professionals (McMullan, 2006; Ahluwalia et al., 2010). To improve patient satisfaction, physicians are forced to become acquainted with information technology, and to keep up with timely evidence-based clinical information.

On the other hand, the translation of knowledge from research to practice has been a major challenge to promote high quality patient care. Early studies in the US and the Netherlands found that 30 to 40% of patients do not receive clinical interventions based on the best existing scientific evidence and up to 25% of care provided is potentially harmful or unnecessary (Schuster et al., 1998; Grol, 2001). A study by Jones et al. (2003) revealed that about two-thirds of child deaths could be prevented by effective

and affordable interventions. Further examples of ineffective and costly treatments that reduce the quality of patient care are reported by Hutin et al. (2003), Attaran (2004) and Corcoran et al. (2010). To ensure that the most effective care is delivered to patients, physicians are encouraged to search for the best available clinical evidence in order to support their clinical decision making processes.

1.1.2. What types of clinical information do physicians need?

The information needs of physicians have been investigated by a considerable number of studies. An early study by Smith (1996) reported that approximately 33% of information needs related to treatment of specific conditions, 25% to diagnosis and 14% to drugs. Similar findings were found by Davies (2007), who reported that the top three categories of information needs were treatment/therapy (38%), diagnosis (24%) and drug therapy/information (11%). A study investigating the use of online evidence-based resources by physicians at the point of care revealed that therapy, prognosis and epidemiology questions were the most common types of inquiries (Schwartz et al., 2003). Yu and Cao (2008), on the hand, analyzed 4654 clinical questions maintained by the National Library of Medicine (NLM). The authors found that 34.3% of the questions were on pharmacology, 30.1% on management, 21.4% on diagnosis and 18.7% on treatment and prevention. A recent systematic review collected a total of 7012 questions raised by clinicians at the point of care (Del Fiol et al., 2014). The study found that 34% of the questions concerned drug treatment, and another 24% concerned physical finding, potential causes of a symptom or diagnostic test finding. In summary, the available evidence indicates that the physicians' greatest information needs is for information about treatment/therapy and drugs.

1.1.3. How do physicians search for clinical information?

Physicians often have very tight schedules. When seeking information for patient care, they are more likely to look for information from readily available resources. There are several options for physicians to search for clinical information: evidence-based medicine databases (e.g. Cochrane Library and UpToDate), medical question-answering systems (e.g. InfoBot and AskHERMES), bibliographic databases (e.g. MEDLINE and EMBASE), or through an intermediary such as a clinical librarian. Despite the ubiquity of electronic resources, textbooks and colleagues remain the most

frequently used resources for medical information (Ely et al., 2005; Davies, 2011; Kosteniuk et al., 2013). Besides, MEDLINE/PubMed is the most widely used electronic resource by junior doctors and physicians for systematic reviews and primary studies (Schilling et al., 2005; Cullen et al., 2011; Davies, 2011).

1.2. MOTIVATION

This section gives an account of the impact of evidence-based medicine (EBM), a scientific approach to teaching the practice of medicine, followed by a discussion of the barriers to implementing EBM. The section ends with a brief introduction of medical question-answering system for EBM practice.

1.2.1. The Impact of EBM training

The term EBM, as described by Sackett et al. (1996), is "*the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patient*". The practice of EBM involves four main steps:

1. Formulating a well-focused clinical question from a patient's problem,
2. Searching medical databases comprehensively for relevant articles,
3. Appraising the validity and clinical applicability of evidence critically, and
4. Implementing the most useful evidence in clinical practice.

The four steps are designed to ensure that the best available evidence can be identified from research studies. The best available evidence, integrated with patient's circumstances and preferences, is then used to support clinical decision-making.

A considerable number of studies have investigated the impact of EBM on physicians' knowledge, attitude and practice. A survey involving 545 physicians found that, following literature search and explicit appraisal, 39% of physicians gained an improved knowledge, 47% had an increased level of confidence in pre-existing clinical decisions, and 5% changed their clinical decisions (Scott et al., 2000). Similar outcomes were found by Markey and Schattner (2001), Lucas et al. (2004) and Straus et al. (2005). A qualitative study involving facilitators and physicians from a large health maintenance organization (Shuval et al., 2007), and a cross sectional survey involving 966 physicians in Norway (Ulvenes et al., 2009) found that the majority of the respondents agreed that EBM leads them towards better practice. A recent survey by

Heighes and Doig (2014) reported that, out of the 130 intensive care specialists, 65.4% of them expressed positive attitudes toward the use of research evidence in clinical practice, and 96.6% of them reported the use of the concepts of EBM at least sometimes. In summary, physicians generally hold positive attitudes towards EBM.

1.2.2. Barriers to implementing EBM

There are numerous barriers to effective evidence-based practices, causing the uptake of clinical evidence by physicians slow and reluctant. Lack of time is the most mentioned barrier to implementing EBM (Davies, 2007; Sadeghi‐Bazargani et al., 2014). A study by Schwartz et al. (2003) found that physicians took about five to ten minutes to obtain an answer from online resources. This time-consuming process limits the use of online resources during patient consultation. Other reported constraints to implementing EBM include limited information technology skills (Lappa, 2005), lack of interest (Ely et al., 2007) and the financial cost of information searches (Andrews et al., 2005; Sadeghi‐Bazargani et al., 2014). Ely and colleagues (2005), on the other hand, investigated the obstacles preventing physician from answering patient-care questions. The obstacles are broadly classified into physician-related and resource-related obstacles. Physician-related obstacles include lack of awareness of an information need, doubt that an answer existed, failure to select the most appropriate resource, and the tendency to formulate unanswerable questions. Resource-related obstacles include excessive time and effort spent searching for answers to clinical questions, deficient access to information resources, failure to identify information need from large volumes of literature, inability of information search engine to answer questions directly, and failure of the selected resources in providing an answer. A systematic review by Zwolsman et al. (2012) concluded that the most commonly reported barriers to the use of best available evidence are insufficient time, deficient EBM skills and the availability of evidence. Similar findings were reported by De Fiol et al. (2014), who found that the main barriers to information seeking by clinicians at the point of care are lack of time and doubt that a useful answer existed.

To conclude, previous studies investigating barriers to evidence-based practice indicate that, in response to a patient-care question, physicians require support and assistance during the search process in order to quickly identify relevant information to answer the question.

1.2.3. Towards medical question-answering system

To better serve the information needs of physicians, medical question answering (MedQA) systems have emerged as a new generation search engine. A question-answering system is an information retrieval application which aims at returning a short and precise answer to a natural language question. An example[1] of question raised by physicians during the point-of-care and the recommended answer is given as follows:

Question: "*Are COX-2 inhibitors and selective nonsteroidal anti-inflammatory drugs safe for adults with aspirin-exacerbated respiratory disease?*"

Answer: "*This review found no evidence of any effect of oral COX-2 inhibitor exposure in adults with asthma and aspirin-sensitivity. Exposure to selective nonsteroidal anti-inflammatory drugs (NSAIDs; e.g. meloxicam) significantly increased respiratory symptoms, but the clinical significance of this effect is uncertain.*"

Although MedQA systems are not the most widely used resource for health information, recent reviews by Athenikos et al. and Bauer et al. (2010; 2012) demonstrated that MedQA systems are improving and are close to becoming valuable tools for the search of quick and reliable information for EBM practice.

1.3. PROBLEM STATEMENT

Transforming an information need into a well-focused question is the first step to practicing EBM. However, when finding answers to clinical questions,

1. Physicians are unaware of their information needs, and
2. Physicians have the tendency to formulate unanswerable questions.

The capability of a QA system in retrieving highly-relevant documents depends on the quality of the input question. Current MedQA systems assume that:

1. Users have a clear understanding of their information needs, and
2. Users are aware of their knowledge deficit and are able to formulate answerable questions or searchable keyword queries.

---

[1] The question-answer pair was delivered as Daily POEM on 4th September 2014 by Essential Evidence Plus, a point-of-care clinical decision support system available at: www.essentialevidenceplus.com.

More specifically, users may encounter the following problems when using current MedQA systems:

1. Difficulty in clearly defining and expressing their information needs, due to:
    - A lack of knowledge of a particular problem domain, and
    - A misplaced expectation that a system is aware of their information needs.
2. Difficulty in formulating well-focused questions or keyword queries, due to:
    - Inability to construct a question using relevant and appropriate vocabulary, especially for specialized domain with specific terminology, and
    - Inability to use advanced query language syntax such as Boolean operators, or a lack of a clear understanding of the query framework used by a system.

When performing a search task using a MedQA system, users need support and assistance in exploring a specific problem domain by understanding the relevant terminology, concepts or topics, and in the process, clarifying and meeting their information needs.

1.4. PROPOSED GOALS

In this thesis, a two-stage approach is proposed to support and assist users in clarifying and meeting their information needs. The two stages are:

- An exploratory stage to visualize a collection of documents for browsing, exploring and searching purposes, and
- A concept stage to visualize useful and important concepts for searching and recognition of the most relevant documents.

The approach is intended to improve the information search process by allowing users to gain a better understanding of the concepts or topics related to a search query, to narrow down, refine or clarify their search intent by exploring a collection of documents, and to quickly locate documents that best match their information needs. It is expected that:

- When users have clear information needs, the approach can assist them in clarifying their search intent, and
- When users have vague information needs, the approach can assist them in exploring a problem domain and guiding them toward their search goal.

1.5. OBJECTIVES

In order to achieve the proposed goals, the studies described in this thesis were undertaken with the following objectives:

- An inter-document similarity analysis that identifies:
  - The most appropriate text field of structured documents for knowledge extraction (semantic extraction of key medical concepts), and
  - The combination of weighting scheme and similarity/distance metric that is most effective for measuring concept-based document similarity.
- A cluster structure analysis that investigates:
  - The most effective concept similarity clustering algorithm to organize a collection of documents into meaningful clusters, and
  - The most appropriate hierarchical structure to visualize the collection as a tree of key medical concepts for information search purposes.
- A known-item search method and a pilot survey that explore:
  - The performance of the proposed clinical question answering engine in retrieving and ranking high quality and evidence-based documents, and
  - The usability and user satisfaction with the proposed engine in supporting and assisting users during the information search process.

1.6. SCOPE of RESEARCH

The scope of the research is limited to:

- The utilization of a clinical question framework as the basis for the extraction of key medical concepts from structured documents – "Knowledge Extraction".
- The development of a concept similarity clustering approach to visualize a collection of document in a graphical form for information seeking and retrieval – "Document Visualization".
- The development of a semi-automated question answering engine with the capabilities to support and assist users in clarifying and recognizing their information needs – "Information Search Support".
- The strategies to improve the information search process for high quality and evidence-based documents that best answer questions that are most frequently asked by physicians – "Therapy Questions".

1.7. CONTRIBUTIONS of RESEARCH

The contributions of the studies presented in thesis include:

- A novel combination of text processing, semantic-based knowledge extraction and pairwise similarity techniques for knowledge extraction and concept-based document similarity measurement,
- A novel concept similarity clustering approach to visualize a collection of documents as a hierarchy of relevant concepts for browsing, exploring and searching purposes, and
- A semi-automated clinical question answering engine with the capabilities to support and assist users in searching and retrieving evidence-based documents.

1.8. THESIS OUTLINE

This section gives an overview of the contents to be found in the following chapters. **Chapter 2** of this thesis starts by introducing different search strategies to identify valid and reliable clinical evidence, followed by reviewing the state of the art of MedQA systems. The proposed solution to the research problem and the architecture of the proposed clinical question answering engine are described in **Chapter 3**.

The next two chapters give details about the proposed concept similarity clustering approach. **Chapter 4** discusses how the test documents were collected and processed for the extraction of key medical concepts. Besides, different weighting schemes and similarity/distance metrics were tested for concept-based similarity between documents. **Chapter 5** explains how different similarity-based hierarchical structures were constructed and evaluated for their effectiveness in grouping documents into meaningful clusters. In addition, a series of information retrieval tests were performed to determine the most appropriate hierarchical structure for document visualization.

In **Chapter 6**, using both well-formulated and poorly-formulated questions, the performance of the proposed engine in retrieving and ranking highly relevant and evidence-based documents were compared to three existing search engines. **Chapter 7** discusses the results of a questionnaire survey, which was conducted among health care providers to determine the usability and user satisfaction with the proposed engine.

Finally, the thesis concludes with summary of findings, limitations of the projects and directions for the future research in **Chapter 8**.

2. CHAPTER II: Literature Review

Increased access to information can lead to more informed and effective decision making. However, information seeking process is a time-consuming and difficult task. Therefore, intelligent information retrieval systems have been extensively studied to summarize relevant and reliable textual sources. Question answering is a form of information retrieval that deals with natural language questions and aims to return precise short answers. This chapter is divided mainly into three sections.

- **Section 2.1** discusses the strategies to search literature effectively in order to answer therapy questions using the best available clinical evidence.
- **Section 2.2** reviews the approaches and resources that have been employed to develop question-answering systems for the clinical domain.
- **Section 2.3** describes how two existing MedQA systems are designed to support and assist users during the information search process.

2.1. STRATEGIES TO FIND THE BEST EVIDENCE

The first step of EBM is to convert an information need into a focused and searchable question. A number of question frameworks have been introduced previously for the formulation of clinical questions and are discussed in **Section 2.1.1**. The second to fourth steps of EBM involve a systematic search of literature, critical appraisal of research evidence and the use of the best evidence in clinical practice. The strategies that have been used to search and determine the quality of information for clinical practice are discussed in **Section 2.1.2**.

2.1.1. Framing Answerable Question

The Question Frameworks

To formulate an answerable question, physicians are recommended to modify their search strategies by rephrasing their questions or to use question frameworks (Ely et al., 2007). The PICO framework proposed by Richardson et al. (1995) has been widely accepted for the formulation of well-defined clinical questions. The framework aims to break down a clinical question into searchable keywords, and is described as follow:

- [P] stands for **P**opulation or **P**roblem that gives information about an individual patient or a group of patient, and/or the primary problem, disease or co-existing conditions that requires clinicians' care.
- [I] stands for **I**ntervention, which describes the treatment, diagnostic test, prognostic factor, or exposure of interest.
- [C] stands for **C**omparison and is usually an alternative to the intervention of interest. In some cases, there is no comparison group.
- [O] stands for **O**utcome that gives information about the result of interest. This can be the outcome of an intervention or an exposure. Generally, patient-oriented outcomes are preferred.

Other question frameworks that have been introduced recently include:

- PESICO: **P**roblem/**P**opulation, **E**nvironment, **S**takeholder, **I**ntervention, **C**omparison and **O**utcome (Schlosser et al., 2007),
- PICOS: **P**roblem/**P**opulation, **I**ntervention, **C**omparison, **O**utcome and **S**tudy Design (Atkins et al. 2011),
- PICOT: **P**roblem/**P**opulation, **I**ntervention, **C**omparison, **O**utcome and **T**ime frame (Rios et al. 2010), and
- SPIDER: **S**ample, **P**henomenon of **I**nterest, **D**esign, **E**valuation and **R**esearch Type (Cooke et al. 2012).

Despite of these different question frameworks, recent studies support the use of PICO for the formulation of clinical questions (Brożek et al., 2009; Rzany, 2009; Sultan et al., 2013; Moyer and Neuspiel, 2014). A recent study by Methley et al. (2014) concluded that PICO is more effective than PICOS and SPIDER for the comprehensive search of systematic reviews. Besides, Nixon et al. (2014) and Schardt et al. (2007) found that the use of PICO can improve the quality of answers or the relevancy of search results. In this regard, it is worthwhile to continue to use PICO for the formulation of answerable questions.

Therapy Questions in PICO Format

Therapy question is a question concerning the effectiveness of a treatment (e.g. medications and surgical procedures) or preventative measure (e.g. immunizations). Two therapy questions[2] in PICO format are given below:

- "In children with acute asthma exacerbations, is oral or injected dexamethasone as effective as predisone or prednisolone?"

  [P]: children with acute asthma exacerbations

  [I]: oral or injected dexamethasone

  [C]: predisone or prednisolone

  [O]: -

- "Is duloxetine effective in reducing pain from chemotherapy-induced peripheral neuropathy in adult cancer survivors?"

  [P]: chemotherapy-induced peripheral neuropathy in adult cancer survivors

  [I]: duloxetine

  [C]: -

  [O]: reducing pain

Huang et al. (2006) investigated the adequacy of PICO framework as a knowledge representation for clinical questions. The authors found that the PICO framework is particularly useful for formulating therapy questions. Five structural patterns of therapy questions identified from the study are presented in **Table 2.1** (Patterns I-II are the most common and Patterns III-V are less common). As shown in the table, a question mark indicates the element that serves as the answer to a question. For example, [O?] indicates that outcome is the desired answer of a question. Besides, each pattern of question contains different combinations of PICO elements and not all patterns of questions have all four PICO elements present. The authors found that there is a lack of elements that comprise a well-formed query in most of the clinical questions. On the other hand, an early study by Bergus et al. (2000) found that questions that contain a proposed intervention, [I] and a relevant outcome, [O] are unlikely to go unanswered. Another study by Staunton (2007) reported that at least 3 of the PICO elements are

[2] The therapy questions were taken from the EBM database, Essential Evidence Plus, a point-of-care clinical decision support system available at http://www.Essentialevidenceplus.com/content/poems.

needed to formulate an answerable question. These studies indicate that the completeness of PICO elements in a question determines whether it is likely to be answered. Despite these findings, it cannot be assumed that forcing a question into PICO will certainly resolve a physician's information need. Booth et al.(2000) demonstrated that PICO-structured questions allowed librarians to conduct more precise searches. However, the questions often included only the [P] and [I] elements. The authors further reported that free-form questions elicit the purpose of the information request improved the relevance of retrieved records. The critical task when developing a PICO question is using appropriate and relevant terminology (Hoogendam et al., 2012; Hastings and Fisher, 2014). A recent literature review by Fourie (2009) identified that health professionals have difficulty articulating and recognizing their information needs, and tend to express a level of uncertainty and anxiety when identifying their information needs. In this aspect, more studies need to be done to investigate the use of PICO framework in assisting health professionals in meeting their information needs.

Table 2-1. Five structural patterns of therapy questions

Pattern	PICO Structure	Question
I	[P][I][O?]	Is enoxaparin useful for moderate renal impairment?
II	[P][I?]	What is the best treatment for acute otorrhea?
III	[I][O?]	Does supplemental vitamin D increase bone mineral density?
IV	[P][I?][O]	Is acupuncture effective in relieving pain in patients with chronic low-back pain?
V	[P][I][C][O?]	What is the comparative effectiveness of ondansetron and metoclopramide for treatment of hyperemesis gravidarum?

2.1.2. Searching for the Best Evidence

Patient-Oriented Evidence.

Clinicians are advised to look for the most useful information based on the strength of evidence provided by a study (Ebell et al., 2004). There are two types of research evidence: disease-oriented evidence (DOEs) and patient-oriented evidence (POEs). DOEs refer to the outcomes of studies that measure intermediate,

histopathology, physiologic or surrogate markers of health. For instance, the measurement of blood pressure, hemoglobin and resting heart rate that may or may not reflect improvement in patient outcomes. POEs refer to the outcomes of studies that matter to patients. These include improvement in symptoms, morbidity, mortality, quality of life and cost that can help patients to live longer or better lives.

Articles containing POEs that have the potential to change practice are called patient-oriented evidence that matters (POEMs). They contain information that has emerging roles in monitoring patients, in operationalizing and evaluating disease management programs, and in quality assessment and improvement. Ebell et al. (1999) reported that busy physicians have to read only 2% of the original studies published each month by focusing on medical journals that publish POEMs. Similar results were found by McKibbon et al. (2004) who investigated the "number of articles needed to be read" (NNR) by physicians in 170 primary healthcare journals. Both studies concluded that POEMs are concentrated in a small subset of journals. On the other hand, MEDLINE[3] provides the "Core Clinical Journals" filter to restrict literature search to 119 journals particularly relevant to practicing physicians (US National Library of Medicine, 2014). The findings suggest that the most useful information for clinical practice can be identified more effectively by focusing on journals that publish POEMs.

Strength of Recommendation.

Clinical recommendation should be made based on the highest quality evidence available. Seven systems were identified by the Agency for Healthcare Research and Quality (AHRQ) that fully addressed the three characteristics: quality, quantity and consistency for grading the strength of a body of scientific evidence (Owens et al., 2010). One of the most popular grading systems is the Strength-of-Recommendation

[3] MEDLINE is the largest and most widely used medical bibliographic database. PubMed is a search engine that offers access to MEDLINE. Both MEDLINE and PubMed are developed and maintained by the NLM. PubMed currently comprises over 24 million citations for biomedical literature from MEDLINE, life science journals and online books. Each citation contains the article title, author(s), publisher, publication date, and if available, MeSH terms, abstract and link(s) to full-text articles. Each of these fields is indexed separately by PubMed. Users can identify potentially interesting articles using appropriate search terms in PubMed, and obtain the full text of a selected article by clicking on a publisher's link. A more specific search can be performed by specifying which fields should be searched. For instance, "warfarin[Title/Abstract]" indicates that the term "warfarin" should be searched only from titles and abstracts. Besides, a more narrow search can be conducted using search filters such as "Clinical Queries" and "Core Clinical Journals" filters in PubMed.

Taxonomy (SORT) (Ebell et al., 2004). A body of evidence could be assigned into three grades based on the quality and consistency of available evidence (**Table 2-2**).

For clinical recommendations regarding treatment, prevention or screening, the quality of POEs from a clinical study can be determined as indicated in **Table 2-3** using Level of Evidence (LoE). As exemplified by Ebell (2005), vitamin E was found in some case-control studies (LoE = 2) to slow functional decline for patients with Alzheimer's disease, but good quality randomized control trials (LoE = 1) have not confirmed this benefit. The greater the level of evidence, the greater the grade of recommendation. Therefore, the intake of vitamin E should be recommended based on the randomized controlled trials but not the case-control studies. The example explains the importance of considering the study design when evaluating the quality of evidence for clinical decision-making.

Table 2-2. Strength-of-Recommendation (SoR) grades

SoR	Definition[t]
A	Recommendation based on high-quality and consistent POEs.
B	Recommendation based on limited-quality and inconsistent POEs.
C	Recommendations based on consensus usual practice, opinion, DOEs or cases series/reports.

[t] POE = "Patient-Oriented Evidence", DOE = "Disease-Oriented Evidence

Table 2-3. Level of Patient-Oriented Evidence (LoE)

Study Quality	Study Design
Level 1	Systematic review, meta-analysis and randomized controlled trial with high quality and consistent findings.
Level 2	Lower quality clinical trial, cohort study and case-control study with lower quality and inconsistent findings.

Clinical Query Filters.

Yu and Cao (2008) categorized 4654 clinical questions maintained by the National Library of Medicine (NLM) into 12 general topics. Some of the most commonly asked topics have been studied extensively by the Hedges Study Group to develop clinical query filters. The filters provide broad (sensitive) and narrow (specific) search of five categories of clinical studies and systematic reviews from the MEDLINE

database. The five categories include: etiology, diagnosis, therapy, prognosis and clinical predication guides (Wong et al., 2003; Wilczynski et al., 2003; Haynes and Wilczynski, 2004; Wilczynski and Haynes, 2004; Haynes et al., 2005; Montori et al., 2005). The following shows the difference between a "broad" and a "narrow" clinical query filters for therapy studies:

- Therapy/Broad*:*
  ((clinical[Title/Abstract] AND trial[Title/Abstract]) OR clinical trials as topic[MeSH Terms] OR clinical trial[Publication Type] OR random[Title/Abstract] OR random allocation[MeSH Terms] OR therapeutic use[MeSH Subheading])*
- Therapy/Narrow:
  (randomized controlled trial[Publication Type] OR (randomized[Title/Abstract] AND controlled [Title/Abstract] AND trial[Title/Abstract]))

The therapy/narrow filter specifies a search for articles reporting randomized controlled trials, or for those that contain the words *randomized AND controlled AND trial* in the titles or abstracts; a more sensitive search can be achieved using the therapy/broad filter to return a higher number of studies about interventions or therapies. On the other hand, the "systematic reviews" clinical query allows the search of systematic reviews and meta-analyses (Montori et al., 2005), which are the highest quality research papers. The use of clinical query filters is intended to retrieve citations related to specific clinical research areas and to avoid information overload.

2.1.3. Summary

Section 2.1 focuses on strategies to find the best available evidence in the literature for the practice of EBM. The main strategies include:

- Converting an information need into a well-focused question or searchable keywords using the PICO framework,
- Focusing on journals particularly relevant to practicing physicians or journals that publish articles addressing outcomes that matter to patients, and
- Determining the level of evidence by study design (randomized controlled trials, case-control studies, etc.) or filtering for specific studies based on the type of question (diagnosis, therapy, etc.)

2.2. MEDICAL QUESTION ANSWERING SYSTEMS

Current MedQA systems focus on providing direct and precise answers to a user's question by employing natural language processing techniques for the automatic extraction of structured information. In this section, a brief introduction of the general architecture of a QA system is provided in **Section 2.2.1**. To identify high-quality and evidence-based information, the approaches and resources that have been used to develop a number of MedQA systems are discussed in **Section 2.2.2**.

2.2.1. General Architecture of QA Systems

As illustrated in **Figure 2-1**, a QA system is based mainly on three phases: question processing, document processing and answer processing.

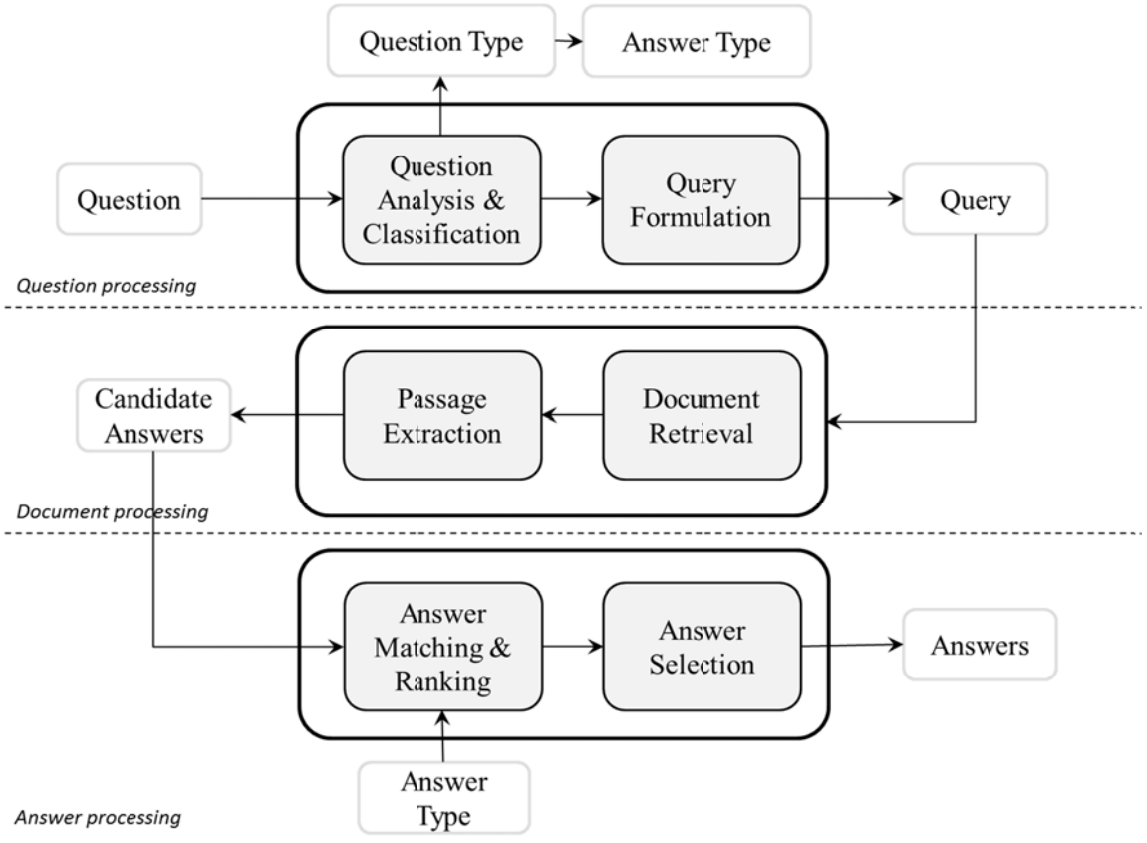


Figure 2-1. The main processing phases of a QA system.

In the question processing phase, a natural language question is generally input to a QA system. The question processing phase performs two steps: question analysis and classification, and query formulation. A natural language question is analyzed to

determine the type of question and the expected type of answer. The output of this phase is a query in canonical form, which serves as the input to a document retrieval engine.

In the document processing phase, the query is submitted to a Web-based or a Corpus-based search engine to retrieve relevant documents. The document retrieved can be narrowed down to the most relevant documents using various document filtering techniques. Candidate answer passages are then extracted from the most relevant documents using various entity recognizers and semantic relatedness techniques.

In the answer processing phase, the candidate answer passages act as the input and are matched with the expected type of answer from the question processing phase. A score is assigned to each candidate document based on its relevant to a query, and same as in the previous two phases, more complex natural language and linguistic processing techniques may be involved in this phase. The output of a QA system is generally displayed as a ranked list of documents.

2.2.2. State-of-the-Art of MedQA systems

Question Processing.

An ideal QA system is expected to be capable of accepting a variety of natural language questions. A recent review by Athenikos and Han (2010) concluded that current MedQA systems are limited by their ability to process only certain types and formats of questions. The Demner-Fushman et al.'s InfoBot system (2008) accepts only PICO-format queries. An example of the PICO query is "*Atrial Fibrillation AND Warfarin AND Aspirin AND Secondary Stroke*". The use of the system may be limited by the ability of users to apply Boolean operators (such as AND and OR). Similar to the Niu et al.'s EpoCare system (2003; 2004), CQA-1.0 (the later version of the InfoBot system) requires users to clearly identify each component of PICO as the input query. A clear understanding of the PICO framework and the terminology of a specialized domain are required to pose a question to the systems. Besides, the medical concepts in the input query must be searchable by the PICO-based systems. For example, the question "*Are high-potency topical corticosteroids more effective than low-potency steroids for alopecia areata in children?*" is broken down and entered into CQA-1.0. As demonstrated in **Figure 2-2**, a more precise description of the "intervention" and "comparison" using "*high-potency topical corticosteroid*" and "*low-potency steroids*"

respectively retrieved no results, whereas using the search terms "*topical corticosteroid*" and "*steroids*" resulted in a maximum of 20 documents. This suggests that users may have problems in describing their information needs using terms or phrases that would be recognized by the PICO-based systems as appropriate vocabulary. On the other hand, the Yu et al.'s AskHERMES system (2007) accepts both well- and poorly-formulated definitional questions. The system also provides over 10 thousand questions with answers that can be searched by category and keyword. In response to a poorly-formulated question, the system attempts to assist users in clarifying their search requests by returning a list of "related questions" from the database. An example is given in **Figure 2-3** by submitting the question "*What is the best treatment for needle stick injury?*" to the AskHERMES system.

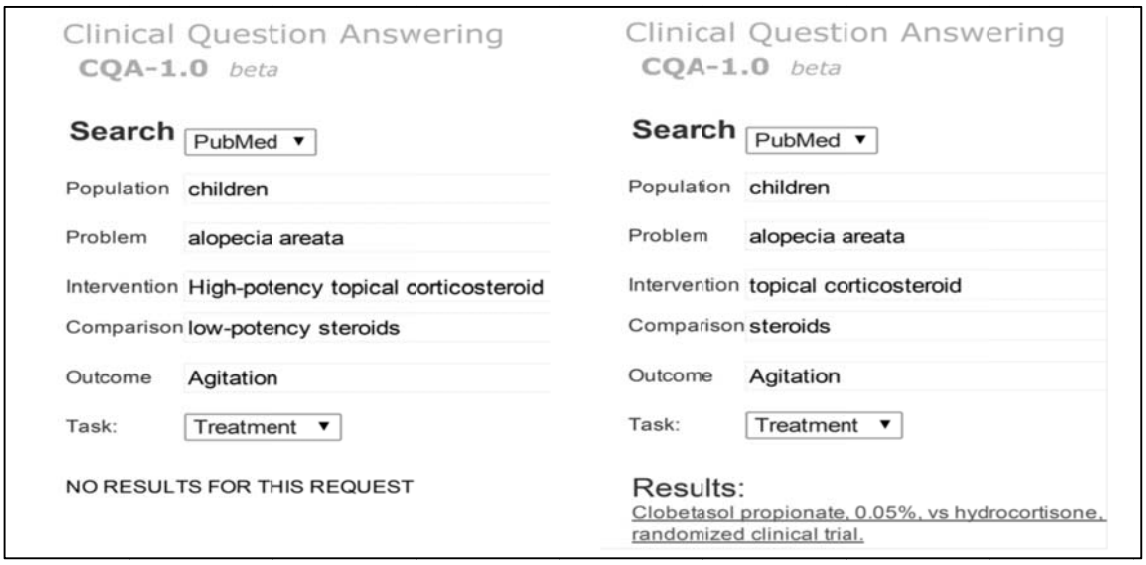


Figure 2-2. Two examples of retrieval results obtained using CQA-1.0.

Figure 2-3. An example of "Related Questions" returned by AskHERMES.

In the question processing phase, a question is processed to identify key query terms. The query terms are used to formulate a search query in canonical form, which is then used as the input of a document retrieval engine. The Delbecque et al.'s (2005),

Niu et al.'s (2006), Demner-Fushman et al.'s (2006) and Weiming et al.'s (2007) QA systems extract UMLS[4] semantic concepts from a natural language question or a PICO-format query as query terms. Yu et al.' QA system generates query terms by identifying and weighting noun phrases in a natural language question. The original query terms are then expanded using different terminological/ontological resources such as UMLS and MeSH[5] for synonymous and related terms. For example, the term "breast cancer" is expanded using MeSH in PubMed as follows to include all possible term forms in the search query. The original query terms and the expanded terms are used to retrieve relevant documents.

> *("breast neoplasms"[MeSH Terms] OR ("breast"[All Fields] AND "neoplasms"[All Fields]) OR "breast neoplasms"[All Fields] OR ("breast"[All Fields] AND "cancer"[All Fields]) OR "breast cancer"[All Fields]*

---

[4] Unified Medical Language System (UMLS) is developed and maintained by the US National Library of Medicine (NLM). It is the largest biomedical terminology system, and is freely available. It is intended to be used for developing computer systems capable of understanding the specialized vocabulary used in biomedicine and health care. The UMLS knowledge resources: Metathesaurus and Semantic Network were used to map biomedical text to UMLS concepts. The 2015 version of the Metathesaurus contains more than 3.1 million concepts and 12 million unique concepts names from over 170 source vocabularies. Examples of source vocabularies are SNOMED CT, LOINC, MeSH and ICD-9CM. In the Metathesaurus, synonymous terms from different source vocabularies are clustered into a single "concept", and are given the same concept unique identifier (CUI). The Semantic Network, on the other hand, is a limited network of 135 Semantic Types and 54 Semantic Relations. It is designed to reduce the complexity of the UMLS. Each Metathesaurus concept is categorized under one or more Semantic Types from the Semantic Network. The 135 Semantic Types are further categorized into 15 Semantic Groups such as "Chemicals and Drugs", "Disorders" and "Gene & Molecular Sequences". The 2013AB version of the UMLS was used to obtain the results described in this thesis.

[5] Medical Subject Headings (MeSH) is the controlled vocabulary created and maintained by the National Library of Medicine (NLM) for indexing MEDLINE citations. Each citation is manually assigned a number of MeSH terms that describe the topics discussed in an article. MeSH consists of sets of term naming descriptors that are arranged alphabetically and hierarchically to allow searching at various levels of specificity. The 2008 version of MeSH has 27,455 descriptors organized in a twelve-level hierarchy of headings. Broad headings such as "Diseases" and "Chemicals and Drugs" are found at the most general level and more specific headings such as "Arbovirus Infections" and "Benzoquinones" are found at more narrow levels of the hierarchy. MeSH also has more than 220,000 entry terms that help find the most appropriate search terms. For example, the entry terms of "Mitomycin" include "Ametycine", "Mitocin-C", "Mitomycin-C" and "Mutamycin". In addition, there are over 224,000 headings called Supplementary Concept Records that account for the large volume of chemical names found in biomedical literature.

Much effort has been put on identifying and expanding query terms to improve the retrieval of relevant documents. However, previous study demonstrated a lack of key medical concepts that comprise a well-formed query in natural language questions posed by physicians (Booth et al., 2000; Huang et al., 2006). More research needs to be done to enable more complicated analysis of poorly-formulated questions. For instance, the question "*What is the best treatment for acute otorrhea?*" contains only the [P] element ("acute otorrhea"). The [I], [C] and [O] elements are not defined in the question, reflecting that a user has a vague information need. Besides expanding a search query to include synonyms related to the [P] element, a QA system should allow a user to refine a search without having to re-enter the search criteria, such as by providing the PICO elements that are related to the initial search query to the users.

Document Processing.

The search engines used for document retrieval are either Web-based (e.g. Google) or Corpus-based (e.g. PubMed). Delbecque et al.'s (2005) and Niu et al.'s (2006) use Google and the Toronto XML (ToX) search engines respectively to retrieve relevant documents. Demner-Fushman et al. (2006) use domain-specific search engine, PubMed, to retrieve medical literature from the MEDLINE database. Weiming et al.'s (2007) use Lucene, a standard information retrieval engine, to retrieve documents from the Web and from the MEDLINE database. Yu and Kaufman (2007) recommend the use of both Web-based and Corpus-based search engines for document retrievals. Besides, there have been a few studies comparing the use of Google Scholar and PubMed for literature searches. Compared to Google Scholar, PubMed provides more powerful tools (such as MeSH and Clinical Query Filters) for users to perform a more efficient search of relevant documents (Henderson, 2005; Anders and Evans, 2010; Bramer et al., 2013). In addition, PubMed remains the most widely used resource by physicians for systematic reviews and original clinical articles (Agoritsas et al., 2012; Shariff et al., 2013). In this regard, it is worthwhile to continue to use PubMed for the retrieval of relevant documents from the MEDLINE database.

The second step of the document processing phase is the extraction of relevant passages. The purpose is to allow an information retrieval system to precisely point out the most relevant parts of a document or to filter out irrelevant documents. Different natural language processing techniques have been used to extract relevant passages. Delbecque et al.'s (2005) identify medically relevant named entities in candidate

documents using the UMLS semantic types. Similarly, Niu et al. (2006) and Demner-Fushman et al. (2006) focus on identifying the semantic roles that correspond to the four fields of PICO frame in both question and candidate documents. Weiming et al. (2007) investigate the relations between question and candidate documents using noun keywords and the UMLS concept mapping rules. A review of the four QA systems shows that both the question processing and document processing phases involve the use of UMLS as a knowledge resource for query formulation and semantic tagging and annotation of candidate documents.

Answer Processing.

In this phase, answers are generated by matching query from the question processing phase with the annotated sentences from the document processing phase. The candidate answers are then ranked based on their matching scores. Answers are generated by providing context from multiple highest-ranked articles using semantic clustering and summarization techniques (Niu et al., 2006; Demner-Fushman and Lin, 2007; Weiming et al., 2007). Delbecque et al. (2005) quantify the co-occurrence of semantic types in candidate documents and select tagged clauses as answers. An ideal answer from a MedQA system should point out the similarities and differences between multiple clinical studies, and integrate the necessary information to generate synthesized answers. This can be achieved by extracting answers from systematic reviews that synthesize information across multiple studies, which however can be limited by the number of systematic reviews available. In current semantic MedQA systems, multiple candidate answers arrive at the same score cannot be compared and analyzed statistically for combination of findings. Similarly, multiple candidate answers disagree on a particular query cannot be compared for differences between findings. In this regard, more research needs to be done for appropriate way to synthesize evidence from multiple primary studies and for a more appropriate presentation of answers.

2.2.3. Summary

Section 2.2 reviewed the approaches and resources that have been used to develop the current MedQA systems. A comparison of four MedQA systems is given in **Table 2-4.** The key findings of the review are summarized as follows:

- Current systems accept multiple types of input, which include PICO-format queries, Boolean search queries and definitional questions. To improve the retrieval of relevant documents, the input is processed by the systems to determine effective query terms and to generate query expansion terms using vocabulary resources such as the UMLS and the MeSH thesaurus.
- Current systems use Google, ToX, PubMed, Lucene, or a combination of the two search engines for document retrieval. Among these, PubMed is the most widely used search engine to retrieve documents from the MEDLINE database. To filter out documents irrelevant to a user's query, most of the current systems utilize the UMLS Metathesaurus for the identification of named entities in candidate document or for the semantic annotation of candidate documents.
- Current systems focus mainly on returning a ranked list of relevant documents. This is achieved most commonly by matching terms in a search query with those in annotated documents. A matching score is given to each document and an answer is generated by providing context from the document using clustering-based text summarization techniques.

Table 2-4. A comparison of four semantic-based QA systems

First Author	***Delbecque (2005)***	***Niu (2003; 2006)***	***Demner-Fushman (2006; 2007)***	***Weiming (2007)***
Query Formulation	Semantic concepts and relations	PICO framed	PICO framed	Semantic concepts and relations
Document Retrieval	Google	ToX engine	MEDLINE	Lucene
Passage Extraction	UMLS concepts tagging, and semantic types and relations for named entity recognition	PICO roles in medical text, and semantic classes and relations	PICO extractors and annotation of text	Noun keywords and UMLS concept mapping rules
Answer Matching & Ranking	Co-occurrence of semantic types	Match query with annotated sentences	Match query with annotated sentences	Match query with annotated sentences
Answer Selection	Semantic relations for selecting tagged clauses as answers	Semantic clustering and summarization	Semantic clustering and summarization	Semantic clustering and summarization

2.3. INFORMATION SEARCH SUPPORT

This section describes and discusses the search support offered by two freely accessible MedQA systems: the CQA-1.0 and the AskHERMES systems, from the process of converting an information need into a well-focused question, to the process of identifying documents that provide the most useful information for clinical practice.

2.3.1. The CQA-1.0 System

The homepage of CQA-1.0 (available at: http://archive.nlm.nih.gov/ridem/cqa.html) provides an interface that requires users to break down their information needs into four components of the PICO framework and is designed to answer complex clinical questions (Demner-Fushman and Lin, 2007). As shown in **Figure 2-4**, two search engines, Essie and PubMed are provided by the system. The search results can be limited to articles from human studies, and to those published with abstracts and written in English. Besides, a more focused search can be achieved by selecting a specific clinical task (such as treatment, prevention or prognosis), or by retrieving articles from one of the following subsets: core clinical journals, nursing journals, systematic reviews, toxicology and Cochrane reviews. The filtering options in CQA-1.0 allow users to limit a search to a specific clinical research area, to a subset of journals, and to a particular type of publication.

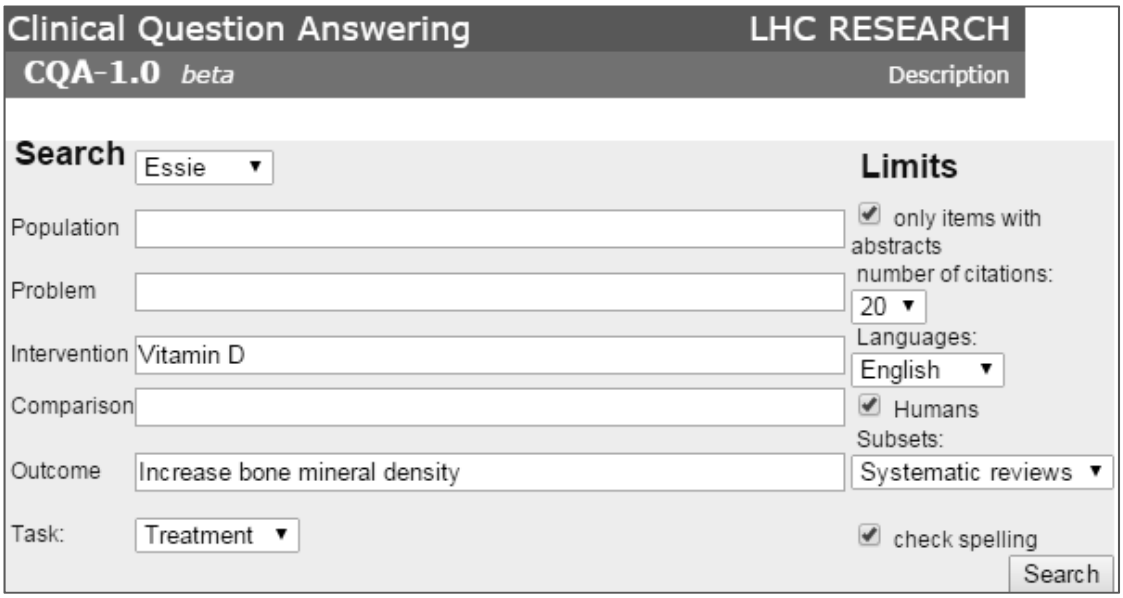
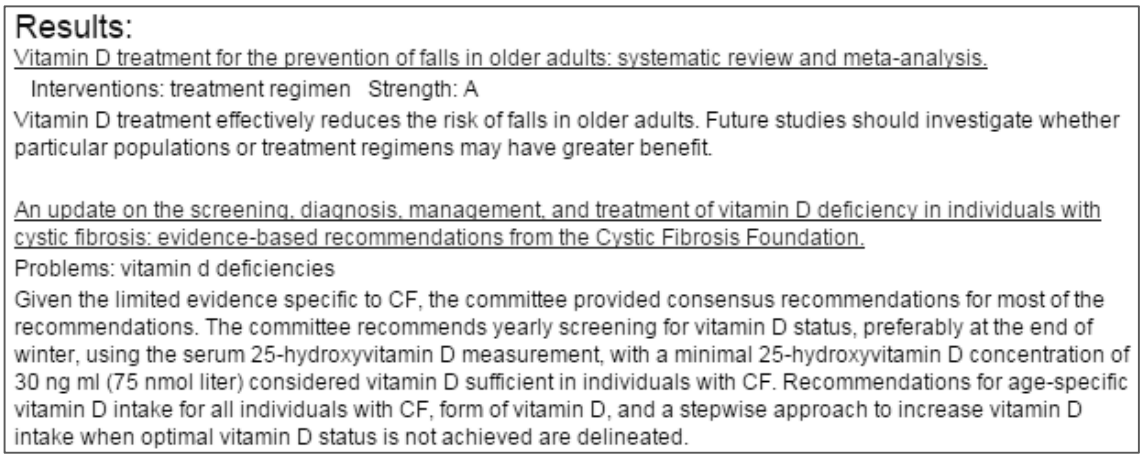


Figure 2-4. Posing a question to CQA-1.0.

A maximum of 20 top-ranked answers are returned by the system in response to an input query. The answers are presented with the relevant PICO elements and the strength of recommendation of A to C, in order to assist users in quickly locating answers to their questions, and in searching the best available evidence. However, this search support function is not consistently applied to all the answers. As seen in **Figure 2-5**, the first answer is supplemented with the relevant [I] element ("treatment regimen") and the strength of recommendation ("Strength: A"), whereas the second answer is presented along with the [P] element ("vitamin d deficiencies") only. A clear understanding of the PICO framework and the terminology of a specialized domain are required to pose a question to the system. The users however may not be able to express their information needs in the vocabulary used in relevant information resources or in the manner expected by the system. If this is the case, the consequence is poor search results.


Results:

Vitamin D treatment for the prevention of falls in older adults: systematic review and meta-analysis.

Interventions: treatment regimen Strength: A

Vitamin D treatment effectively reduces the risk of falls in older adults. Future studies should investigate whether particular populations or treatment regimens may have greater benefit.

An update on the screening, diagnosis, management, and treatment of vitamin D deficiency in individuals with cystic fibrosis: evidence-based recommendations from the Cystic Fibrosis Foundation.

Problems: vitamin d deficiencies

Given the limited evidence specific to CF, the committee provided consensus recommendations for most of the recommendations. The committee recommends yearly screening for vitamin D status, preferably at the end of winter, using the serum 25-hydroxyvitamin D measurement, with a minimal 25-hydroxyvitamin D concentration of 30 ng ml (75 nmol liter) considered vitamin D sufficient in individuals with CF. Recommendations for age-specific vitamin D intake for all individuals with CF, form of vitamin D, and a stepwise approach to increase vitamin D intake when optimal vitamin D status is not achieved are delineated.

Figure 2-5. An example of answers generated by CQA-1.0.

2.3.2. The AskHERMES System

The homepage of askHermes (available at: http://www.AskHERMES.org/) provides a simple and clean interface for the submission of question. The system processes both well-defined and ill-formulated questions (Cao et al., 2011). At the top of the result page (**Figure 2-6**) are links to several clinical question answering tools, which include utilities to browse questions by category and keyword, to classify questions into the top five most frequent question categories (such as "diagnosis" and "treatment and prevention"), and to generate query terms from *ad hoc* questions and

then apply the terms for information retrieval. The utilities aim to assist users in understanding how a question is answered by the system.

In response to a question, short passages extracted from the MEDLINE abstracts are presented as answers, with the query terms from *ad hoc* questions formatted in bold. Three different arrangements of answers are presented by the system. Clustered answers are grouped based on different combinations of query terms and expanded query terms from the UMLS Metathesaurus. Topic labels are assigned to each cluster to enable users to easily locate information of interest. The system also provides a ranked list of answers. Classified answers are grouped according to the common labels appear in answer passages. The system allows users to perform a search based on the presentation of answers that they prefer.

Figure 2-6. The result page of AskHermes.

Compared to CQA-1.0, AskHERMES provides a more complicated result page for the search of clinical evidence. Besides, according to a study by Bauer and Berleant (2012), the system returns passages that could potentially answer all types of questions, causing the retrieval of high number of results. This may in turn result in information overload, which is one of the main obstacles that prevents physicians from answering patient-care questions. CQA-1.0, on the other hand, assumes that users have a clear understanding of their search targets and are able to convert their information needs into searchable PICO queries. In response to a poorly-formulated question, CQA-1.0 does not assist users in refining their search, while a list of “related questions” is displayed by AskHERMES in attempt to satisfy users’ information needs.

2.3.3. Summary

Combining the findings from **Section 2.2** and **Section 2.3**, the problems that users may encounter when performing a search task using the current MedQA systems are summarized as follows:

- Inability to formulate a well-focused question due to a lack of terminology of a specialized domain or a lack of knowledge of a new area of interest,
- Inability to describe an information need using terms and phrases that would be recognized by a system as appropriate vocabulary,
- Inability to break down an information need into searchable keywords in order to fit the question framework used by a system, and
- Inability to use advanced query syntax such as Boolean operators when formulating a search query.

2.4. CONCLUSION

Multiple literature search strategies have been developed to support physicians in finding the best available clinical evidence for the practice of EBM. MedQA systems are designed to allow users to quickly identify the most useful clinical information with minimal effort. Most of the current MedQA systems assume that users have clear information needs, have sufficient knowledge of a subject domain, and have the ability to formulate answerable questions using appropriate vocabulary when performing a search task. There is a lack of studies that focus on assisting users in clarifying and recognizing their information needs by promoting the interaction between users and a MedQA system.

3. CHAPTER III: Thesis Proposal

When searchers have a clear understanding of the information they are looking for, converting an information need into a well-focused question is a fairly simple task. But when searchers have vague information needs or are unfamiliar with a subject domain, they encounter difficulties articulating their information needs and translating them into well-focused questions. A well-focused question warrants a high quality answer. In contrast, when a question is poorly-formulated, a search task can be very difficult and time consuming as the range of covered topics becomes larger. This chapter is divided into two sections.

- **Section 3.1** explains how the proposed framework can assist users in clarifying and recognizing their information needs, and
- **Section 3.2** gives a brief description of the architecture of the clinical question answering engine proposed in this thesis.

3.1. PROPOSED FRAMEWORK

In response to a user's query, document clustering can be used to organize a collection of documents into a number of meaningful clusters. Clustering approach has been shown to be more effective than traditional ranked list approach for interactive information retrieval (Leuski, 2001; Leuski and Allan, 2004; Zhu et al., 2008). In terms of information seeking, document clustering allows users to quickly locate related documents or to filter out irrelevant documents (Zhao and Karypis, 2002; Punitha et al., 2011). Besides, when a query is vague or ill-defined, the clustering approach can help focus a search to a specific cluster of documents (Eaton and Zhao, 2001; Lechtenfeld and Fuhr, 2012). In short, document clustering has the potential to support the search of relevant documents, especially when users have difficulties expressing precisely their information needs.

A two-stage approach was proposed to improve the process of searching the most relevant documents to answer clinical questions: the **exploratory stage** and the **concept stage**. As illustrated in **Figure 3-1**, in response to a clinical question, users are allowed to explore a particular subject domain using a hierarchy of medical interventions displayed in the user interface of the proposed clinical question answering engine during the exploratory stage. By selecting a cluster of interest, a list of relevant documents presented along with the most useful and important medical concepts (i.e.

the [P-O] and [I/C] elements) are returned to the users in the concept stage. It was hypothesized that:

- Hypothesis 1: "*A hierarchical structure of medical interventions can assist users in narrowing down and better understanding their search intent*", and
- Hypothesis 2: "*The visualization of PICO elements can facilitate the recognition of relevant documents that best answer an information need.*"

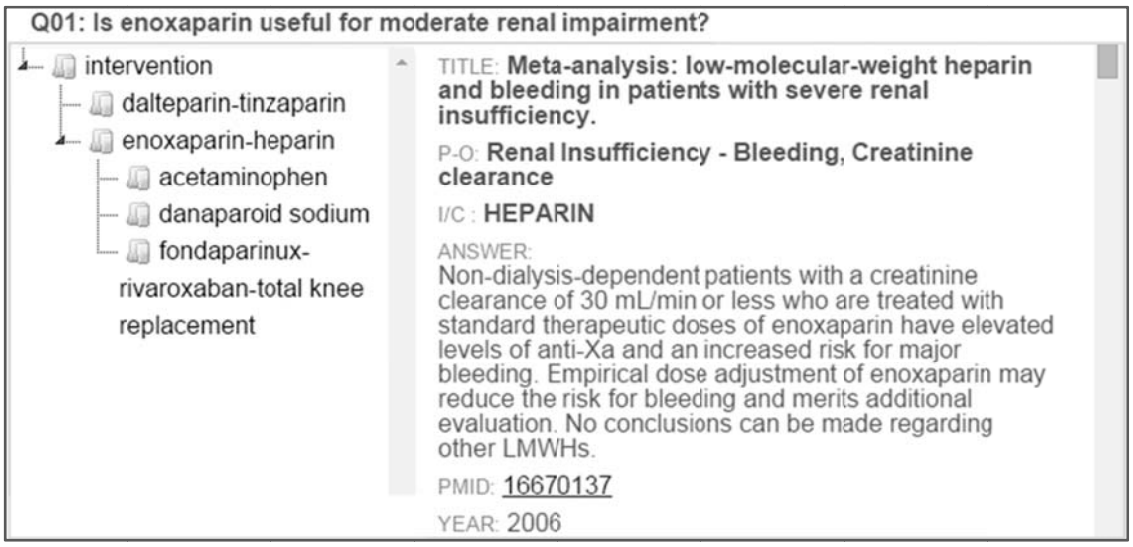


Figure 3-1. The proposed user interface

The two-stage approach is further explained as followed using **Figure 3-2**:

Exploratory Stage.

The purpose of this stage is to support and assist users in meeting their information needs. Two main steps are involved in this stage. Firstly, a collection of documents that match an information need are grouped into different clusters based on the similarity of medical interventions, which are the [I] and [C] elements of the PICO framework, between documents. Secondly, the clusters are organized and visualized as a hierarchical structure of medical interventions. The aims are to allow users to:

- Gain a better understanding of the terminology, concepts or topics related to a particular domain of interest, and
- Narrow down, refine or clarify their search intent by browsing, exploring and searching a collection of documents.

Concept Stage.

The purpose of this stage is to allow users in quickly identifying documents that best described their information needs. The hierarchy of interventions generated from the exploratory stage acts as a mediator to support the concept stage. Each document in the answer field is presented along with the most useful and important PICO elements. The aim is to facilitate the search and identification of documents that best match an information need.

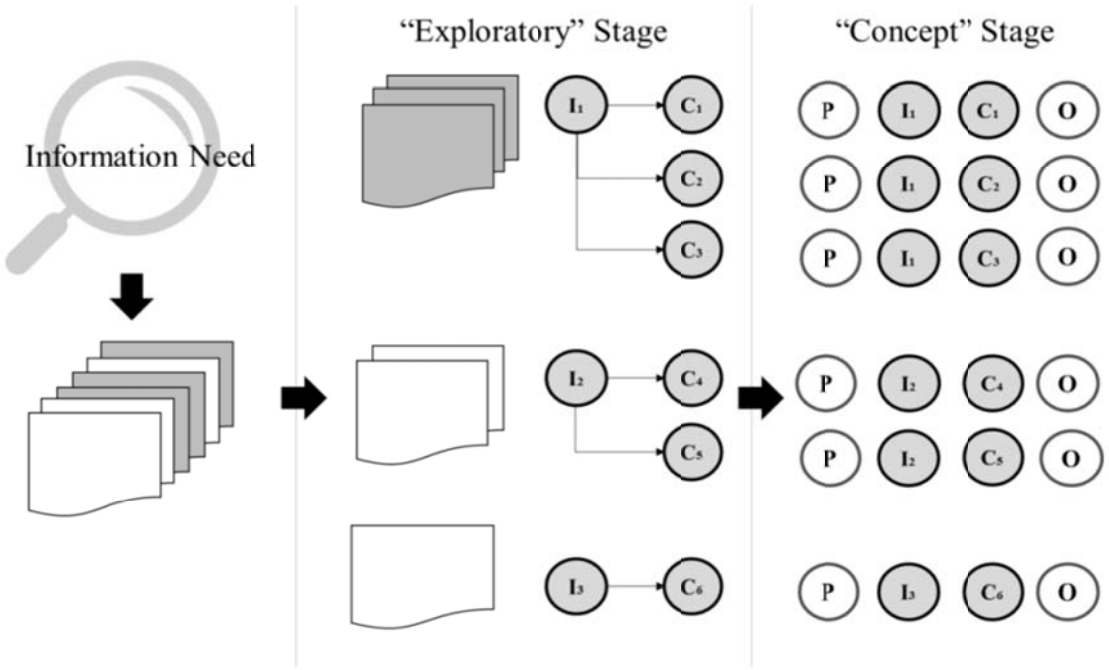


Figure 3-2. The two stages of the proposed solution.

In summary, the proposed framework aims to improve the search of the most relevant documents by offering support and assistance to the users during the information search process. It is expected that users can gain a better understanding of a problem domain, clarify or refine their search interest, and recognize their information needs through the interaction with the proposed clinical question answering engine.

3.2. ARCHITECTURE OF THE PROPOSED ENGINE

A pilot study was conducted to test the feasibility of the proposed framework (Vong and Then, 2014). The study showed that the proposed concept similarity clustering approach has the potential to assist users in clarifying or refining a vague information need. The architecture of the proposed clinical question-answering engine, based on the pilot study, is depicted in **Figures 3-3** to **3-5**, and is described briefly as follows:

Question Processing.

A clinical question in natural language is submitted to the proposed clinical question answering engine. The question processing phase consists of knowledge extraction and query formulation. The processing of the question identifies medical concepts that represent the four elements of the PICO framework. The identified elements are then used to construct a search query.

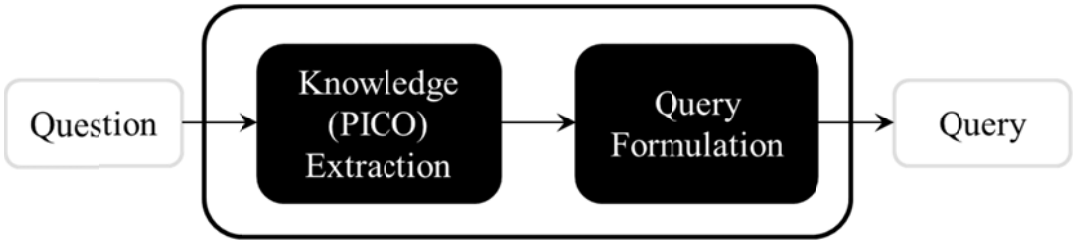


Figure 3-3. Question processing phase of the proposed engine.

Document Processing.

In the document processing phase, the search query is entered into a search engine to retrieve relevant documents from a medical literature database such as MEDLINE via PubMed. The clinical query filters are applied to improve the retrieval of therapy studies, particularly randomized controlled trials and systematic reviews/meta-analyses. PICO elements and candidate answers are then extracted from candidate passages, which are different fields such as the titles, abstracts and MeSH terms of candidate documents.

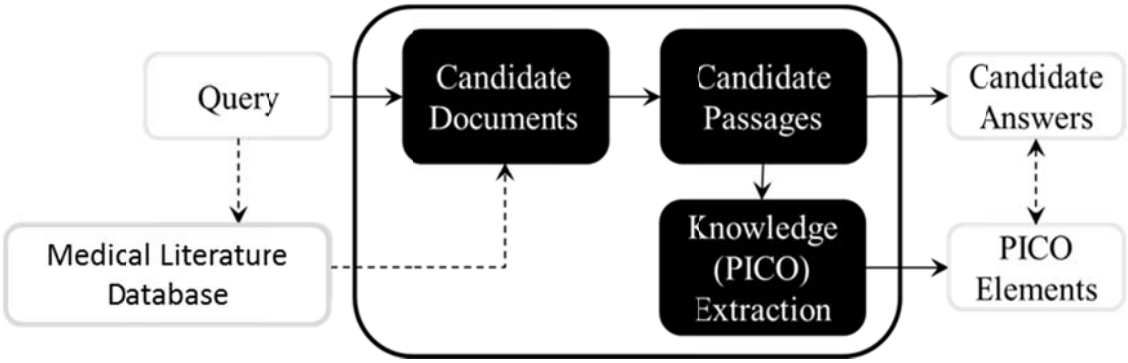


Figure 3-4. Document processing phase of the proposed engine.

Answer Processing.

Using the PICO elements extracted from the document processing phase, a hierarchy of medical interventions is constructed and displayed in the user interface. Each cluster of the hierarchy contains documents with similar medical interventions and is labeled with the therapy topic that appears the most frequent among the documents.

By selecting a cluster of interest from the hierarchy, a ranked list of answers along with their associated PICO elements is presented in the answer field of the user interface. The answers are ranked so that the most recent studies published in core clinical journals and with the highest quality study design appear at the top position of the result list.

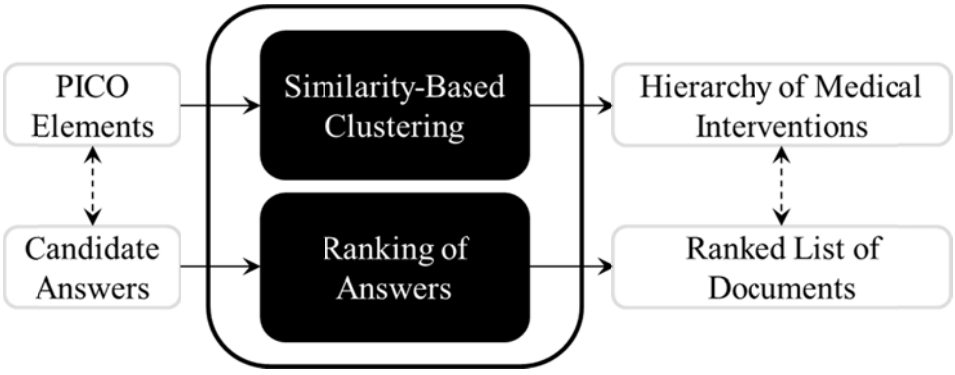


Figure 3-5. Answer processing phase of the proposed engine.

Table 3-1. Three processing phases of the proposed engine

Processing Phase		**Strategy**
Question Processing	Query Formulation	Knowledge (PICO) Extraction
Document Processing	Document Retrieval	MEDLINE via PubMed & Clinical Query Filter
	Passage Extraction	Knowledge (PICO) Extraction from Text Fields
Answer Processing	Answer Matching & Selection	Concept-based Similarity & Agglomerative Hierarchical Clustering
	Answer Ranking	Strength of Evidence

3.3. OUTLINE of the FOLLOWING CHAPTERS

A detailed description of the proposed framework is provided in the following chapters of this thesis.

- **Chapter 4** is related to the question processing and the document processing phases, which includes the extraction of effective search terms for the retrieval of relevant documents and the extraction of PICO elements for concept-based inter-document similarity analysis.
- **Chapter 5** is related to the answer processing phase, which contains information about how a collection of documents is clustered and visualized as a hierarchy of medical interventions to support the information search process.
- **Chapter 6** evaluates and compares the performance of the proposed clinical question answering engine with three existing search engines in ranking highly-relevant and evidence-based documents at higher positions in the lists of search results.
- **Chapter 7** investigates the usability and user satisfaction with the proposed clinical question answering engine, in terms of its capability in improving the information search process.

4. CHAPTER IV: Inter-Document Similarity Analysis

It was proposed that, in response to a therapy question: (i) a hierarchy of medical interventions can support users in narrowing down and better understanding their search intent and (ii) the visualization of PICO elements can facilitate the search of the most relevant documents. In this chapter, using a set of 10 therapy questions, PICO elements were extracted from different text fields such as the titles or abstracts of the resulting MEDLINE documents. Each document was then converted into a bag of weighted medical interventions. The similarity between two bags of interventions was computed using 42 similarity and distance metrics, and was compared to the similarity ratings provided by human experts using a series of statistical separation tests. The objectives of this chapter are:

- To determine the most appropriate text field of MEDLINE documents for the extraction of PICO elements, and
- To identify the most optimal combination of weighting scheme and similarity/distance metric for concept-based similarity measurement between documents.

Section 4.1 presents the methodologies used for inter-document similarity analysis. The results presented in this chapter are preliminary and serve to guide the cluster structure analysis in the following chapter. Despite this, the results of the separation tests, presented and discussed in **Section 4.2**, indicate that the 10 therapy questions are sufficient to achieve the objectives of this chapter. The top four similarity metrics were further compared and evaluated for their performance in clustering similar documents using a collection of 100 therapy questions in **Chapter 5**.

4.1. METHODOLOGY

4.1.1. Collection of MEDLINE Documents

10 therapy questions posed by clinicians at the point of care were selected randomly from the NLM and processed as described in **Section 4.1.3** to derive medical

concepts. The questions[6] are listed below with the medical concepts formatted in *Italics* (4 of Pattern I, 3 of Pattern II, 1 of each of Patterns III-V, as described in **Table 2-1**):

1. Is *Vitamin E* useful for the treatment or prevention of *Alzheimer's disease*?
2. Are *leukotriene inhibitors* effective for *allergic rhinitis*?
3. What are the indications for doing a *thrombectomy* or using *thrombolytics* for a patient with a *deep vein thrombosis*?
4. What is the treatment for *hyperthyroidism* due to *Grave's disease*?
5. What are we going to do for this *child* with *cellulitis*?
6. What are the latest recommendations for the treatment of childhood *enuresis*?
7. Does *celebrex* (*celecoxib*) or *vioxx* (*rofecoxib*) cause *heart disease* and *myocardial infarction*?
8. What drug should be used for chemical *cardioversion* of *atrial fibrillation*?
9. Is *carvedilol* better than *propranolol* for *congestive heart failure*?
10. Is the combined use of *zyba*n with *nicotine replacement* better than either one alone?

The medical concepts were used as the main search terms. **Table 4-1** shows an example of the search terms and strategies used for the retrieval of relevant documents from the MEDLINE database. Query expansion using the MeSH Metathesaurus in PubMed has been shown to improve the retrieval of relevant documents (Lu et al., 2009). Therefore, the medical concepts were allowed to be expanded in PubMed and all possible term forms were included using Boolean operators to refine the search query. The therapy/broad and systematic review clinical query filters were used to maximize the sensitivity of the search strategy. The search was limited to articles with abstract, written in English and human studies published before 16th Feb 2014.

The same search strategy was used for the ten questions. The identified articles were sorted by publication date to collect the latest studies. Users generally look for the first 10 or 20 articles retrieved by a system only (Wang et al., 2004). Therefore, to avoid information overload, for Questions 1-6, 8 and 10, the latest 50 articles were collected, and all the identified articles were collected for Questions 7 and 9 (**Table 4-2**). Overall, a total of 458 MEDLINE articles were collected.

[6] The questions were maintained by the NLM and can be collected from the AskHERMES system available at: http://www.askhermes.org/qaseam/NlmquestionList.seam.

Table 4-1. Search terms and search strategies used for Question 1.

Criteria	Search Strategy
Database	MEDLINE
Search Term	Vitamin E ; Alzheimer's Disease *("vitamin e"[MeSH Terms] OR "vitamin e"[All Fields]) AND ("alzheimer disease"[MeSH Terms] OR ("alzheimer"[All Fields] AND "disease"[All Fields]) OR "alzheimer disease"[All F0ields] OR ("alzheimer's"[All Fields] AND "disease"[All Fields]) OR "alzheimer's disease"[All Fields])*[t]
Search filter	Therapy/Broad[filter] AND systematic[sb]
Text Availability	Abstract
Species	Humans
Language	English
Publication Date	Before 16th Feb 2014

[t] The *Italics* in the table show part of the search query used to retrieve relevant documents.

Table 4-2. Number of articles collected from the MEDLINE database.

Question	No. of articles retrieved	No. of articles collected
1.	119	50
2.	98	50
3.	198	50
4.	192	50
5.	471	50
6.	127	50
7.	42	42
8.	86	50
9.	16	16
10.	183	50
Total	**1532**	**458**

4.1.2. Generation of PICO Sentences

Previous studies demonstrated that the position of a sentence within an abstract is useful in determining the PICO elements that the sentence carries (Demner-Fushman and Lin, 2007; Boudin et al., 2010). Two types of abstracts were identified from the collected articles: structured abstracts with internal section headings and unstructured abstracts written in paragraph format without the headings. Both types of abstracts were cut into three segments based on the headings and the position of the sentences in the abstracts (**Table 4-3**). The segmented sentences are called in the remainder of this thesis the "**PICO sentences**".

Table 4-3. Derivation of PICO sentences.

Representation	Internal Section Heading	Position of Sentence
[P]	Introduction, Background, Objective	First 3 sentences
[I]/[C]	Method	Sentences in between the first and the last 3 sentences
[O]	Result, Conclusion	Last 3 sentences

4.1.3. Generation of PICO Elements

The ten question, the PICO sentences, and the titles, full abstracts, chemicals and MeSH terms resulting from the 458 MEDLINE documents were processed by the MetaMap Transfer[7] (MMTx) program. The purpose is to identify medical concepts semantically from the UMLS Metathesaurus (Aronson, 2001). As shown in **Figure 4-1**, the program tokenizes a sentence into separate phrases, and returns two types of mapped concepts with their concept unique identifier (CUI) numbers in *Italics* and associated semantic types in square bracket. `Meta Mapping` concepts were extracted from the MMTx outputs and processed using Rapidminer 5.2[8] to generate PICO elements (Ertek

[7] MetaMap is a program developed by the NLM to map biomedical text to UMLS Metathesaurus concepts. Two types of mapped concepts are produced by the program: Meta Candidates, which are a list of mapped concepts, and Meta Mapping, which are the highest scoring concepts from the list.

[8] RapidMiner is a code-free analytic platform for data mining, machine learning and predictive analytics. The RapidMiner Text Processing Package provides different operators to load and process non-structural textual data and to transform nonstructural data into structural forms for further analysis.

et al., 2013). Concepts with semantic types listed in **Table 4-4** were recognized as PICO elements whereas those with other semantic types were excluded. Duplicate terms, synonyms and stopwords were removed by identifying their CUI numbers. For instance, "blood sugar" and "blood glucose" are synonyms with the same CUI number (i.e. C0005802). Examples of stopwords are "find", "release", "peer support", "still", "little" and "inform". PICO elements extracted from the ten questions were used to build the search queries (**Section 4.1.1**). For each of the 458 documents, a set of PICO elements was collected respectively from the PICO sentences, titles, full abstracts, chemicals and MeSH terms. The aim is to identify the most appropriate source of PICO elements for the subsequent inter-document similarity (**Section 4.1.5**) and cluster structure analyses (**Chapter 5**). An example of the different fields of a MEDLINE document with PubMed unique identifier (PMID) of 23583234 is shown in **Figure 4-2**.

Processing 00000000.tx.2: The aromatase inhibitor anastrozole inhibits estrogen synthesis.

Phrase: "The aromatase inhibitor anastrozole"
Meta Candidates (4)
1.827 *C0290883*: anastrozole [Organic Chemical, Pharmacologic Substance]
2.734 *C0593802*: Aromatase inhibitor (Aromatase Inhibitors) [Pharmacologic Substance]
3.660 *C0003805*: Aromatase [Amino Acid, Peptide, or Protein, Enzyme]
4.627 *C0243077*: inhibitors [Chemical Viewed Functionally]
Meta Mapping (901)
734 *C0593802*: Aromatase inhibitor (Aromatase Inhibitors) [Pharmacologic Substance]
827 *C0290883*: anastrozole [Organic Chemical, Pharmacologic Substance]

Phrase: "inhibits"
Meta Candidates (4)
1.966 *C0311403*: Inhibited [Qualitative Concept]
2.928 *C0237477*: Arrest inhibitor (Arrested progression) [Temporal Concept]
3.928 *C0392351*: arrest (Law enforcement arrest) [Governmental or Regulatory Activity]
4.928 *C0521111*: Retarded [Qualitative Concept]
Meta Mapping (966)
966 *C0311403*: Inhibited [Qualitative Concept]

Phrase: "estrogen synthesis."
Meta Candidates (5)
1.861 *C0869032*: Synthesis [Phenomenon or Process]
2.694 *C0014939*: Estrogen (Estrogens) [Hormone, Pharmacologic Substance, Steroid]
3.623 *C0720298*: Estrogenic [Hormone, Pharmacologic Substance, Steroid]
4.594 *C0014949*: Estrus [Organism Function]
5.594 *C0323166*: Oestrus [Invertebrate]
Meta Mapping (888)
694 *C0014939*: Estrogen (Estrogens) [Hormone, Pharmacologic Substance, Steroid]
861 *C0869032*: Synthesis [Phenomenon or Process]

Figure 4-1. An example of MMTx output.

TI - Vitamin E and memantine in Alzheimer's disease: clinical trial methods and baseline data

AB - BACKGROUND: Alzheimer's disease (AD) has been associated with both oxidative stress and excessive glutamate activity. A clinical trial was designed to compare the effectiveness of (i) alpha-tocopherol, a vitamin E antioxidant; (ii)memantine (Namenda), an N-methyl-D-aspartate antagonist; (iii) their combination; and (iv) placebo in delaying clinical progression in AD. METHODS: The Veterans Affairs Cooperative Studies Program initiated a multicentre, randomized, double-blind, placebo-controlled trial in August 2007, with enrolment through March 2012 and follow-up continuing through September 2012. Participants with mild-to-moderate AD who were taking an acetylcholinesterase inhibitor were assigned randomly to 2000 IU/day of alpha-tocopherol, 20 mg/day memantine, 2000 IU/day alpha-tocopherol plus 20 mg/day memantine, or placebo. The primary outcome for the study is the Alzheimer's Disease Cooperative Study/Activities of Daily Living Inventory. Secondary outcome measures include the Mini-Mental State Examination; the Alzheimer's Disease Assessment Scale, cognitive portion; the Dependence Scale; the Neuropsychiatric Inventory; and the Caregiver Activity Survey. Patient follow-up ranged from 6 months to 4 years. RESULTS: A total of 613 participants were randomized. The majority of the patients were male (97%) and white (86%), with a mean age of 79 years. The mean Alzheimer's Disease Cooperative Study/Activities of Daily Living Inventory score at entry was 57 and the mean Mini-Mental State Examination score at entry was 21. CONCLUSION: This large multicentre trial will address the unanswered question of the long-term safety and effectiveness of alpha-tocopherol, memantine, and their combination in patients with mild-to-moderate AD taking an acetylcholinesterase inhibitor. The results are expected in early 2013.

MH - Aged
Aged, 80 and over
Alzheimer Disease/*drug therapy
Antioxidants/*therapeutic use
Double-Blind Method
Excitatory Amino Acid Antagonists/*therapeutic use
Female
Humans
Longitudinal Studies
Male
Memantine/*therapeutic use
Psychiatric Status Rating Scales
Veterans
Vitamin E/*therapeutic use

RN - Antioxidants
Excitatory Amino Acid Antagonists
Vitamin E
Memantine

Figure 4-2. Different fields of a MEDLINE article with PMID of 23583234.

(TI = Title, AB = Abstract, MH = MeSH Terms, RN = Chemicals)

Table 4-4. Identification of PICO elements by semantic types.[9]

Representation	Semantic Type
[P]/[O]	Age group, Family group, Group, Human, Patient or disabled group, Population group, Acquired abnormality, Anatomical abnormality, Cell or molecular dysfunction, Congenital abnormality, Disease or syndrome, Experimental model of disease, Finding, Injury or poisoning, Mental or behavioral dysfunction, Neoplastic process, Pathologic function, Sign or symptom.
[I]/[C]	Daily or recreational activity, Amino acid, peptide, or protein, Antibiotic, Clinical drug, Eicosanoid, Enzyme, Hormone, Inorganic chemical, Lipid, Neuroreactive substance or biogenic amine, Nucleic acid, nucleoside, or nucleotide, Organic chemical, Organophosphorus compound, Pharmacologic substance, Receptor, Steroid, Vitamin, Diagnostic procedure, Therapeutic or preventive procedure.

4.1.4. Text processing of the [I] and [C] elements

The text processing was achieved using Rapidminer 5.2 and includes four steps. The purpose is to create word vectors based on the derivation of the [I] and [C] elements and the weighting schemes applied to them. The [I] and [C] elements stand for "intervention" and "comparison" respectively. Both of the elements indicate the therapeutic or preventive procedures or medications described in the original articles. Therefore, the two elements are called jointly the "**interventions**" in the remainder of this thesis.

Step 1: Extraction.

The interventions resulting from different fields of MEDLINE documents were collected (**Table 4-5**). Interventions from two or three different fields such as "Titles + Chemicals" were combined, regardless of the occurrence of identical interventions. PICO sentences were extracted from the full abstracts. Therefore, the combination of the two fields, "PICO sentences" and "Full-abstracts", were excluded from the study.

[9] The semantic types used for the identification of PICO elements were adapted from a previous work by Boundin et al. (2010).

Table 4-5. Derivation of interventions.

Code	Source				
A	PICO sentences				
B	Full abstracts				
C	MeSH terms				
D	Titles				
E	Chemicals				
F	PICO sentences	+	MeSH terms		
G	PICO sentences	+	Titles		
H	PICO sentences	+	Chemicals		
I	Full abstracts	+	MeSH terms		
J	Full abstracts	+	Titles		
K	Full abstracts	+	Chemicals		
L	Titles	+	MeSH terms		
M	Titles	+	Chemicals		
N	Titles	+	MeSH terms	+	PICO sentences
O	Titles	+	MeSH terms	+	Full abstracts
P	Titles	+	Chemicals	+	PICO sentences
Q	Titles	+	Chemicals	+	Full abstracts
R	Titles	+	Chemicals	+	MeSH terms

Step 2: Tokenization.

Multi-word interventions were tokenized by whitespace and hyphenated words were kept intact. For instance, the intervention “anti-inflammatory agents” is tokenized into “anti-inflammatory” and “agents”.

Step 3: Stemming.

The resulting words were stemmed using the Snowball algorithm (Wurst and Mierswa, 2007) in order to map different grammatical forms of a word to a common term. For instance, the words “therapy” and “therapies” are stemmed into “therapi”, and “vitamin” and “vitamins” into “vitamin”.

Step 4: Weighting.

The stemmed words were weighted using normalized term frequency (TF), binary occurrences (BO), term occurrences (TO) or term frequency-inverse document frequency (TFIDF), and were represented as word-vectors. For a term i in document j, if

f_{ij} = the number of occurrences of term i in document j,

f_{dj} = the total number of terms occurring in document j, and

f_{ti} = the number of documents in a collection that contains term i,

the weight of term i in document j, as denoted by w_{ij}, can be computed using four weighting schemes described in **Table 4-6**. Both the TF- and TFIDF-weighted word vectors were expressed in numerical form, while BO- and TO-weighted word vectors were expressed respectively in binomial and nominal forms.

Table 4-6. Four weighting schemes.

Scheme	Description	Formula
TF	The ratio of the frequency of term i in document j to the total number of terms in document j.[t]	$w_{ij} = \frac{f_{ij}}{f_{dj}}$
BO	The occurrence of term i in document j with a binary value of 0 or 1.	$w_{ij} = \begin{cases} 1, & f_{ij} > 0 \\ 0, & else \end{cases}$
TO	The absolute number of occurrence of term i in document j.	$w_{ij} = f_{ij}$
TFIDF	The frequency of term i in document j multiplies by the inverse of the number of documents in which term i appears at least once. $\|D\|$ is the total number of documents.[t]	$w_{ij} = \frac{f_{ij}}{f_{dj}} \log\left(\frac{\|D\|}{f_{ti}}\right)$

[t] The resulting vectors were normalized to the Euclidean unit length (a value between 0 and 1).

4.1.5. Inter-Document Similarity

For each of the ten questions, a collection of interventions was collected from the resulting documents. Each document was represented as a bag of weighted medical interventions, as described in **Section 4.1.4**. The resulting documents were assembled into pairs. The similarity between each pair of documents was computed using the "dist" and "simil" functions available in the R package "proxy"[10] (Meyer and Buchta, 2014). A total of 42 similarity/distance metrics were utilized to compute the similarity or distance between each pair of documents.

Bags-of-Binary Word Vectors.

Suppose that two documents, u and v derived are represented respectively by a bag-of-**binary** word vectors, and if

$a =$ the number of vectors where the values of u and v are both 1 ("positive matches"),

$b =$ the number of vectors where the values of u and v are 0 and 1 respectively, ("mismatches"),

$c =$ the number of vectors where the values of u and v are 1 and 0 respectively, ("mismatches"),

$d =$ the number of vectors where the values of u and v are both 0 ("negative matches"), and

$n =$ the sum of a, b, c and d,

the similarity between u and v, as denoted by S_{uv}, can be computed using the 20 binary similarity metrics shown in **Table 4-7**.

[10] R, also called "GNU S", is a free software environment for statistical computation and graphics. It provides a programming language, high levels graphics and a debugger environment. The root of R is the S language, which was developed by John Chambers and colleagues at Bell Laboratories. It is a software package with pre-programmed statistical procedures such as generalized linear models and time series analysis, and capability for programming tailored statistical analyses. The R "proxy" package provides functions for computing similarity/distance matrix between either rows or columns of a matrix/data frame. The package was used in this chapter to compute the similarity between two bags of medical interventions extracted from two documents.

Table 4-7. Binary similarity metrics.

Metric	Formula
Braun-blanquet	$S_{uv} = \frac{a}{\max[(a+b),(a+c)]}$
Dice	$S_{uv} = \frac{2a}{2a+b+c}$
Fager	$S_{uv} = \frac{a}{\sqrt{(a+b)(a+c)}} - \frac{\sqrt{(a+c)}}{2}$
Faith	$S_{uv} = \frac{(a+d)/2}{n}$
Hamman	$S_{uv} = \frac{a-(b+c)+d}{n}$
Jaccard	$S_{uv} = \frac{a}{a+b+c}$
Kulczynski1	$S_{uv} = \frac{a}{b+c}$
Kulczynski2	$S_{uv} = \frac{\frac{a}{(a+b)} + \frac{a}{(a+c)}}{2}$
Michael	$S_{uv} = \frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$
Mountford	$S_{uv} = \frac{2a}{ab+ac+2bc}$
Mozley	$S_{uv} = \frac{a \times n}{(a+b)(a+c)}$
Ochiai	$S_{uv} = \frac{a}{\sqrt{(a+b)(a+c)}}$
Phi	$S_{uv} = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
Russel	$S_{uv} = \frac{a}{n}$
Simple Matching	$S_{uv} = \frac{(a+d)}{n}$
Simpson	$S_{uv} = \frac{a}{\min[(a+b),(a+c)]}$
Stiles	$S_{uv} = \log_{10} \frac{n(\lvert ad-bc\rvert - \frac{n}{2})^2}{(a+b)(a+c)(b+d)(c+d)}$
Tanimoto	$S_{uv} = \frac{a+d}{a+2b+2c+d}$
Yule	$S_{uv} = \frac{ad-bc}{ad+bc}$
Yule2	$S_{uv} = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$

Bags-of-Numerical Word Vectors.

If two documents, u and v, are represented respectively by a bag-of-**numerical** word vectors, X and Y, then

$X = (x_1, x_2, \dots, x_n)$,

$Y = (y_1, y_2, \dots, y_n)$, and

$n =$ the total number of word vectors.

The similarity and distance between u and v, as denoted by S_{uv} and D_{uv} respectively, can be computed using the 17 numerical similarity/distance metrics shown in **Table 4-8**.

Bags-of-Nominal Word Vectors.

If two documents, u and v, are represented respectively by a bag-of-**nominal** word vectors, X and Y, then

$X = (x_1, x_2, \dots, x_n)$,

$Y = (y, y_2, \dots, y_n)$, and

$n =$ the total number of word vectors.

The similarity between u and v can be computed using the 5 nominal similarity metrics shown in **Table 4-9**.

Distance-to-Similarity Conversion.

A distance value, D_i, was converted to a similarity value, S_i, using:

$$S_i = \frac{1}{D_i + 1}$$

The resulting similarity values were normalized to a scale of 0 to 1. Suppose that:

$G = (G_1, G_2, \dots, G_t)$ are the similarity values of t pairs of documents,

$G_{min} =$ the minimum value of G, and

$G_{max} =$ the maximum value of G

the normalized similarity value of each pair of documents, as denoted by N_i, was calculated using:

$$N_i = \frac{G_i - G_{min}}{G_{max} - G_{min}}$$

where $i = 1, 2, \dots, t$.

Table 4-8. Numerical similarity and distance metrics.

Metric	Formula[t.]
Correlation	$S_{uv} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$
Cosine	$S_{uv} = \frac{\sum(x_i y_i)}{\sqrt{\sum(x_i)^2 \cdot \sum(y_i)^2}}$
eJaccard	$S_{uv} = \frac{\sum(x_i y_i)}{\sum(x_i^2 + y_i^2 - x_i y_i)}$
fJaccard	$S_{uv} = \sum \frac{\min[x_i, y_i]}{\max[x_i, y_i]}$
Bhattacharyya	$D_{uv} = \sqrt{\sum(\sqrt{x_i} - \sqrt{y_i})^2}$
Bray	$D_{uv} = \frac{\sum\lvert x_i - y_i\rvert}{\sum(x_i + y_i)}$
Canberra	$D_{uv} = \sum \frac{\lvert x_i - y_i\rvert}{\lvert x_i + y_i\rvert}$
Chord	$D_{uv} = \sum \sqrt{2(\frac{1 - x_i y_i}{\sqrt{x_i^2 y_i^2}})}$
Divergence	$D_{uv} = \sum \frac{(x_i - y_i)^2}{(x_i + y_i)^2}$
Euclidean	$D_{uv} = \sqrt{\sum(x_i - y_i)^2}$
Geodesic	$D_{uv} = \sum \arccos(\frac{x_i y_i}{\sqrt{x_i^2 y_i^2}})$
Hellinger	$D_{uv} = \sqrt{\sum(\sqrt{\frac{x_i}{\hat{x}}} - \sqrt{\frac{y_i}{\hat{y}}})^2}$
Manhattan	$D_{uv} = \sum\lvert x_i - y_i\rvert$
Soergel	$D_{uv} = \frac{\sum\lvert x_i - y_i\rvert}{\sum \max[x_i, y_i]}$
Supremum	$D_{uv} = max\lvert x_i - y_i\rvert$
Wave	$D_{uv} = \sum(1 - \frac{\min[x_i, y_i]}{\max[x_i, y_i]})$
Whittaker	$D_{uv} = \frac{\sum\left\lvert\frac{x_i}{\hat{x}} - \frac{y_i}{\hat{y}}\right\rvert}{2}$

[t.] Where, $\boldsymbol{i = 1, 2, \ldots, n}$, $\boldsymbol{\hat{x} = \sum x_i}$ $\boldsymbol{and}$ $\boldsymbol{\hat{y} = \sum y_i}$.

Table 4-9. Nominal similarity metrics.

Metric	Formula[t.]
Chi-squared	$S_{uv} = \sum \frac{(x_i - y_i)^2}{y_i}$
Cramer	$S_{uv} = \sqrt{\frac{\frac{\sum \frac{(x_i - y_i)^2}{y_i}}{n}}{\min[(a-1),(b-1)]}}$
Pearson	$S_{uv} = \sqrt{\frac{\sum \frac{(x_i - y_i)^2}{y_i}}{n + \sum \frac{(x_i - y_i)^2}{y_i}}}$
Phi-squared	$S_{uv} = \frac{\sum \frac{(x_i - y_i)^2}{y_i}}{n}$
Tschuprow	$S_{uv} = \sqrt{\frac{\frac{\sum \frac{(x_i - y_i)^2}{y_i}}{n}}{\sqrt{(a-1)(b-1)}}}$

[t.] Where, $i = 1, 2, \ldots, n$, a = number of rows and b = number of columns.

4.1.6. Paired and Unpaired Documents

The documents retrieved for each of the ten questions were divided into pairs. The similarity between each pair of documents was judged by two raters with medical background based on the [I] and [C] elements (i.e. the interventions) appear in the titles and abstracts. The similarity-rating task involved two key steps.

Step 1: Identification of Interventions.

Figures 4-3 to **4-5** show three articles with PMIDs of 24381967, 23583234 and 19528519, respectively. The interventions, which include therapeutic/preventive procedures and medications, were identified by Rater 1 and Rater 2 from the titles and abstracts of the articles.

TITLE: Effect of vitamin E and memantine on functional decline in Alzheimer disease: the TEAM-AD VA cooperative randomized trial.

ABSTRACT: IMPORTANCE: Although vitamin E and memantine have been shown to have beneficial effects in moderately severe Alzheimer disease (AD), evidence is limited in mild to moderate AD. OBJECTIVE: To determine if vitamin E (alpha tocopherol), memantine, or both slow progression of mild to moderate AD in patients taking an acetylcholinesterase inhibitor. DESIGN, SETTING, AND PARTICIPANTS: Double-blind, placebo-controlled, parallel-group, randomized clinical trial involving 613 patients with mild to moderate AD initiated in August 2007 and concluded in September 2012 at 14 Veterans Affairs medical centers. INTERVENTIONS: Participants received either 2000 IU/d of alpha tocopherol (n = 152), 20 mg/d of memantine (n = 155), the combination (n = 154), or placebo (n = 152). MAIN OUTCOMES AND MEASURES: Alzheimer's Disease Cooperative Study/Activities of Daily Living (ADCS-ADL) Inventory score (range, 0-78). Secondary outcomes included cognitive, neuropsychiatric, functional, and caregiver measures. RESULTS: Data from 561 participants were analyzed (alpha tocopherol = 140, memantine = 142, combination = 139, placebo = 140), with 52 excluded because of a lack of any follow-up data. Over the mean (SD) follow-up of 2.27 (1.22) years, ADCS-ADL Inventory scores declined by 3.15 units (95% CI, 0.92 to 5.39; adjusted P = .03) less in the alpha tocopherol group compared with the placebo group. In the memantine group, these scores declined 1.98 units less (95% CI, -0.24 to 4.20; adjusted P = .40) than the placebo group's decline. This change in the alpha tocopherol group translates into a delay in clinical progression of 19% per year compared with placebo or a delay of approximately 6.2 months over the follow-up period. Caregiver time increased least in the alpha tocopherol group. All-cause mortality and safety analyses showed a difference only on the serious adverse event of "infections or infestations," with greater frequencies in the memantine (31 events in 23 participants) and combination groups (44 events in 31 participants) compared with placebo (13 events in 11 participants). CONCLUSIONS AND RELEVANCE: Among patients with mild to moderate AD, 2000 IU/d of alpha tocopherol compared with placebo resulted in slower functional decline. There were no significant differences in the groups receiving memantine alone or memantine plus alpha tocopherol. These findings suggest benefit of alpha tocopherol in mild to moderate AD by slowing functional decline and decreasing caregiver burden.

RATER 1: Vitamin E (alpha tocopherol), memantine, acetylcholinesterase inhibitor

RATER 2: Vitamin E (alpha tocopherol), memantine, acetylcholinesterase inhibitor, placebo

Figure 4-3. An article with PMID of 24381967.

TITLE: Vitamin E and memantine in Alzheimer's disease: clinical trial methods and baseline data.

ABSTRACT: BACKGROUND: Alzheimer's disease (AD) has been associated with both oxidative stress and excessive glutamate activity. A clinical trial was designed to compare the effectiveness of (i) alpha-tocopherol, a vitamin E antioxidant; (ii) memantine (Namenda), an N-methyl-D-aspartate antagonist; (iii) their combination; and (iv) placebo in delaying clinical progression in AD. METHODS: The Veterans Affairs Cooperative Studies Program initiated a multicenter, randomized, double-blind, placebo-controlled trial in August 2007, with enrollment through March 2012 and follow-up continuing through September 2012. Participants with mild-to-moderate AD who were taking an acetylcholinesterase inhibitor were assigned randomly to 2000 IU/day of alpha-tocopherol, 20 mg/day memantine, 2000 IU/day alpha-tocopherol plus 20 mg/day memantine, or placebo. The primary outcome for the study is the Alzheimer's Disease Cooperative Study/Activities of Daily Living Inventory. Secondary outcome measures include the Mini-Mental State Examination; the Alzheimer's Disease Assessment Scale, cognitive portion; the Dependence Scale; the Neuropsychiatric Inventory; and the Caregiver Activity Survey. Patient follow-up ranged from 6 months to 4 years. RESULTS: A total of 613 participants were randomized. The majority of the patients were male (97%) and white (86%), with a mean age of 79 years. The mean Alzheimer's Disease Cooperative Study/Activities of Daily Living Inventory score at entry was 57 and the mean Mini-Mental State Examination score at entry was 21. CONCLUSION: This large multicenter trial will address the unanswered question of the long-term safety and effectiveness of alpha-tocopherol, memantine, and their combination in patients with mild-to-moderate AD taking an acetylcholinesterase inhibitor. The results are expected in early 2013.

RATER 1: Vitamin E antioxidant (alpha tocopherol), N-methyl-D-aspartate antagonist (memantine/Namenda), acetylcholinesterase inhibitor

RATER 2: Vitamin E (alpha tocopherol), placebo, memantine, acetylcholinesterase inhibitor

Figure 4-4. An article with PMID of 23583234.

TITLE: Donepezil delays progression to AD in MCI subjects with depressive symptoms.

ABSTRACT: OBJECTIVE: To determine whether the presence of depression predicts higher rate of progression to Alzheimer disease (AD) in patients with amnestic mild cognitive impairment (aMCI) and whether donepezil treatment beneficially affect this relationship. METHODS: The study sample was composed of 756 participants with aMCI from the 3-year, double-blind, placebo-controlled Alzheimer's Disease Cooperative Study drug trial of donepezil and vitamin E. Beck Depression Inventory (BDI) was used to assess depressive symptoms at baseline and participants were followed either to the end of study or to the primary endpoint of progression to probable or possible AD. RESULTS: Cox proportional hazards regression, adjusted for age at baseline, gender, apolipoprotein genotype, and NYU paragraph delayed recall score, showed that higher BDI scores were associated with progression to AD ($p = 0.03$). The sample was stratified into depressed (BDI score > or =10; $n = 208$) and nondepressed (BDI <10; $n = 548$) groups. Kaplan-Meier analysis showed that among the depressed subjects, the proportion progressing to AD was lower for the donepezil group than the combined vitamin E and placebo groups at 1.7 years ($p = 0.023$), at 2.2 years ($p = 0.025$), and remained marginally lower at 2.7 years ($p = 0.070$). The survival curves among the three treatment groups did not differ within the nondepressed participants. CONCLUSIONS: Results suggest that depression is predictive of progression from amnestic mild cognitive impairment (aMCI) to Alzheimer disease (AD) and treatment with donepezil delayed progression to AD among depressed subjects with aMCI. Donepezil appears to modulate the increased risk of AD conferred by the presence of depressive symptoms.

RATER 1: Donepezil, vitamin E

RATER 2: Donepezil, placebo, vitamin E

Figure 4-5. An article with PMID of 19528519.

Step 2: Rating of Similarity.

Using the examples given in Step 1, the raters determined the similarity of interventions between two documents. A score of 1 was assigned to documents with highly similar interventions; 0, otherwise. **Table 4-10** shows the similarity of three pairs of documents determined by Rater 1 and Rater 2. A total of 1225 pairs of documents were rated respectively for Questions 1-6, 8 and 10. For Questions 7 and 9, 861 and 120 pairs of documents respectively were rated.

Table 4-10. Similarity rating of three pairs of documents.

PMID	-	PMID	Rater 1	Rater 2
24381967	-	23583234	1	1
23583234	-	19528519	1	0
19528519	-	24381967	1	0

Inter-Rater Agreement.

The two raters performed the similarity task independently and agreed as shown in **Table 4-11**. Assume that,

a = the number of pairs that both raters agree to be similar,

d = the number of pairs that both raters agree to be dissimilar, and

b, c = the number of pairs that both raters disagree on,

the inter-rater agreement was assessed using Cohen's kappa statistic, κ:

$$\kappa = \frac{2(ad - bc)}{(a + c)(c + d) + (a + b)(b + d)}$$

Table 4-11. Agreement between two raters.

		Rater 1's Judgment	
		Positive	***Negative***
Rater 2's Judgment	***Positive***	a	b
	Negative	c	d

The strength of agreement was interpreted according to the guidelines by Landis and Koch (1977) (**Table 4-12**). Documents with highly similar interventions were identified as "paired documents" whereas those with low similarity of interventions were identified as "unpaired documents". In case of disagreement, the pairs of documents were excluded for the separation tests described in **Section 4.1.7**.

Table 4-12. Strength of agreement by kappa statistic.

Kappa	Agreement
< 0	Less than change agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 0.99	Alomost perfect agreement

4.1.7. Separation Tests

A combination of descriptive and inferential statistical techniques was used to analyze the similarity distributions of paired and unpaired documents.

Mean Difference.

The mean difference between paired and unpaired similarities, S_{MD}, was calculated using:

$$S_{MD} = \overline{S_{paired}} - \overline{S_{unpaired}}$$

where $\overline{S_{paired}}$ is the mean of paired similarity values and $\overline{S_{unpaired}}$ is the mean of unpaired similarity values. The greater the value of S_{MD}, the better the two similarity distributions are separated from each other.

One-Way ANOVA.

This was performed to determine whether there is a significant difference between the means of paired and unpaired similarities. The test was performed on similarity metrics that achieved the highest mean differences (referred to hereafter as the "**top similarity metrics**"). A p-value less than 0.05 indicates a statistically significant difference exists between the two means.

Histogram.

All similarity values were normalized to a scale from 0 to 1. Paired documents have a similarity value close to 1 and unpaired documents have a similarity value close to 0. The distributions of paired and unpaired similarities were presented respectively by a histogram with intervals of equal length. The two histograms were merged to examine

the performance of the top similarity metrics in separating paired documents from unpaired documents. As shown in **Figure 4-6**, paired documents are separated far apart from unpaired documents in plot A. This indicates good separation. In plot B, the two distributions overlapped each other, indicating poor separation.

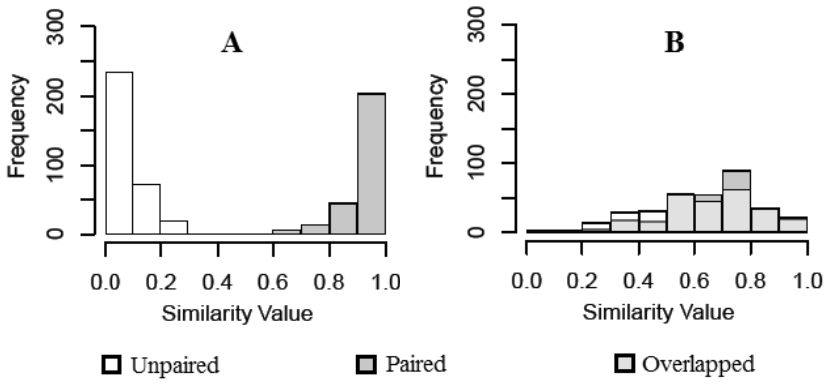


Figure 4-6. Histograms of paired and unpaired similarities.

Boxplot.

Similar to histograms, boxplots were created to assess the effectiveness of the top similarity metrics in separating paired documents from unpaired documents. As shown in **Figure 4-7**, boxplots represent the median (the middle line), the 25th and 75th percentiles (the lower and upper edges of the boxes, respectively), the range (the whisker) and the outliers (the circles). In plot A, the median of paired documents is close to 1 whereas for unpaired documents, the median is close to 0. This indicates good separation. In plot B, the median and range of paired similarities are similar to those for unpaired similarities. There is also a high number of outliers, indicating poor separation.

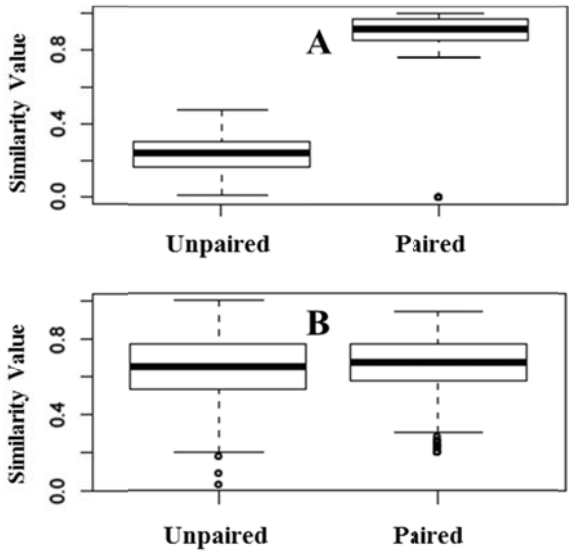


Figure 4-7. Boxplots of paired and unpaired similarities.

ROC Curve.

The relative operating characteristic (ROC) curves were constructed to evaluate the performance of top similarity metrics in measuring the similarity of paired and unpaired documents. As shown in **Figure 4-8**, a ROC curve plots true positive rate (TPR) against false positive rate (FPR). Let $TP =$ true positive, $FP =$ false positive, $FN =$ false negative and $TN =$ true negative predictions. The values of TPR and FPR were computed using:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{TN + FP}$$

The TPR is also known as sensitivity, and the FPR is equal to $(1 -$ specificity). The figure shows the ROC curve for each of the questions under study in dotted line, and the average of the ten curves in solid line. The average curve was achieved by calculating the means of $TPRs$ and $FPRs$. The areas under the curves (AUC) were measured to indicate the degree of accuracy. An area of 1.0 represents perfect discrimination or 100% accuracy whereas an area of 0.5 indicates performance no better than chance. The greater the AUC, the better the ability of a similar metric in differentiating paired documents from unpaired documents.

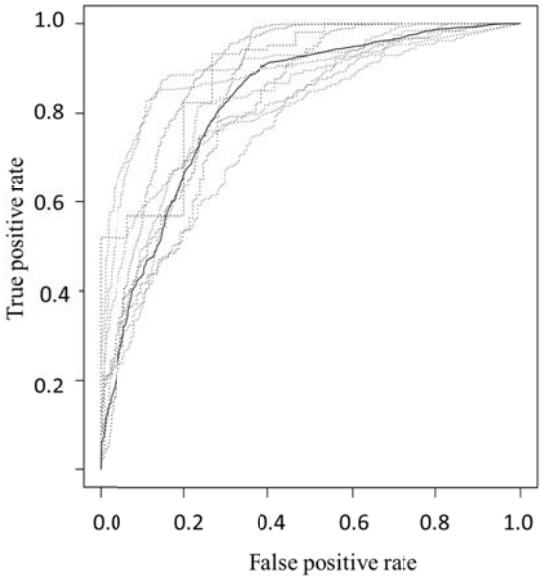


Figure 4-8. An example of ROC curves.

The histograms, boxplots and ROC curves were produced using R statistical software and the mean differences and one-way ANOVA were calculated using SPSS (version 20.0.0, IBM Corporation, New York, USA).

4.2. RESULTS and DISCUSSION

4.2.1. Inter-Rater Agreement

The kappa statistic measures the degree of agreement between two raters. The kappa values for the ten questions range from 0.810 to 0.970. This gives an average of 0.894 (**Table 4-13**). The values indicate that there is a strong agreement between Rater 1 and Rater 2. 10323 out of 10781 pairs of documents agreed by both raters were included for the separation tests, of which 5243 were paired and 5080 were unpaired documents.

Table 4-13. Kappa values for ten questions.

Question	Weighted Kappa ± S.E.
1	0.874 ± 0.015
2	0.901 ± 0.012
3	0.932 ± 0.010
4	0.903 ± 0.015
5	0.911 ± 0.013
6	0.970 ± 0.008
7	0.810 ± 0.027
8	0.829 ± 0.016
9	0.895 ± 0.016
10	0.894 ± 0.016

4.2.2. Mean Difference

The S_{MD} indicates the difference between the mean of paired and the mean of unpaired similarities. The larger the S_{MD}, the greater the difference and the less overlap between the distributions of paired and unpaired similarities. For each of the ten questions under analysis, 4 weighting schemes and 42 similarity or distance metrics were employed to calculate the similarity of paired and unpaired documents using interventions derived from 18 different sources. Therefore, a total of (4 × 42 × 18) = 3024 sets of similarity or distances values were generated for each question. The resulting distance values were converted to similarity values. The similarity values were then normalized to a score between 0 and 1. A value close to 1 indicates strong similarity whereas a value close to 0 means low similarity. The findings obtained are as follows:

Finding 1: The weighting scheme BO gave the highest S_{MD}, followed by TO and TF, and the lowest by TFIDF. Each document was represented as a bag of weighted word vector. The greater the weight of a word, the more important the word to a document. Given that the word "*aspirin*" occurs 5 times in a bag of 10 words, BO returns a value of 1 and TO returns a value of 5. TF takes into account the total number of words in the bag of words and gives a weight of $\frac{5}{10} = 0.50$. Assume that the word occurs at least once in 30 out of 50 documents, TFIDF returns a value of $\frac{5}{10}\log\frac{50}{30} = 0.11$. TFIDF reduces the relative importance of high frequency words. Low frequency words are assigned a higher weight in order to distinguish a document from other documents. TO and TF, in contrast, assign a higher weight to high frequency words. BO depends only on the occurrence of a word, regardless of whether it is a high or low frequency words. **Table 4-14** summarizes the metrics that produced the highest S_{MD} by weighting scheme. As shown in the table, BO performed better than the other three weighting schemes with the highest S_{MD}. The results indicate that the frequency of words has less influence on the separation of paired and unpaired documents.

Table 4-14. Mean differences (S_{MD}) by 4 weighting schemes.[t]

Question	BO		TF		TFIDF		TO	
	Metric	S_{MD}	**Metric**	S_{MD}	**Metric**	S_{MD}	**Metric**	S_{MD}
1	Yule	0.54	Correlation	0.30	eJaccard	0.12	Pearson	0.32
2	Yule	0.73	Cosine	0.45	Cosine	0.19	Pearson	0.48
3	Yule	0.42	Correlation	0.18	Correlation	0.07	Pearson	0.21
4	Stiles	0.55	Correlation	0.34	Cosine	0.19	Pearson	0.39
5	Yule	0.56	Correlation	0.27	Correlation	0.15	Pearson	0.33
6	Yule	0.63	Cosine	0.45	Cosine	0.30	Pearson	0.49
7	Yule	0.46	Cosine	0.37	Whittaker	0.25	Pearson	0.32
8	Yule	0.69	Correlation	0.34	Divergence	0.10	Pearson	0.45
9	Yule	0.66	Whittaker	0.52	Whittaker	0.48	Tschuprow	0.50
10	Stiles	0.32	Cosine	0.22	Supremum	0.06	Cramer	0.23

[t] The table shows only the similarity metric that yielded the highest mean difference.

Finding 2: The binary similarity metric, Yule, performed the best among the 42 similarity/distance metrics. As shown in **Table 4-14**, Yule yielded the highest S_{MD} for Questions 1-3 and 5-9. For Questions 4 and 10, Stiles achieved the highest S_{MD}. For each question, 10 out of the 3024 sets of similarity values with the highest S_{MD} were chosen. This gives 100 sets of similarity values for ten questions. The metrics used to compute the similarity values were identified and are summarized as depicted in **Figure 4-9**. Binary metrics performed better than nominal and numerical metrics. The metrics, ranked from the highest to the lowest frequency, are 39 for Yule, 20 for Yule2, 19 for Stiles, 11 for Simpson, 3 for Ochiai, 2 for Fager and Pearson, and 1 for Correlation, Cosine, Dice and Kulczynski2. These eleven metrics are defined as the "**top similarity metrics**". 89 out of the 100 sets of similarity values were achieved by using the top four binary metrics: Yule, Yule2, Stiles and Simpson. The results indicate that Yule performed the best in separating paired documents from unpaired documents.

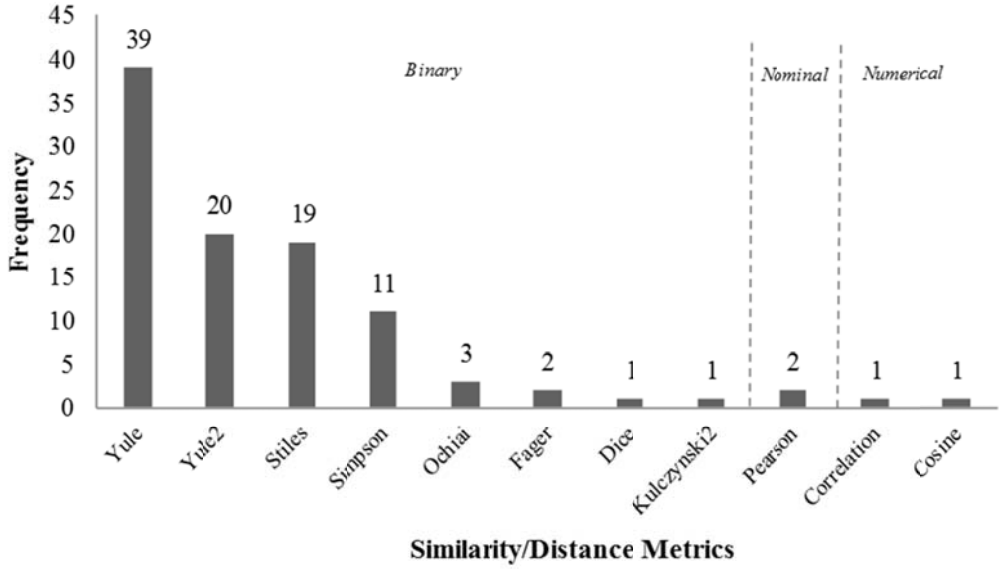


Figure 4-9. Top similarity or distance metrics by frequency.

Finding 3: Interventions derived from "Full abstracts" or "Titles + Full Abstracts" performed better than those derived from other sources. Interventions generated from five different fields of MEDLINE documents were evaluated. **Table 4-15** summarizes the metrics that produced the highest S_{MD} by the source of interventions. As shown in the table, the highest values were contributed by four sources of interventions: B = "Full abstracts" for Questions 3, 5, 8 and 10, D = "Titles" for Question 2, J = "Titles + Full abstracts" for Questions 1, 6 and 7 and M = "Titles + Chemicals" for Questions 4 and 9.

Table 4-15. Mean difference (S_{MD}) by 18 sources of interventions.[t]

Question	Source Indicator							
	B	**D**	**G**	**J**	**K**	**M**	**Q**	**R**
1	Yule (0.53)			**Yule (0.54)**	Yule (0.32)			
2		**Yule (0.73)**	Yule (0.68)			Simpson (0.54)		
3	**Yule (0.42)**		Yule (0.31)	Yule (0.39)				
4						**Stiles (0.55)**	Stiles (0.48)	Stiles (0.45)
5	**Yule (0.56)**			Yule (0.52)	Yule (0.43)			
6	Yule (0.61)		Yule (0.59)	**Yule (0.63)**				
7	Yule (0.43)		Yule2 (0.32)	**Yule (0.46)**				
8	**Yule (0.70)**			Yule (0.68)			Yule (0.41)	
9					Yule (0.54)	**Yule (0.66)**		
10	**Stiles (0.32)**		Yule (0.29)	Stiles (0.29)				

[t] The table shows the top three metrics that yielded the highest mean difference by each question.
B = "Full abstracts", D = "Titles", G = "Titles + PICO sentences",
J = "Titles + Full abstracts", K = "Chemicals + Full abstracts",
M = "Titles + Chemicals", Q = "Titles + Chemicals + Full abstracts",
and R = "Titles + Chemicals + MeSH terms".

For each question, 10 out of the 3024 sets of similarity values with the highest S_{MD} were chosen. This gives 100 sets of similarity values for ten questions. The sources of interventions for the 100 sets of similarity values were identified and illustrated in **Figure 4-10**. The top three sources of interventions, ranked from the highest to the lowest frequency, are 25 for "Titles + Full abstracts", 24 for "Full abstracts" and 11 for "Titles + PICO sentences". The results showed that interventions derived from "Full abstracts" and "Titles + Full abstracts" performed remarkably better than those derived from other sources in separating paired documents from unpaired documents.

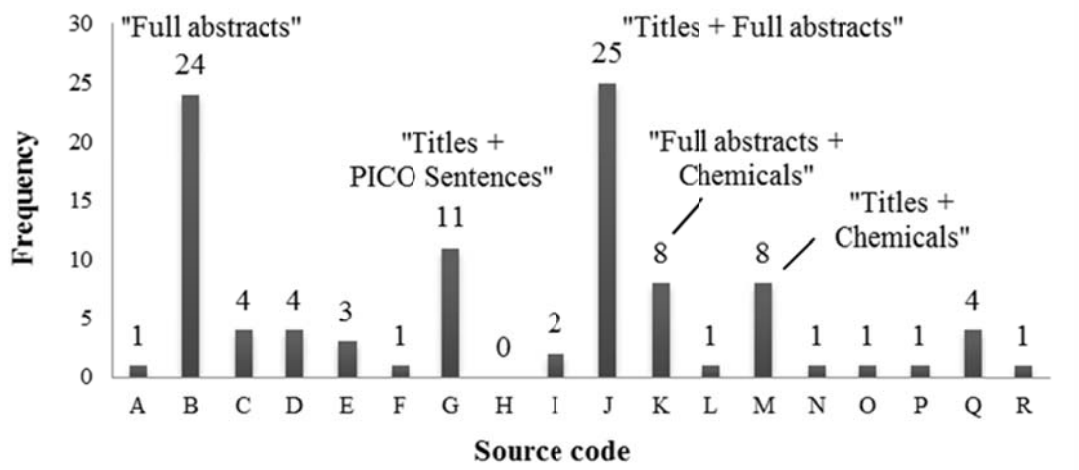


Figure 4-10. Performance of 18 sources of interventions by frequency.

Finding 4: "Full abstracts" play main role in separating paired documents from unpaired documents. The [I] and [C] elements extracted from five different fields of MEDLINE documents: titles, full abstracts, PICO sentences, MeSH terms and chemicals were used to form the 18 sources of interventions. The mean differences (S_{MD}) of similarity values calculated using Yule metric for Questions 1-10 were averaged. **Figure 4-11** illustrates the influence of the five fields to the separation of paired and unpaired documents by average S_{MD}. The highest average S_{MD} were obtained by "Full abstracts" (0.484) and "Titles + Full abstracts" (0.470). The combinations of two fields using "Titles", "PICO sentences", "MeSH terms" and "Chemicals" (except of "Full abstracts") improved the separation slightly. A combination of three fields did not improve the separation. For instance, the addition of "Chemicals" to "Titles + PICO sentences" decreased the average S_{MD} from 0.337 to 0.286. However, the combinations of "Full abstracts" with the other four fields improved the separation. For example, the average S_{MD} of "Titles" increased from 0.201 to 0.470 upon the addition of "Full abstracts". In brief, the results indicate that the [I] and [C] elements generated from full abstracts are useful for the separation of paired and unpaired documents.

Finding 5: The Yule similarity metric performed better than common similarity metrics. Common similarity metrics used in text mining include Cosine, Correlation, eJaccard and Euclidean (Strehl et al., 2000). The mean differences (S_{MD}) of similarity values calculated using Yule and common similarity metrics for the ten questions were averaged. The average S_{MD} scores of the five metrics are illustrated in **Figure 4-12** with an increase in number of pairs of interventions. The figure shows that

the number of pairs has little influence on the performance of the five metrics. The average S_{MD} of Yule (0.50 ± 0.02) was evidently higher than the average S_{MD} of Cosine (0.23 ± 0.01), Correlation (0.20 ± 0.01), eJaccard (0.13 ± 0.02) and Euclidean (0.02 ± 0.01).

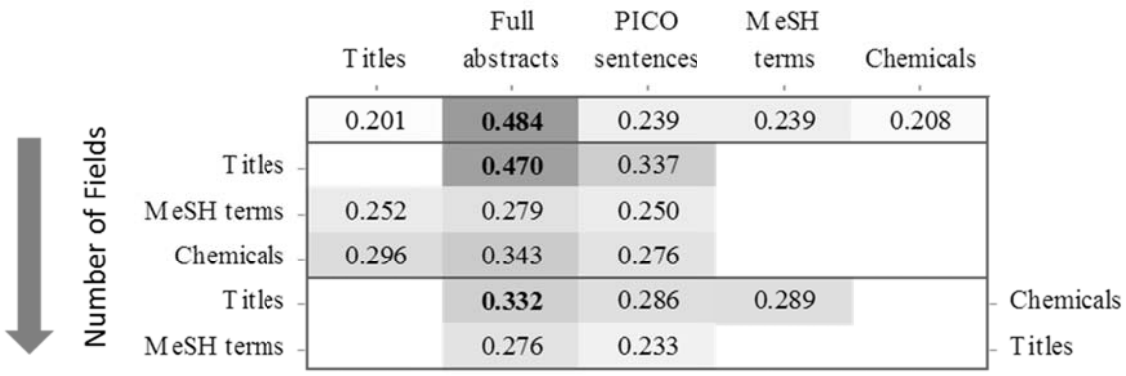


Figure 4-11. Effects of five fields of MEDLINE documents on average S_{MD}. Similarity metric: Yule

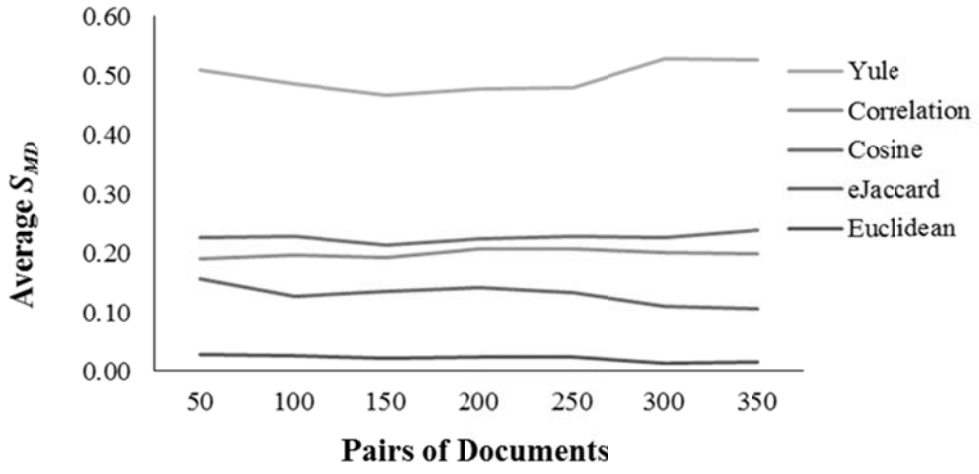


Figure 4-12. Average S_{MD} against number of pairs of documents. Derivation of interventions: "Titles + Full abstracts"

In summary, in terms of S_{MD}, the paired and unpaired documents were best separated using the binary similarity metric, Yule, and the interventions derived from "Full abstracts" or "Titles + Full abstracts".

4.2.3. One-Way ANOVA

One-way ANOVA (analysis of variance) was employed to determine whether the means differences (S_{MD}) between paired and unpaired similarities were significant at $p < 0.05$. A significant difference indicates that paired and unpaired documents are well separated. The same analysis was carried out for the ten questions under study. **Tables 4-16** and **4-17** show the results for Question 1 and Question 9. The results for Question 1 showed that, using interventions derived from "Titles", "Full abstracts" and "Titles + Full abstracts", the mean differences of similarity values calculated using the top similarity metrics were all statistically significant ($p < 0.05$). An insignificant difference was found ($p \geq 0.05$) when the similarity values were calculated using Dice, Kulcynzki2 and Ochiai for interventions derived from "PICO sentences" and "Titles + PICO sentences". The results for Question 9 showed a higher number of insignificant differences between paired and unpaired similarities.

Figure 4-13 illustrates the performance of the top similarity metrics and the five sources of interventions by the number of questions with significant ($p < 0.05$) and insignificant ($p \geq 0.05$) mean differences. The combination of "Titles" with "PICO sentences" did not result in a large increase in number of questions with significant mean differences. The PICO sentences were extracted from the full abstracts, as described in **Section 4.1.2**. The results indicate that "PICO sentences" and "Titles" provide insufficient or discrete [I] and [C] elements for the separation of paired documents from unpaired documents. "Full abstracts" and "Titles + Full abstracts" performed the best with a higher number of questions with p values < 0.05, indicating that a good separation can be achieved using the interventions generated from the two sources. The performance of the top similarity metrics in separating paired and unpaired documents using interventions derived from "Titles + Full abstracts" are the same (No. of questions with $p < 0.05 = 10$), except for Stiles that returned a small number of insignificant mean differences. A quite similar finding was obtained by using interventions derived from "Full abstracts" for the separation of paired and unpaired documents.

The use of the binary metric, Yule, for the measurement of similarity between documents, and the use of "Titles + Full abstracts" or "Full abstracts" as the source of interventions were further supported by the results from the analysis of variance (one-way ANOVA).

Table 4-16. One-way ANOVA analysis of paired and unpaired similarities for Question 1.[t]

Metric	Source of interventions				
	Titles	PICO sentences	Full abstracts	Titles + PICO sentences	Titles + Full abstracts
Yule	■	■	■	■	■
Yule2	■	■	■	■	■
Dice	■	□	■	□	■
Fager	■	■	■	■	■
Kulczynski2	■	□	■	□	■
Ochiai	■	□	■	□	■
Simpson	■	□	■	■	■
Stiles	■	■	■	■	■
Correlation	■	■	■	■	■
Cosine	■	■	■	■	■
Pearson	■	■	■	■	■

[t] The symbols "■" indicates $p < 0.05$ and "□" indicates $p \geq 0.05$.

Table 4-17. One-way ANOVA analysis of paired and unpaired similarities for Question 9.[t]

Metric	Source of interventions				
	Titles	PICO sentences	Full abstracts	Titles + PICO sentences	Titles + Full abstracts
Yule	■	□	■	□	■
Yule2	■	□	■	□	■
Dice	□	□	■	□	■
Fager	□	□	■	□	■
Kulczynski2	□	□	■	□	■
Ochiai	□	□	■	□	■
Simpson	□	□	■	□	■
Stiles	□	□	□	□	□
Correlation	□	□	■	■	■
Cosine	□	□	■	□	■
Pearson	□	□	□	■	■

[t] The symbols "■" indicates $p < 0.05$ and "□" indicates $p \geq 0.05$.

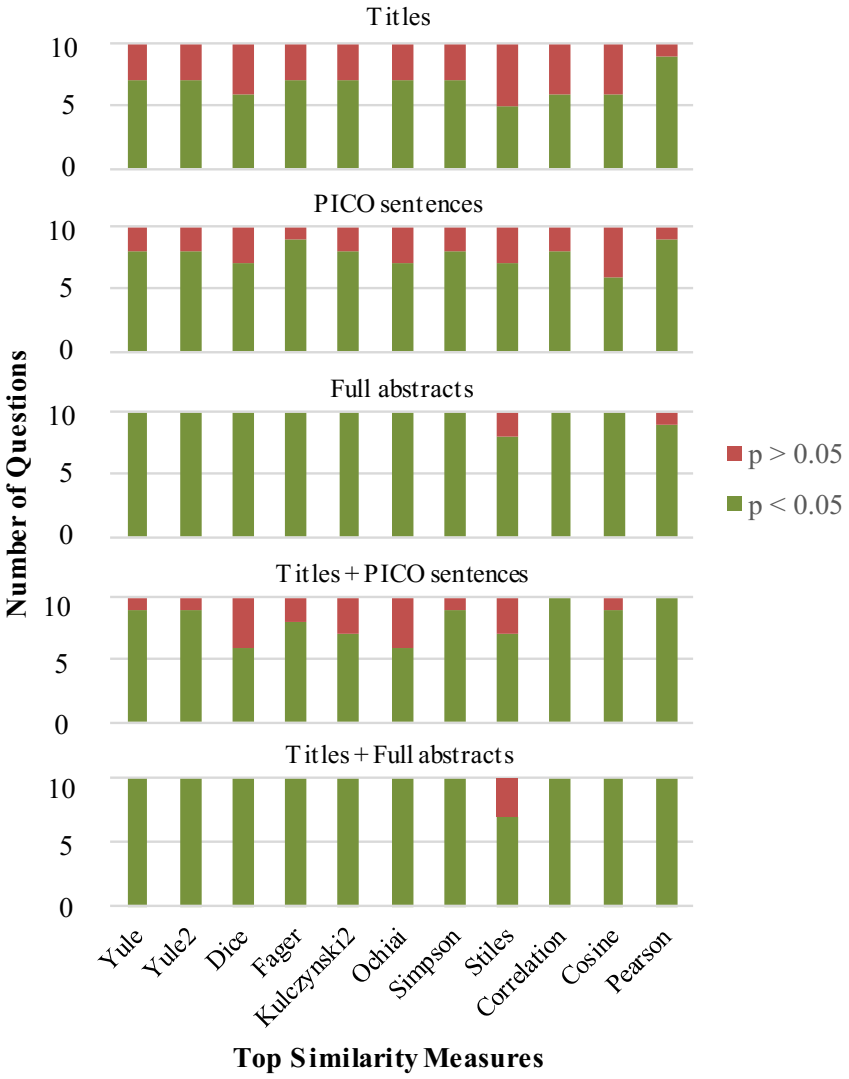


Figure 4-13. Number of questions with significant mean differences.

4.2.4. Histograms and Boxplots

Histograms and boxplots were created to investigate the frequency distributions of paired and unpaired similarities. A value close to 1 indicates strong similarity whereas a value close to 0 means low similarity. As described in **Section 4.1.7**, the less overlap between two histograms, the better the separation between paired and unpaired documents. The range and median of similarity values and the outliers were examined using boxplots. The interventions derived from "Titles + Full abstracts" were used to analyze the performance of the top similarity metrics. For each metric, the most common distribution patterns out of the 10 questions analyzed were identified. The common patterns of distributions are illustrated using histograms and boxplots in **Figure 4-14** and are described as follow:

Pattern 1: Paired histogram was skewed strongly to the right and unpaired histogram was skewed sharply to the left. This indicates that for paired documents, the similarity values were close or equal to 1 (median ~ 0.9, range ~ 0.7 to 1.0) whereas for unpaired documents, the similarity values were close to 0 (median 0.0, range 0.0 to 1.0). A small region of overlap was found between the two histograms (between similarity values of 0.6 and 1.0). **The two distributions were well separated with overlap in high similarity region.**

Pattern 2: Paired histogram was relatively flat with no sharp peak and unpaired histogram was skewed massively to the left. The median of paired documents was 0.4 (range ~ 0.1 to ~ 0.9) whereas for unpaired documents, the median was 0.0 (range 0.0 to ~ 0.5) with high number of outliers. The two histograms overlapped largely between similarity values of 0.1 and 0.5. Although the similarity values of most of unpaired documents were close to 0, **the two distributions were not distinctly separated with paired documents occurring in high similarity region**.

Pattern 3: Paired histogram was skewed to the left with no sharp peak whereas for unpaired histogram, the distribution was skewed significantly to the left. The similarity values of most of the paired documents were less than 0.5 (median ~ 0.2, range 0.0 to ~ 0.6). The median of unpaired documents was close to 0 with high number of outliers. The zone of overlap was found mainly between similarity values of 0.0 and 0.4. **The two distributions overlapped each other in low similarity region and were poorly separated**.

Pattern 4: The distribution of paired histogram was uneven whereas for unpaired histogram, the distribution was skewed slightly to the left. Both of the histograms were considered flat with no sharp peaks. The medians of paired and unpaired boxplots were fairly close to each other (~ 0.2 and ~ 0.4, respectively). A low number of paired documents (which were identified as the outliers) occurred in high similarity region. **The two distributions overlapped each other in low similarity region and were very poorly separated**.

Yule and Yule2 tended to produce pattern 1 that caused a wide separation of paired and unpaired similarities. Kulczynski2, Correlation, Cosine and Pearson were more likely to produce pattern 2. In pattern 2, unpaired documents were assigned mostly to low similarity region whereas paired documents were distributed from low to high similarity region. Fager tended to produce pattern 4 which failed to separate the two distributions. Dice, Ochiai and Simpson produced different patterns of distributions,

which include patterns 2, 3 and 4 that demonstrated moderate to poor separation. The performance of Stiles was inconsistent, which produced wide, moderate and poor separations (patterns 1, 2 and 4 respectively). The degree of overlap between paired and unpaired histograms for the four patterns looked apparently the same.

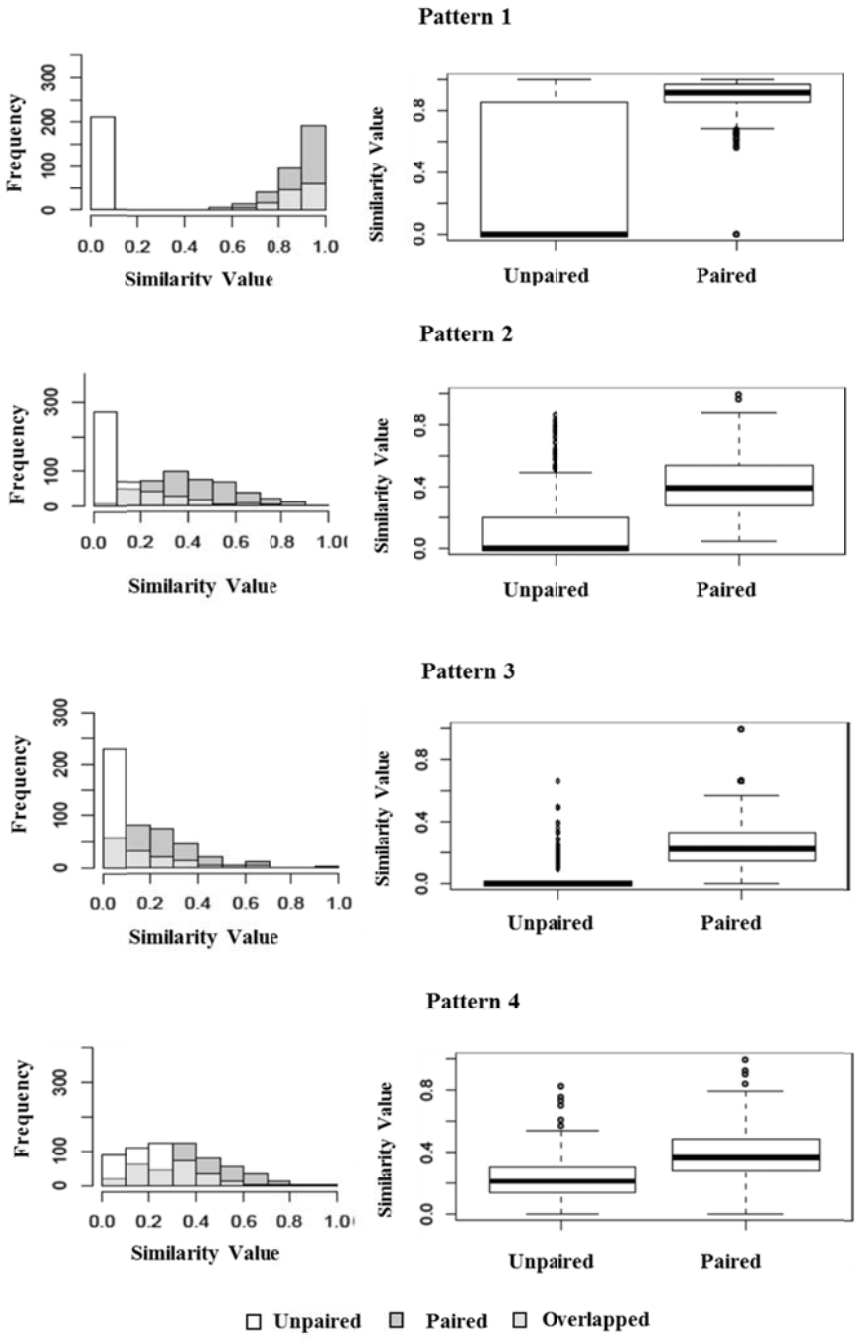


Figure 4-14. Histograms and boxplots showing the patterns of distributions of paired and unpaired similarities.

In summary, in terms of classifiability, Yule and Yule2 resulted in a more clear-cut separation of paired and unpaired similarities in histograms than other similarity metrics.

4.2.5. ROC Curves

The performance of top similarity metrics were further evaluated using ROC curves. A ROC curve was constructed for each of the metrics using average true positive rate (TPR) and average false positive rate (FPR). **Figure 4-15** shows a zoomed in view of the ROCs of top similarity metrics. The closer a ROC curve is to the top left, the better the performance of a similarity metric. The four similarity metrics: Yule, Yule2, Correlation and Cosine performed the best with AUCs of 0.82, 0.82, 0.85 and 0.85 respectively. The values, interpreted based on the guidelines by Hosmer and Lemeshow (2000), indicate excellent discrimination.

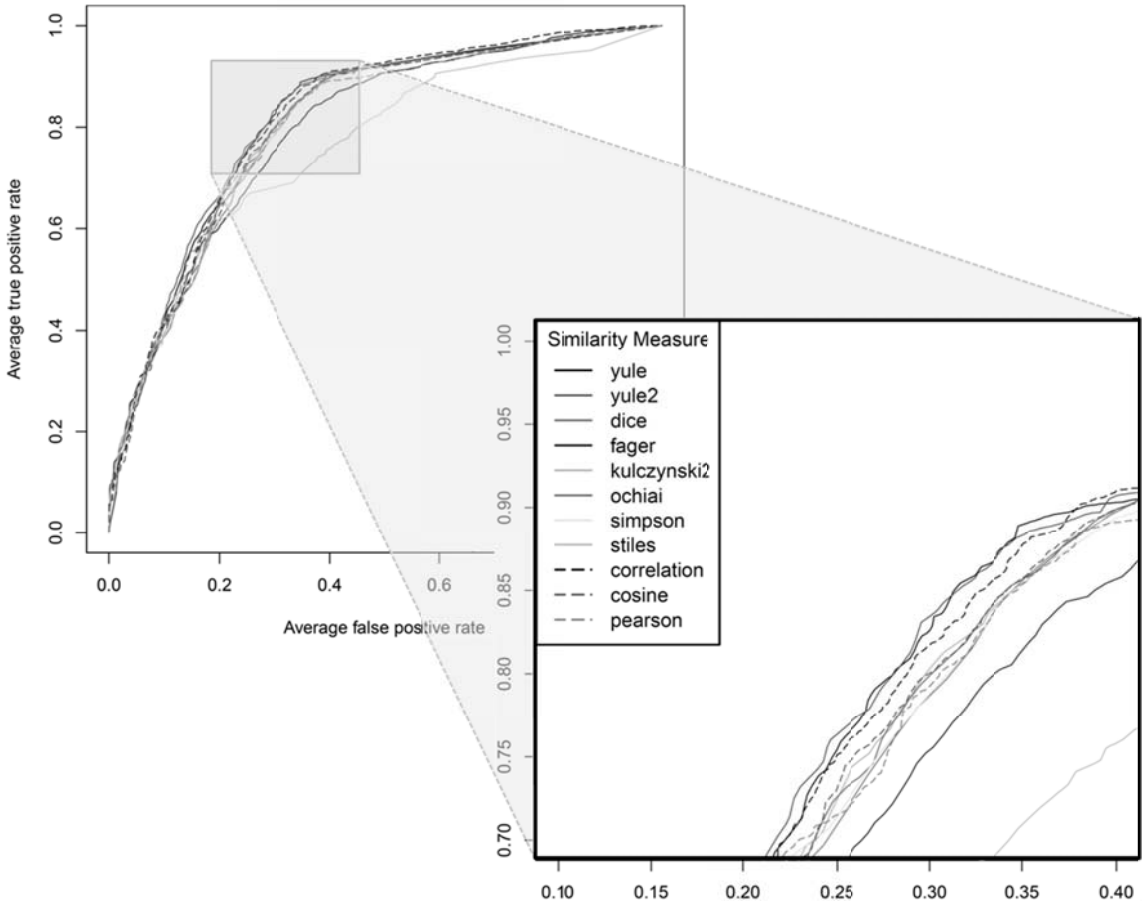


Figure 4-15. ROC curves of top similarity metrics.

Sensitivity is defined as TPR, which reflects the ability of a metric to correctly identify true positives (i.e. paired documents). 1-FPR is defined as specificity, which reflects the ability of a metric to correctly identify true negatives (i.e. unpaired documents) or to avoid false positives. As shown in **Table 4-18**, Yule and Yule2 were superior to Correlation and Cosine in terms of sensitivity and specificity. Yule achieved

the highest specificity when the sensitivity was fixed at 0.80 or higher. The highest sensitivity was shown by Yule2 when the specificity was fixed at 0.80. However, the differences in specificity and sensitivity values between the four similarity metrics were small, suggesting that the performance of the four metrics was comparable.

Table 4-18. Sensitivity and specificity of top four similarity metrics.

		Sensitivity =			
	Metric	**0.90**	**0.85**	**0.80**	**0.75**
Specificity	Yule	0.62	0.69	0.73	0.76
	Yule2	0.60	0.68	0.72	0.77
	Correlation	0.61	0.67	0.71	0.75
	Cosine	0.59	0.65	0.69	0.71
		Specificity =			
	Metric	**0.90**	**0.80**	**0.70**	**0.60**
Sensitivity	Yule	0.40	0.66	0.84	0.91
	Yule2	0.40	0.68	0.84	0.91
	Correlation	0.43	0.66	0.83	0.91
	Cosine	0.41	0.63	0.80	0.91

4.3. CONCLUSION

In this chapter, the performance of 4 weighting schemes and 42 similarity/distance metrics was evaluated based on their ability to separate paired documents from unpaired documents. The key results of the separation tests are listed as follows:

1. In terms of S_{MD}, the weighting scheme, BO, performed better than TO, TF and TFIDF,
2. The binary similarity metric, Yule, gave the highest S_{MD}, and performed better than the common similarity metrics,
3. Interventions derived from "Full abstract" and "Titles + Full abstracts" performed the best, in terms of S_{MD} and one-way ANOVA analysis,
4. "Full abstracts" provide crucial [I] and [C] elements (i.e. medical interventions) for similarity measurement,
5. The top similarity metrics, as measured by S_{MD}, include: Yule, Yule2, Stiles, Simpson, Ochiai, Fager, Dice, Kulzynski2, Pearson, Correlation and Cosine,

6. Yule and Yule2 gave a more clear cut separation in histograms with minor overlap in high similarity region,
7. The top four similarity metrics, based on the ROC curves, include: Yule, Yule2, Cosine and Correlation, and
8. Yule and Yule2 showed a slightly higher sensitivity and specificity than Cosine and Correlation in correctly identifying paired and unpaired documents.

Among the top four similarity metrics found in this study, the two metrics, Cosine and Correlation, have commonly been used in document clustering and short-text clustering (Huang, 2008; Subhashini and Kumar, 2010; Rangrej et al., 2011; Lin et al., 2013). It was shown in this study that the Yule and Yule2 similarity metrics performed better than Cosine and Correlation metrics. Though not as well studied as the common similarity metrics, an improvement in clustering performance using the Yule metric was reported by Malik and Kender (2006). On the other hand, abstracts provide more detail about the contents of a document than the titles alone. Therefore, PICO elements should be extracted from "Full abstracts" or "Titles + Full abstracts".

To conclude, the overall results support the combination of the weighting scheme, BO, the binary similarity metrics, Yule or Yule2, and the interventions derived from "Full abstract" or "Titles + Full abstracts" for concept-based similarity between documents. The results obtained from this chapter were used to group documents into different clusters based on the similarity of medical interventions that they contain and for the subsequent visualization of the most useful and important medical concepts (i.e. the PICO elements) in the answer field of the proposed clinical question answering engine. A detailed description of the clustering analysis is presented in **Chapter 5**.

5. CHAPTER V: Cluster Structure Analysis

As described in **Chapter 4**, each document was represented as a bag of medical interventions. In this chapter, using both poorly-formulated and well-formulated therapy question, the similarities between documents were calculated using top similarity metrics (Cosine, Correlation, Yule and Yule2) and were then clustered using different agglomerative hierarchical clustering algorithms (Complete-link, Average link and Ward-link). If a hierarchy is too narrow and deep, users will have to click through an inordinate number of levels to reach the topic of interest. Relatively, if a hierarchy is too flat, a parent cluster will contain many child nodes that may increase the time and difficulty for users to define their topics of interest. Therefore, the purpose of this chapter is to identify the most appropriate hierarchical structure to cluster and visualize a collection of documents for browsing, searching and exploring purposes. This was achieved by:

1. Exploring the number of hierarchy levels in different similarity-based hierarchies,
2. Identifying the average location of the best clusters, i.e. clusters with high precision and high recall, by hierarchy level,
3. Calculating the average percentage of relevant documents in the best clusters by expanding the hierarchies level by level, and
4. Measuring the performance of different similarity-based hierarchies in visualizing documents relevant to a set of therapy topics using standard information retrieval metrics such as mean average precision and precision at k.

Section 5.1 presents the methodologies used for cluster structure analysis. The results of the analysis are presented and discussed in **Section 5.2**.

5.1. METHODOLOGY

5.1.1. Collection of Test Questions

Cao et al. (2011) selected 60 questions randomly from the ClinicalQuestions Collection (US National Library of Medicine). The authors aimed at investigating the performance of AskHERMES in answering long and complex questions. Demner-Fushman et al. (2006), on the other hand, evaluated the performance of CQA-1.0 using 30 questions of the type "*What is the current opinion on the best pharmacotherapy for disease X?*" in the June 2004 issue of Clinical Evidence. Previous studies suggest that

the completeness of PICO elements in a question determines whether it is likely to be answered (Bergus et al., 2000; Staunton, 2007). The AskHERMES and CQA-1.0 systems were not evaluated based on the completeness of PICO elements in a clinical question. Therefore, in this chapter, a total of 100 therapy questions were classified based on the completeness of PICO elements that they contain into 50 "poorly-formulated" and 50 "well-formulated" questions.

The first set of questions is maintained by the NLM and can be downloaded from the ClinicalQuestions Collection (US National Library of Medicine). A total of 50 questions that contain only one or two PICO elements were collected and are defined as "poorly-formulated". For instance, the question "*What is the treatment for hyperthyroidism due to Grave's disease?*" can be broken down as follows:

[P]: hyperthyroidism due to Grave's disease
[I]: -
[C]: -
[O]: -

The second set of questions is derived from an EBM database called Essential Evidence Plus (2015). A total of 50 questions that contain three to four PICO elements were collected and are considered as "well-formulated". For example, the question "*Are epidural corticosteroid injections effective in decreasing pain and improving function in patients with sciatica?*" can be broken down as follows:

[P]: patients with sciatica
[I]: epidural corticosteroid injections
[C]: -
[O]: decreasing pain and improving function

The results obtained using the two sets of questions were compared in this chapter. The purpose is to investigate the effects of the number of PICO elements in a question to the retrieval of relevant documents and the construction of similarity-based hierarchies.

5.1.2. Construction of Hierarchy

Agglomerative hierarchical clusterings differ in the metric used to compute the distance between interventions and the linkage method used to determine the distance

between two clusters. Based on the results identified from **Chapter 4**, interventions were extracted from the "Titles and Abstracts" of MEDLINE documents. Each document was represented by a bag of interventions. A matrix that contains the distances between all the documents was created respectively using the top similarity metrics: Cosine, Correlation, Yule and Yule2. The distance matrix was used as the input of a hierarchical clustering algorithm. Similar interventions were clustered together using three clustering algorithms implemented in the "hclust" function in the R "stats" package: average-link (AL), complete-link (CL) and ward-link (WL). The clustering algorithms can be described using the Lance-Williams dissimilarity update formula (Murtagh and Contreras, 2012). If two existing clusters, C_i and C_j, are merged to form a new cluster, C_{ij}, the dissimilarity (or distance) d between the new cluster and any existing cluster C_k is given by:

$$d_{C_{ij}C_k} = \propto_i d_{C_iC_k} + \alpha_j d_{C_jC_k} + \beta\, d_{C_iC_j} + \gamma \left| d_{C_iC_k} - d_{C_jC_k} \right|$$

where the values of $\propto$, β and γ are dependent on the clustering strategy presented in **Table 5-1**. The resulting hierarchical clusters were displayed as dendrograms using the "plot" function in R. The heights of the dendrograms were adjusted and cut at a specified level using the "rank_branches" function in the R "dendextend" package and the "cut" function in the R "stats" package, respectively (Galili, 2014).

Table 5-1. Parameters in the Lance-Williams update formula for three clustering methods.[t]

	Parameter			
Strategy	$\propto_i$	$\propto_j$	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$
Average-link	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	0	0
Complete-link	0.5	0.5	0	0.5
Ward-link	$\frac{n_i + n_k}{n_i + n_j + n_k}$	$\frac{n_j + n_k}{n_i + n_j + n_k}$	$-\frac{n_k}{n_i + n_j + n_k}$	0

[t.] n_i, n_j and n_k are the number of interventions in C_i, C_j and C_k respectively.

12 types of hierarchies were computed using the following combinations of similarity metrics and clustering methods:

Correlation – AL Cosine – AL Yule – AL Yule2 – AL

Correlation – CL	Cosine – CL	Yule – CL	Yule2 – CL
Correlation – WL	Cosine – WL	Yule – WL	Yule2 – WL

The hierarchies are intended to be used to group a collection of documents into meaningful clusters and to visualize the medical interventions relevant to a given query. The performance of each hierarchy was assessed by identifying the number of levels and the number of documents that a user will need to explore to obtain all the relevant documents for a total of 1000 test topics. Questions formulated with greater number of PICO elements allow a more precise search. Therefore, an average of 5 topics was identified from each of the hierarchies generated using well-formulated questions (5 topics x 50 questions = 250 topics). In contrast, questions formulated with lower number of PICO elements results in documents of a wide range of topics. Therefore, an average of 15 topics was identified from each of the hierarchies generated using poorly-formulated questions (15 topics x 50 questions = 750 topics). The test topics were selected randomly from their respective clusterings. A sample of the hierarchy is shown in **Figure 5-1** and can be explained as follows:

1. Each bag of interventions (e.g. “Thyroid drug - Irradiation”) represents a single document (as indicated with a PMID number).
2. Each [I] or [C] element (e.g. “Irradiation”, “Propylthiouracil” and “Thiamazole”) is referred to as a topic.
3. Similar elements are grouped under the same clusters. For example, the topic “Methylprednisolone” is grouped in a cluster of five articles at Level 1.
4. The similarity between interventions becomes stronger as the number of level increases. For instance, the interventions “Methylprednisolone pulse therapy-Glucocorticoid - Alendronate” are located in a cluster of two articles at Level 3.

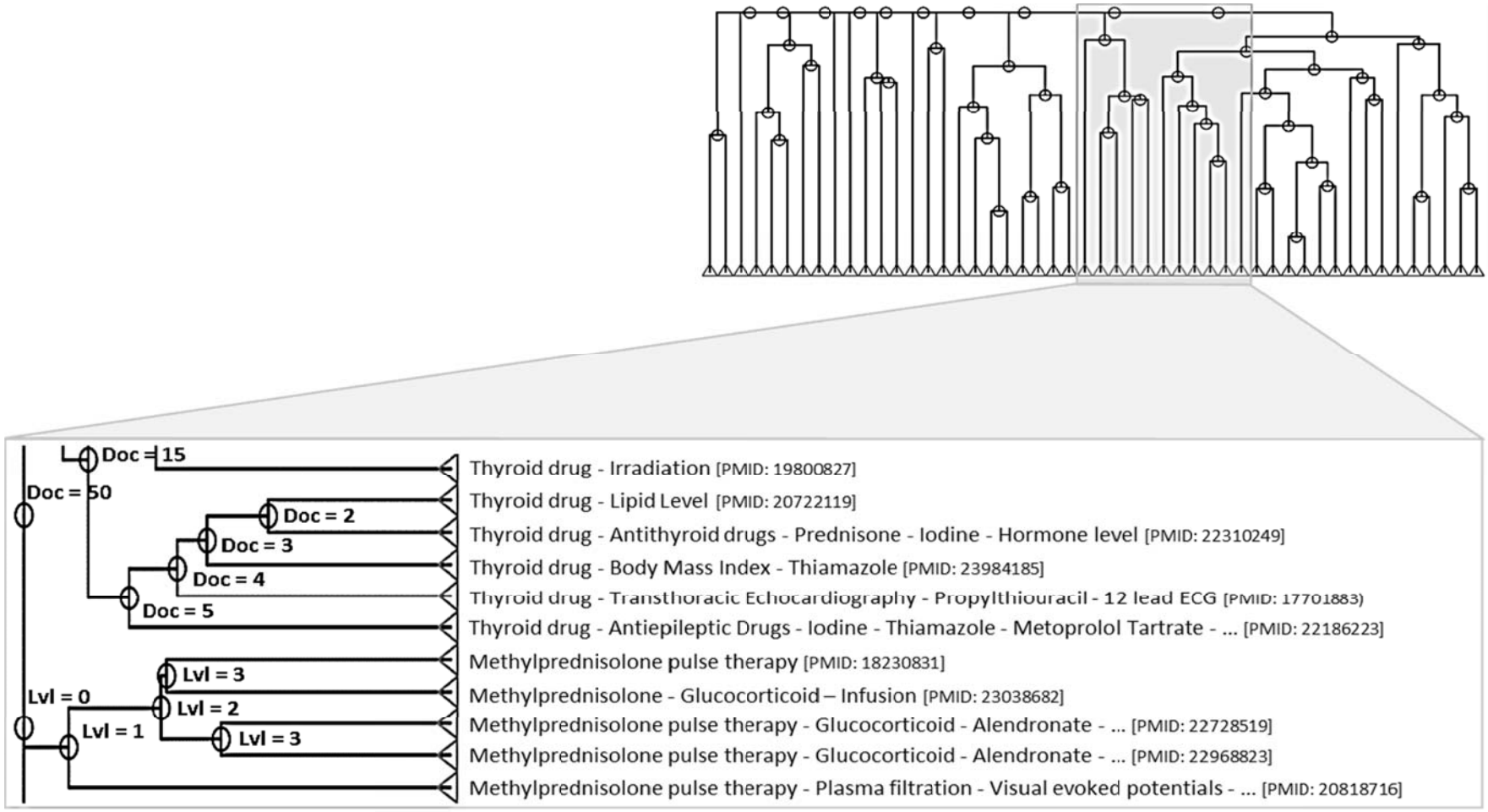


Figure 5-1. An example of agglomerative hierarchical clustering.

(Doc = No. of documents, Lvl = Hierarchy level)

5.1.3. Identification of Best Clusters

The precision, recall and F-measure of each cluster were calculated. Precision is the ratio of relevant documents retrieved for a given topic (N_{Rel}) over the total number of relevant and irrelevant documents retrieved ($N_{Rel} + N_{Irrel}$).

$$Precision\ (P) = \frac{N_{Rel}}{N_{Rel} + N_{Irrel}}$$

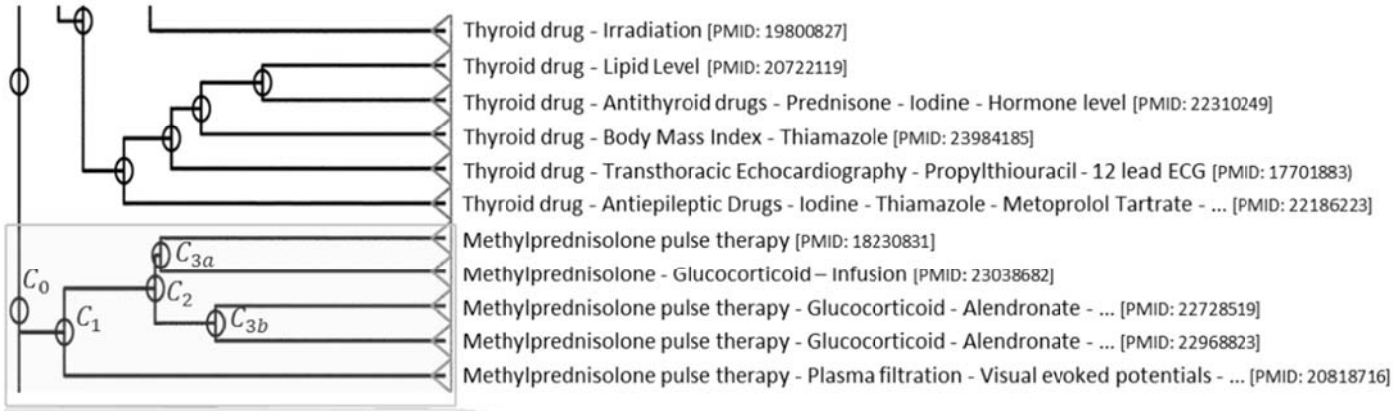
Recall is the ratio of relevant documents retrieved for a given topic (N_{Rel}) over the total number of relevant documents retrieved and not retrieved ($N_{Rel} + M_{Rel}$). The actual number of relevant documents was determined by two human raters, as described in **Chapter 4 Section 4.1.6**.

$$Recall\ (R) = \frac{N_{Rel}}{N_{Rel} + M_{Rel}}$$

A good cluster is supposed to contain as many relevant documents as possible with high precision and high recall. The F-measure is the harmonic mean of precision and recall.

$$F - measure\ (F) = 2 \times \frac{P \times R}{P + R}$$

Continuing the example given in **Figure 5-1**, **Figure 5-2** shows two examples of how the best clusters were identified from the hierarchy. As shown in the figure, Topic 1 ("Methylprednisolone") and Topic 2 ("Alendronate") are grouped under C_1. The best cluster is determined by the highest F-measure. Suppose that the actual number of documents relevant to Topic 1 and Topic 2 are 5 and 2 respectively out of a total of 50 documents. Topic 1 is best represented by C_1 at Level 1 with the highest precision, recall and F-measure. 5 out of the 5 documents relevant to Topic 1 appear in C_1. Topic 2 is best represented by C_{3b} at Level 3. As exemplified by the examples, the precision increases and the recall decreases with an increase in number of hierarchy level, and the F-measure quantifies the balance between precision and recall. The same analysis was performed on 1000 test topics in order to identify the average location of the best clusters in 12 types of similarity-based clusterings by hierarchy level.



	Cluster	N_{Rel}	$N_{Rel} + N_{Irrel}$	$N_{Rel} + M_{Rel}$	Precision	Recall	F-measure
Topic 1:	C_0	5	50	5	0.10	1.00	0.18
Methylprednisolone	**C_1**	**5**	**5**	**5**	**1.00**	**1.00**	**1.00**
	C_2	4	4	5	1.00	0.80	0.89
	C_{3a}	2	2	5	1.00	0.40	0.57
	C_{3b}	2	2	5	1.00	0.40	0.57
Topic 2:	C_0	2	50	2	0.04	1.00	0.08
Alendronate	C_1	2	5	2	0.40	1.00	0.57
	C_2	2	4	2	0.50	1.00	0.67
	C_{3a}	0	2	2	0.00	0.00	0.00
	C_{3b}	**2**	**2**	**2**	**1.00**	**1.00**	**1.00**

Figure 5-2. Identification of the best clusters.

(N_{Rel} = No. of relevant documents, $N_{Rel} + N_{Irrel}$ = Sum of relevant and irrelevant documents, $N_{Rel} + M_{Rel}$ = Sum of relevant documents retrieved and not retrieved)

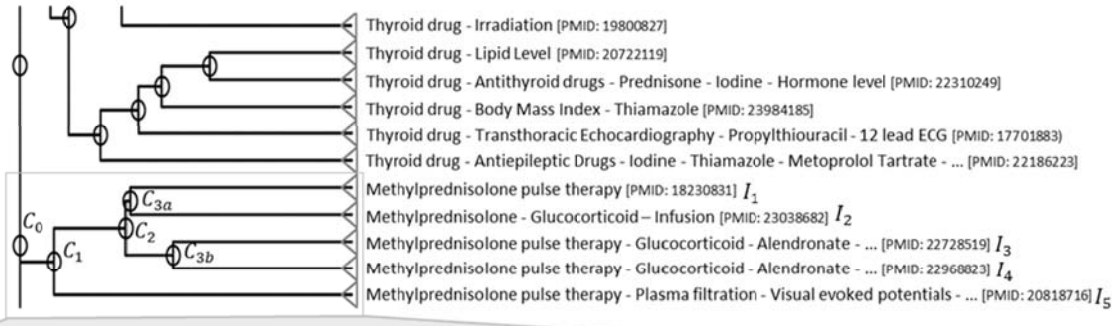
5.1.4. Percentage of Relevant Documents

This section aims to identify the percentage of documents relevant to a test topic (% Rel) in the best cluster. The purpose is to determine the number of hierarchy levels that should be expanded to obtain a certain amount of relevant documents from different similarity-based clusterings. If N_{Rel} and N_{Irrel} are the numbers of relevant and irrelevant documents respectively in a cluster, the % Rel is calculated by:

$$\% \, Rel = \frac{N_{Rel}}{N_{Rel} + N_{Irrel}} \times 100\%$$

An example of the calculation is given in **Figure 5-3**. As shown in the figure, the cluster that best represents a topic, as determined by the highest F-measure, is identified level by level by increasing the number of branches of a hierarchy from Structure A to Structure C. Structure A with one level depth supports the exploration of Topic 1 ("Methylprednisolone") but not for Topic 2 ("Alendronate") (% Rel of $C_1 = 100\%$ and 40%, respectively). The percentage of relevant documents increases with an increase in number of levels. Half of the documents included in C_2 of Structure B are irrelevant to Topic 2. However, by dividing C_2 to C_{3a} and C_{3b}, 2 out of 2 of the documents relevant to Topic 2 are grouped into C_{3b} of Structure B. C_{3b} of Structure C gives % Rel of 100% for both Topic 1 and Topic 2, suggesting that the two topics are best presented by a hierarchy with a depth of three levels.

The same analysis was performed on 1000 test topics and the average percentages of relevant documents (% $\overline{Rel}$) at different hierarchy levels were calculated to compare the overall performance of 12 types of similarity-based hierarchies.



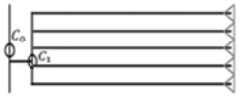
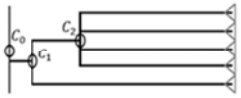
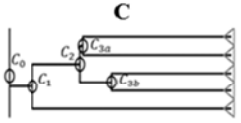
			Topic 1: Methylprednisolone				Topic 2: Alendronate			
Level	**Structure**	**Cluster**	**F-measure**	N_{Rel} :	$N_{Rel} + N_{Irrel}$	**% *Rel***	**F-measure**	N_{Rel} :	$N_{Rel} + N_{Irrel}$	**% *Rel***
1	**A** 	C_0	0.18	5 :	50	10	0.08	2 :	50	4
		C_1	1.00	5 :	5	100	0.57	2 :	5	40
2	**B** 	C_0	0.18	5 :	50	10	0.08	2 :	50	4
		C_1	1.00	5 :	5	100	0.57	2 :	5	40
		C_2	0.89	4 :	4	100	0.67	2 :	4	50
3	**C** 	C_0	0.18	5 :	50	10	0.08	2 :	50	4
		C_1	1.00	5 :	5	100	0.57	2 :	5	40
		C_2	0.89	4 :	4	100	0.67	2 :	4	50
		C_{3a}	0.57	2 :	2	100	0.00	0 :	2	0
		C_{3b}	0.57	2 :	2	100	1.00	2 :	2	100

Figure 5-3. Percentage of relevant documents by hierarchy level.

(N_{Rel} = No. of relevant documents, $N_{Rel} + N_{Irrel}$ = Sum of relevant and irrelevant documents, % Rel = Percentage of relevant documents)

5.1.5. Visualization Performance

In **Sections 5.1.3** and **5.1.4**, the general location of the best clusters and the number of hierarchy levels that should be expanded were explored by identifying the number of relevant documents in each cluster. In this section, a hierarchy was expanded level by level in a top-down manner, and the proportion of relevant documents visualized for a given topic was identified. For instance, as shown in **Figure 5-3**, I_1 to I_5 are five documents relevant to the topic "Methylprednisolone". Starting from the top of the hierarchy, by clicking C_1, I_5 is visualized whereas I_1 to I_4 grouped in C_2 are invisible. A user will have to click C_2 and its child clusters, C_{3a} and C_{3b}, to make them visible. In other words, a user will have to explore three levels to collect the five documents. As shown in **Table 5-2**, the documents visualized and not visualized can be presented as a list of document based on the number of hierarchy level that has been expanded. A score of strength of evidence described in **Chapter 6 Section 6.3.3** was assigned to each document to ensure that multiple lists of documents for the same query were ranked similarly. The performance of a hierarchy in visualizing a set of test topics was evaluated using the trec_eval[12] program. Each topic was treated as a query and two input files were passed to the trec_eval program: "trec_top_file" and "trec_rel_file".

1. "trec_top_file" contained a ranked list of documents visualized and not visualized by a hierarchy.
2. "trec_rel_file" is a list of documents judged by human raters as relevant or non-relevant to a given query, as described previously in **Chapter 4 Section 4.1.6**.

Table 5-2. Lists of documents visualized by expanding a hierarchy level by level.[t]

Rank	Level 0	Level 1	Level 2	Level 3
I_1	0	0	0	1
I_2	0	0	0	1
I_3	0	0	0	1
I_4	0	0	0	1
I_5	0	1	1	1

[t.] A value of 0 indicates a document is visible whereas 1 indicates a document is hidden.

[12] The trec_eval program supplied by TREC (Text Retrieval Conference) is designed for evaluating the information retrieval of an information retrieval system or program (Voorhees, 2003).

For each query, the following information retrieval metrics were computed as the output of the program: average precision, 11-point interpolated precision, precision at k and R-precision.

Mean Average Precision (MAP).

The average precision (AP) for a single topic was computed by averaging the precision values calculated after each relevant document is visualized. This was performed on a set of 750 topics. The AP scores of these topics were averaged to derive mean average precision (MAP), a single measure of the overall quality of a hierarchy. Fixed recall levels were not chosen for MAP and there is not interpolation.

Suppose that 20 documents are retrieved for a query Q, in which 7 are known to be relevant to a topic i. As shown in **Table 5-3**, recall and precision are calculated each time a relevant document is visualized. A value of 0 is assigned to any relevant documents not visualized. The AP for i is calculated as follow

$$AP_i = \frac{\frac{1}{1}+\frac{2}{3}+\frac{3}{5}+\frac{4}{6}+\frac{5}{9}+\frac{6}{14}+\frac{7}{20}}{7} = 0.61$$

If n is a batch of topics relevant to Q, MAP is the average of AP across different recall levels and over all the test topics evaluated.

$$MAP = \frac{1}{n}\sum_{n} AP_n$$

11-Point Interpolated Average Precision.

For each topic, this metric computes precision at 11 levels of recall: 0.0, 0.1, …, 0.9 and 1.0. As presented in **Table 5-3**, the first document is relevant. The recall and precision are $\frac{1}{7}$ $(= 0.14)$ and $\frac{1}{1}$ (= 1.0) respectively. This value is entered into **Table 5-4** for the recall level of 0.1 and is interpolated back from the recall level of 0.1 to the 0.0 level. The recall of the third document is $\frac{2}{7}$ $(= 0.28)$ with precision of $\frac{2}{3}$ $(= 0.67)$. The value is entered into the recall level of 0.2. The fourth document is not relevant. The fifth document increases the recall level to $\frac{3}{7}$ $(= 0.43)$ with new precision of $\frac{3}{5}(= 0.60)$. The precision value at the 0.4 level is interpolated back to the 0.3 level. Therefore, the interpolated precision is defined as the maximum precision for a given recall level. The

measurement was performed across 750 topics. The arithmetic mean of the interpolated precisions at each recall level was calculated to compare the performance of different similarity-based hierarchies.

Table 5-3. Recall and precision of 20 documents with 7 known to be relevant

Rank	Relevance[t]	Recall	Precision
1	Rel	1/7	1/1
2	NRel		0
3	Rel	2/7	2/3
4	NRel		0
5	Rel	3/7	3/5
6	Rel	4/7	4/6
7	NRel		0
8	NRel		0
9	Rel	5/7	5/9
10	NRel		0
11	NRel		0
12	NRel		0
13	NRel		0
14	Rel	6/7	6/14
15	NRel		0
16	NRel		0
17	NRel		0
18	NRel		0
19	NRel		0
20	Rel	7/7	7/20

[t] Rel = relevant document, NRel = nonrelevant document.

Table 5-4. 11-point interpolated precision

Recall	Precision
0.0	1.00
0.1	1.00
0.2	0.67
0.3	0.60
0.4	0.60
0.5	0.67
0.6	0.56
0.7	0.56
0.8	0.43
0.9	0.35
1.0	0.35

Precision at k (P@k).

$P@k$ is the precision of relevant documents after k documents have been visualized. For example, a topic has 20 relevant documents. **Table 5-5** shows the top 20 documents visualized by a hierarchy. The $P@5$, $P@10$, $P@15$ and $P@20$ of the topic are $\frac{3}{5} = 0.60$, $\frac{5}{10} = 0.50$, $\frac{6}{15} = 0.40$ and $\frac{7}{20} = 0.35$, respectively. End-users generally look for the first 10 or 20 documents retrieved only (Wang et al., 2004). The metric has the advantage of measuring precision at fixed low levels of retrieved results such as 10 or 20 documents.

Table 5-5. Precision at fixed document cut-off value

k	Relevance	$P@k$
1	Rel	1/1
2	NRel	1/2
3	Rel	2/3
4	NRel	2/4
5	Rel	3/5
6	Rel	4/6
7	NRel	4/7
8	NRel	4/8
9	Rel	5/9
10	NRel	5/10
11	NRel	5/11
12	NRel	5/12
13	NRel	5/13
14	Rel	6/14
15	NRel	6/15
16	NRel	6/16
17	NRel	6/17
18	NRel	6/18
19	NRel	6/19
20	Rel	7/20

Average R-Precision (ARP).

If R is the number of relevant documents for a topic i, R-precision (RP) is the precision after R documents have been visualized. Average R-precision (ARP) is the arithmetic mean of the RP values over a batch of topics. It can be expressed as follows:

$$ARP = \frac{1}{n}\sum_{n} RP_n$$

where n is the number of topics relevant to a query Q. As an example, assume that a run consists of two topics (Topic A and Topic B), which have 10 and 20 relevant documents respectively. If 7 out of the top 10 documents visualized are relevant to Topic A, and 15 out of the top 20 documents visualized are relevant to Topic B, then the run's ARP is calculated by averaging the RP values of the two topics.

$$ARP = \frac{1}{2}\left(\frac{7}{10} + \frac{15}{20}\right) = 0.73$$

5.2. RESULTS and DISCUSSION

The results obtained using poorly-formulated questions are presented and discussed in **Sections 5.2.1** to **5.2.4**, and are compared to the results obtained using well-formulated questions in **Section 5.2.5**.

5.2.1. Structure of Hierarchies

The structures of (50 poorly-formulated questions × 3 clustering methods × 4 similarity metrics) = 600 hierarchies were evaluated. **Figure 5-4** shows the general structures of Cosine-based and Yule2-based clusterings generated using average-link (AL), complete link (CL) and ward-link (WL) methods. A cluster is assumed to be stable if it has lower than 7 child nodes. The purpose is to simplify the description of the structures of different clusterings.

- For AL clusterings, the root nodes split from the top left of the hierarchies into a stable cluster (to the left) and a large cluster (to the right). A stable cluster is split off gradually from their parent clusters until two stable clusters are formed at the very right side of the hierarchies. This indicates that, by exploring the hierarchies level by level, the interventions are visualized gradually from the left side to the right side of the hierarchies. A longer time is required to visualize all the interventions or to obtain the topics located at the right bottom of the hierarchies.
- The roots of CL clusterings are wide. For example, 11 out of the 49 nodes are aligned on the root of the Cosine-CL clustering. This reduces the number of child nodes appear under the root nodes. The root nodes are split into multiple branches, producing stable child nodes which are distributed evenly across the

hierarchies. Compared to AL clusterings, a shorter time is required to search for a topic in CL clusterings as the number of nodes or branches are lower.

- Compared to AL clusterings, the root nodes of WL clusterings split into two parent nodes with child nodes distributed more evenly to the left and right sides of the hierarchies. Besides, a lower number of hierarchy levels are found on the upper levels of the hierarchies. Compared to CL clusterings, each of the WL clusterings has only a single root node. The root node splits gradually from two large clusters to multiple stable clusters. Additionally, the branches of WL clusterings are located mainly at the bottom of the hierarchies.

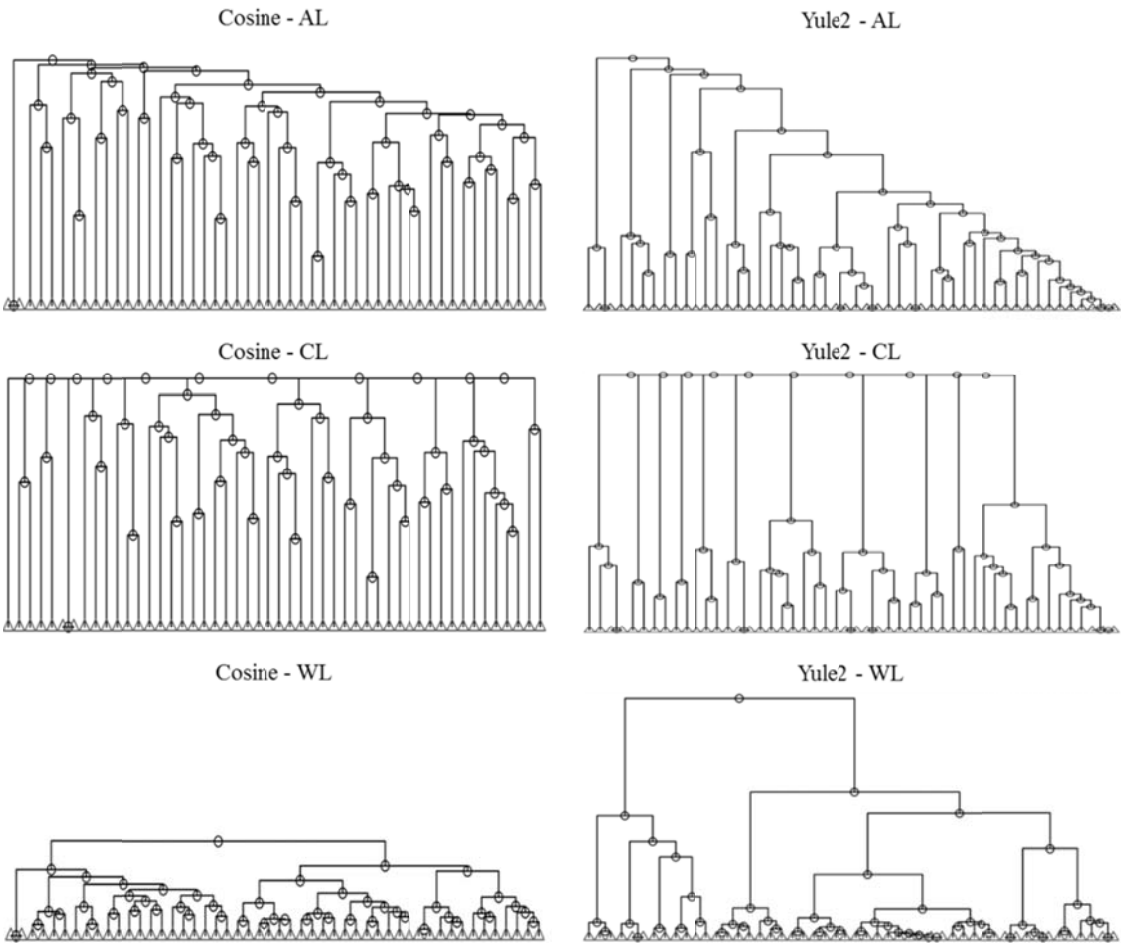


Figure 5-4. General structures of average-link (AL), complete-link (CL) and ward-link (WL) hierarchical clusterings.

The results suggest that either CL or WL algorithm is more suitable for visualizing a hierarchy of medical interventions. In addition, the greater the number of hierarchy levels, the longer it takes for a user to browse and search for a therapy topic in a hierarchy. **Table 5-6** presents the average number of hierarchy levels, $\overline{HLvl}$, in each of

the 12 types of hierarchies. An analysis of the different similarity-based clusterings showed that WL algorithm produced the lowest number of levels, followed by CL algorithm and the highest by AL algorithm. A slightly lower number of levels were found in Yule-WL and Yule2-WL clusterings ($\overline{HLvl}$ = 11 ± 3.57 and 11 ± 3.17, respectively) than Cosine-WL clusterings ($\overline{HLvl}$ = 12 ± 4.12). The overall results suggest that WL algorithm produces a more appropriate hierarchical structure than AL and CL algorithms, in terms of the number of levels that a user will need to explore during the search process.

Table 5-6. Average number of hierarchy levels ($\overline{HLvl}$) in 12 types of hierarchies.

	Clustering algorithm		
Similarity metric	**AL**	**CL**	**WL**
Cosine	19± 6.20	16 ± 5.83	12 ± 4.12
Correlation	25± 7.90	15 ± 3.95	13 ± 3.87
Yule	20± 5.90	14 ± 4.27	11 ± 3.57
Yule2	20± 5.50	14 ± 4.10	11 ± 3.17

5.2.2. Location of Best Clusters

The precision (P), recall (R) and F-measure (F) of each cluster were calculated to identify the clusters that best represent 750 topics selected randomly from the 600 hierarchies generated using poorly-formulated questions. The best cluster was determined by identifying the cluster that yielded the highest F-measure (F_{max}). The higher the F value, the greater the quality of a cluster. **Table 5-7** shows the locations of the best clusters for 15 test topics in a Correlation-AL clustering. It can be seen from the table that:

1. The maximum number of levels in the hierarchy is 23.
2. Relevant documents are grouped in one (e.g. Level 1 of Topic 1, R = 1.00) or two clusters (e.g. Level 5 of Topic 2, R = 0.33 and 0.67 respectively),
3. The best clusters appear at Levels 15, 5, 10 and 19 respectively for Topics 1, 2, 3 and 15 (F_{max} = 0.69, 0.80, 1.00 and 0.67, respectively), and
4. The best clusters contain all of the relevant documents (e.g. Level 10 of Topic 3, P = 1.00 and R = 1.00) or part of the relevant documents (e.g. Level 19 of Topic 15, P = 0.50 and R = 1.00).

Table 5-7. Distribution of best clusters in a Correlation-AL.

Topic	Hierarchy Level	Precision (*P*)	Recall (*R*)	F-measure (*F*)
1	0	0.32	1.00	0.48
	1	0.33	1.00	0.49
	⋮	⋮	⋮	⋮
	11	0.46	0.69	0.55
	⋮	⋮	⋮	⋮
	15	0.69	0.69	0.69
	⋮	⋮	⋮	⋮
	23	0.5	0.06	0.11
2	0	0.06	1.00	0.11
	1	0.06	1.00	0.12
	⋮	⋮	⋮	⋮
	4	0.08	1.00	0.14
	5	0.03	0.33	0.05
	5	1.00	0.67	0.80
	⋮	⋮	⋮	⋮
	23	0.00	0.00	0.00
3	0	0.10	1.00	0.18
	⋮	⋮	⋮	⋮
	7	0.14	1.00	0.25
	⋮	⋮	⋮	⋮
	10	1.00	1.00	1.00
	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
15	0	0.02	1.00	0.04
	1	0.02	1.00	0.04
	⋮	⋮	⋮	⋮
	19	0.50	1.00	0.67
	⋮	⋮	⋮	⋮
	23	0.00	0.00	0.00

The locations of the best clusters for the same topics in a Yule-CL and a Cosine-WL clusterings were discovered and are shown in **Table 5-8** and **Table 5-9** respectively. The three tables show that the best clusters located at different levels of the hierarchies. For instance, the best clusters for Topic 1 are located at Levels 15, 5 and 3 respectively of the AL, CL and WL clusterings.

Table 5-8. Distribution of best clusters in a Yule-CL clustering.

Topic	Hierarchy Level	Precision (*P*)	Recall (*R*)	F-measure (*F*)
1	0	0.32	1.00	0.48
	1	0.33	1.00	0.50
	⋮	⋮	⋮	⋮
	3	0.36	0.94	0.52
	4	0.37	0.94	0.53
	5	0.38	0.94	0.54
	⋮	⋮	⋮	⋮
	17	0.05	0.06	0.11
2	0	0.06	1.00	0.11
	1	0.06	1.00	0.12
	⋮	⋮	⋮	⋮
	6	0.08	1.00	0.15
	7	0.03	0.33	0.05
	7	1.00	0.67	0.80
	⋮	⋮	⋮	⋮
	17	0.50	0.33	0.40
3	0	0.02	1.00	0.04
	1	0.05	1.00	0.04
	2	0.02	1.00	0.04
	3	0.20	1.00	0.33
	4	0.50	1.00	0.67
	⋮	⋮	⋮	⋮
	17	0.00	0.00	0.00
⋮	⋮	⋮	⋮	⋮
15	0	0.02	1.00	0.04
	1	0.02	1.00	0.04
	⋮	⋮	⋮	⋮
	10	0.14	1.00	0.25
	11	0.25	1.00	0.40
	12	0.50	1.00	0.67
	⋮	⋮	⋮	⋮
	17	0.00	0.00	0.00

Table 5-9. Distribution of best clusters in a Cosine-WL clustering.

Topic	Hierarchy Level	Precision (*P*)	Recall (*R*)	F-measure (*F*)
1	0	0.32	1.00	0.48
	1	0.39	0.75	0.51
	⋮	⋮	⋮	⋮
	2	0.42	0.69	0.52
	⋮	⋮	⋮	⋮
	3	0.52	0.69	0.59
	⋮	⋮	⋮	⋮
	8	0.50	0.06	0.11
2	0	0.06	1.00	0.11
	1	0.03	0.33	0.06
	⋮	⋮	⋮	⋮
	3	0.22	0.67	0.33
	⋮	⋮	⋮	⋮
	5	1.00	0.67	0.80
	⋮	⋮	⋮	⋮
	8	0.50	0.33	0.40
3	0	0.10	1.00	0.18
	1	0.16	1.00	0.28
	⋮	⋮	⋮	⋮
	2	1.00	1.00	1.00
	⋮	⋮	⋮	⋮
	3	1.00	0.60	0.75
	⋮	⋮	⋮	
	8	0.00	0.00	0.00
⋮	⋮	⋮	⋮	⋮
15	0	0.02	1.00	0.04
	1	0.03	1.00	0.06
	⋮	⋮	⋮	⋮
	3	0.05	1.00	0.09
	⋮	⋮	⋮	⋮
	6	0.50	1.00	0.67
	⋮	⋮	⋮	⋮
	8	0.00	0.00	0.00

As the best clusters located at different hierarchy levels, the following measurements were performed to determine the general location of the best clusters:

1. The hierarchy levels of the best clusters over the 750 topics were averaged. As presented in **Table 5-10**, the best clusters located on average at Level 4-5 of WL clusterings. Compared to WL clusterings, a greater number of levels would need to be discovered to reach the best clusters in CL and AL clusterings.
2. The percentages of best clusters at different ranges of hierarchy levels were calculated (**Figure 5-5**). More than 60% of the best clusters in WL clusterings and 30-35% of the best clusters in CL and AL clusterings were located at Level 0-5. Up to 98% of the best clusters were located at Levels 0-5 and 6-10 of WL clusterings. Compared to CL clusterings, higher percentages of best clusters were located at deeper levels of AL clusterings.
3. The percentages of best clusters were identified by expanding the hierarchies level by level. As illustrated in **Figure 5-6**, the top ten levels of WL clusterings contained the highest percentages of best clusters (above 90%). About 60-70% of the best clusters were identified from the top ten levels of AL and CL clusterings.

The overall results from this section showed that the best clusters were located on average at Level 4-5 and more than 90% of the clusters could be obtained from the top ten levels of the WL clusterings evaluated.

Table 5-10. Average location of best clusters by hierarchy level.

	Clustering algorithm		
Similarity metric	**AL**	**CL**	**WL**
Cosine	9.45 ± 4.25	8.40 ± 3.54	4.55 ± 2.20
Correlation	9.32 ± 5.21	8.20 ± 3.24	4.65 ± 2.51
Yule	9.10 ± 4.75	7.45 ± 3.45	4.20 ± 2.50
Yule2	9.05 ± 4.55	7.34 ± 3.25	4.15 ± 2.34

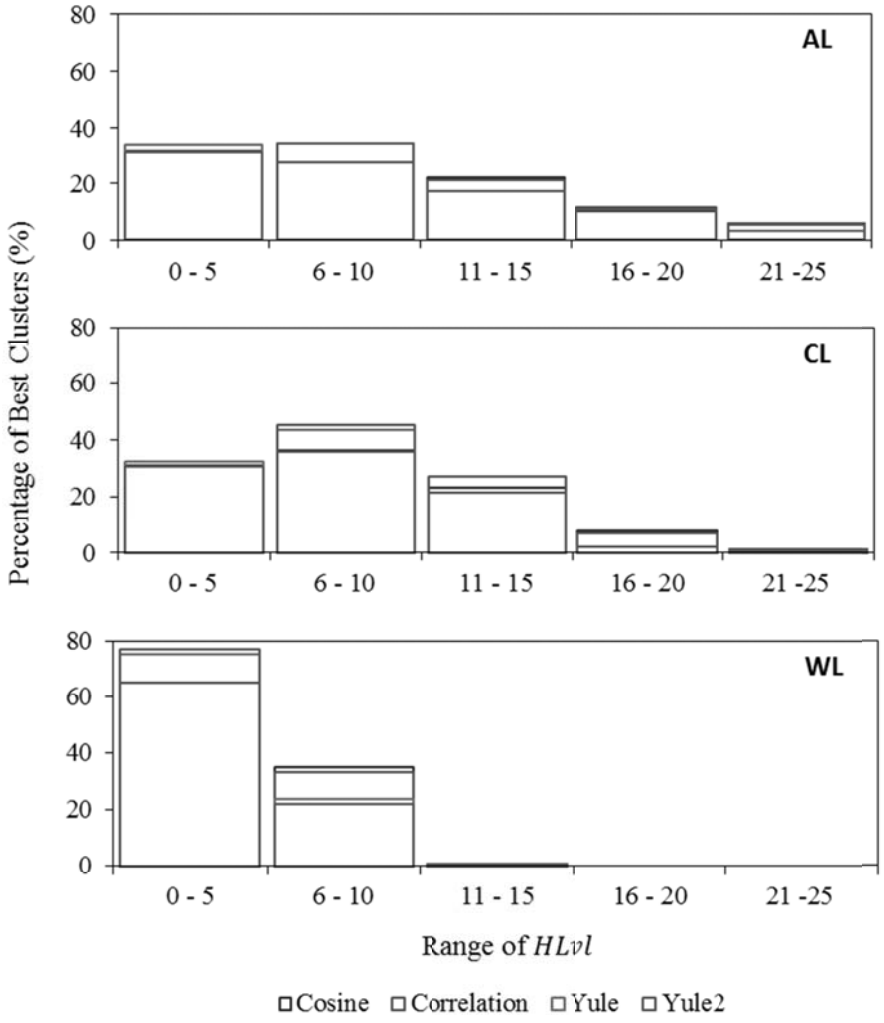


Figure 5-5. Percentage of best clusters by different ranges of hierarchy levels (*HLvl*).

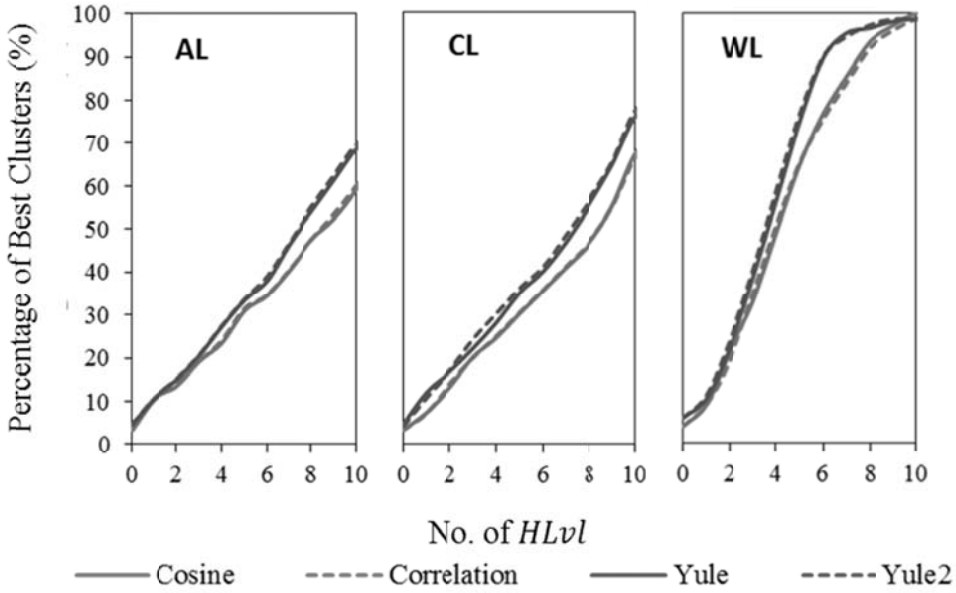


Figure 5-6. Percentage of best clusters located on the top ten hierarchy levels (*HLvl*).

5.2.3. Percentage of Relevant Documents

This section describes the extent to which a hierarchy should be expanded to facilitate the exploration of interventions in different clusters. The clusters that best represent a topic, as measured by F_{Max}, were identified hierarchically from the top 1, top 2, ... and top 10 levels. **Table 5-11** shows the percentage of relevant documents ($\%\ Rel$) in the best clusters when the depth of three Yule-based clusterings was increased from 1 to 10 levels. As shown in the table, the $\%\ Rel$ increases with an increase in hierarchy level. The $\%\ \overline{Rel_{15}}$ is the average percentage of relevant documents over 15 topics. The best clusters from the top 10 levels of Yule-AL, Yule-CL and Yule-WL clusterings contained, respectively, an average of 53%, 55% and 61% of relevant documents.

Table 5-11. Percentage of relevant document ($\%\ Rel$) in top 10 hierarchy levels of three Yule-based clusterings.

Clustering Algorithm	Hierarchy Level	$\%\ Rel$					$\%\ \overline{Rel_{15}}$
		Topic 1	Topic 2	Topic 3	…	Topic 15	
AL	1	33	6	11	…	2	8
	⋮	⋮	⋮	⋮		⋮	⋮
	4	33	8	13	…	3	17
	5	33	100	14	…	3	30
	6	33	100	14	…	3	38
	⋮	⋮	⋮	⋮		⋮	⋮
	9	48	100	100	…	4	52
	10	53	100	100	…	5	**53**
CL	1	33	6	10	…	2	14
	⋮	⋮	⋮	⋮		⋮	⋮
	4	37	7	12	…	2	25
	5	38	8	13	…	3	26
	6	38	8	14	…	3	26
	⋮	⋮	⋮	⋮		⋮	⋮
	9	38	100	26	…	10	50
	10	38	100	83	…	14	**55**
WL	1	32	8	13	…	6	13
	⋮	⋮	⋮	⋮		⋮	⋮
	4	52	100	83	…	5	53
	5	61	100	100	…	6	57
	6	61	100	100	…	8	58
	⋮	⋮	⋮	⋮		⋮	⋮
	9	61	100	100	…	50	61
	10	61	100	100	…	50	**61**

The clusters that best represent 750 topics were identified by expanding 600 hierarchies generated using poorly-formulated questions level by level. The average percentages of relevant documents in the best clusters (% $\overline{Rel_{750}}$) are given in **Table 5-12** and **Figure 5-7**. As shown in **Table 5-12**:

1. The best clusters identified from the top 5 levels of AL and CL clusterings contained approximately 30% of relevant documents, whereas for WL clusterings, more than 50% of relevant documents were found in the best clusters.
2. By expanding the hierarchies to a depth of 10 levels, about 70% of relevant documents were identified from the best clusters in Yule/Yule2-WL clusterings.

Table 5-12. Average percentage of relevant document over 750 topics (% $\overline{Rel_{750}}$) in top 10 hierarchy levels.

Clustering Algorithm	Hierarchy Level	% $\overline{Rel_{750}}$			
		Cosine	Correlation	Yule	Yule2
AL	1	17	16	18	18
	⋮	⋮	⋮	⋮	⋮
	4	24	23	25	26
	5	27	26	28	28
	6	30	30	31	32
	⋮	⋮	⋮	⋮	⋮
	9	42	41	45	46
	10	45	43	49	50
CL	1	17	17	18	18
	⋮	⋮	⋮	⋮	⋮
	4	26	25	29	30
	5	28	27	31	32
	6	30	28	33	34
	⋮	⋮	⋮	⋮	⋮
	9	43	43	50	51
	10	48	48	55	55
WL	1	21	21	22	22
	⋮	⋮	⋮	⋮	⋮
	4	48	50	52	54
	5	55	56	58	59
	6	60	62	63	64
	⋮	⋮	⋮	⋮	⋮
	9	64	65	68	69
	10	64	66	69	71

A further analysis of the WL clusterings revealed that the percentages of relevant documents improved weakly after Level 6, as illustrated in **Figure 5-7.** This can be explained by the findings in the previous section that the best clusters were located on

average at Level 4-5 of WL clusterings (refer to **Table 5-10**). Besides, the % $\overline{Rel_{750}}$ increased with an increase in hierarchy level. Compared to AL and CL clustering, a lower number of clusters would need to be further explored to obtain all the relevant documents from the WL clusterings. The results also suggest that a WL clustering should be expanded to a minimum depth of 5 levels to obtain more than 50% of relevant documents from the best clusters.

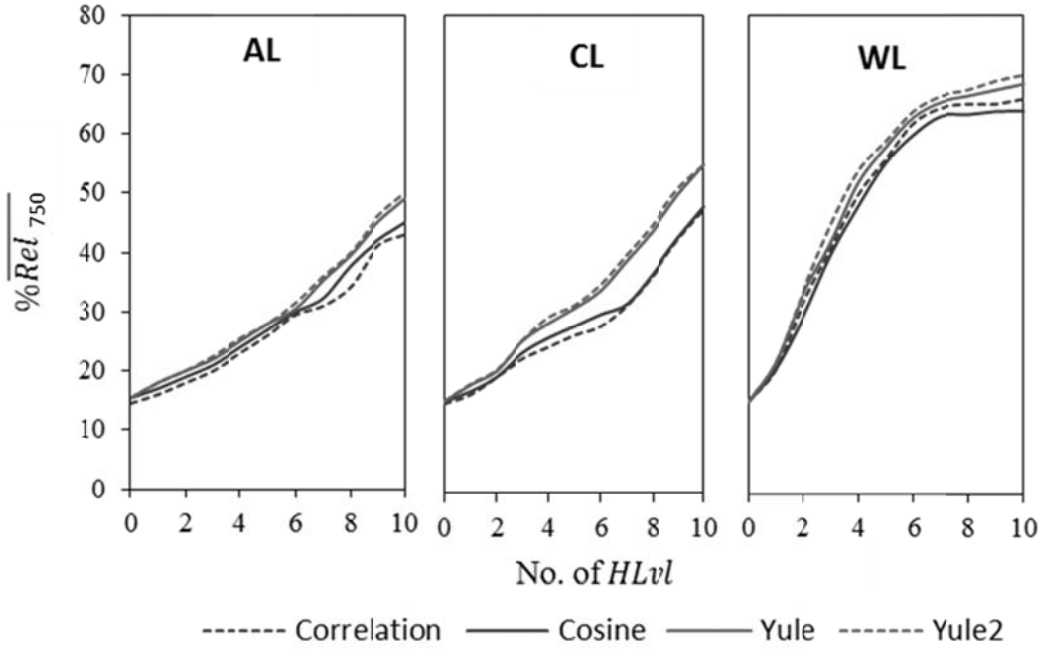


Figure 5-7. Average percentage of relevant documents over 750 topics (% $\overline{Rel_{750}}$) when the hierarchy level increased from 1 to 10.

5.2.4. Visualization Performance

The purpose of this section is to evaluate the performance of a hierarchy in visualizing the relevant documents for a set of test topics. The higher the MAP score, the higher the number of relevant documents that are visualized by a hierarchy. A MAP score of 1.00 indicates that all of the relevant documents are visualized by a hierarchy. As shown in **Figure 5-8**, the MAP increased with an increase in hierarchy level, and by expanding the hierarchies to a maximum depth of 11 levels, a more significant increase in MAP was achieved by WL clusterings than by AL and CL clusterings. When the hierarchies were expanded from a maximum of 6 to 11 levels, the MAP increased from 0.78 to 0.98 for Yule2-WL clusterings and from 0.77 to 0.96 for Yule-WL clusterings (**Table 5-13**).

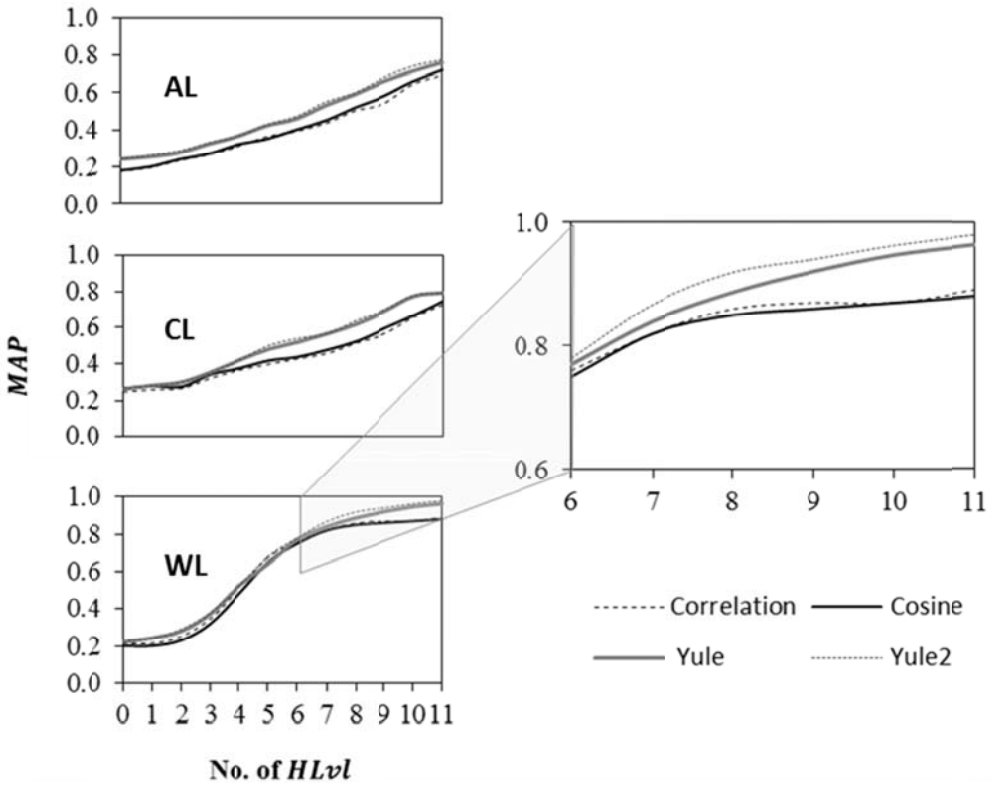


Figure 5-8. Mean average precision (MAP) of 750 topics for 12 types of hierarchies

Table 5-13. Mean average precision (MAP)

Hierarchy Level	Clustering algorithm	Similarity metric			
		Cosine	Correlation	Yule	Yule2
6	**AL**	0.40	0.39	0.46	0.47
	CL	0.44	0.43	0.52	0.54
	WL	0.75	0.76	0.77	0.78
11	**AL**	0.72	0.69	0.76	0.78
	CL	0.75	0.73	0.79	0.79
	WL	0.88	0.89	0.96	0.98

Similar to MAP, $\overline{P11}$ provides a statistical measure of the visualization performance of a hierarchy. The $\overline{P11}$ scores of WL clusterings at different hierarchy levels are given in **Table 5-14**. By expanding the hierarchies to a maximum of 11 levels, the $\overline{P11}$ scores of Yule-WL and Yule2-WL clusterings were close to 1.00, indicating that most of the relevant documents were visible.

Table 5-14. 11-point interpolated average precision ($\overline{P11}$). Clustering algorithm: WL

Hierarchy Level	Similarity metric			
	Cosine	**Correlation**	**Yule**	**Yule2**
1	0.21	0.22	0.25	0.25
2	0.24	0.26	0.29	0.30
3	0.33	0.36	0.38	0.39
4	0.49	0.52	0.53	0.53
5	0.66	0.69	0.65	0.69
6	0.76	0.77	0.78	0.79
7	0.83	0.83	0.85	0.88
8	0.86	0.87	0.90	0.93
9	0.87	0.88	0.93	0.95
10	0.88	0.89	0.96	0.97
11	0.89	0.90	**0.97**	**0.98**

As shown in **Table 5-15**, the highest P@5, P@10 and ARP were achieved by Yule-WL and Yule2-WL clusterings. Yule2-WL clustering yielded P@5 of 0.51. The value indicates that half of the first 5 documents visualized were relevant to the topics evaluated. Besides, an ARP of 0.98 indicates that a higher number of relevant documents were visualized by Yule2-WL clusterings, when compared to other types of clusterings.

Table 5-15. P@5, P@10 and Average R Precision (ARP)

Performance Indicator	Clustering Algorithm	Similarity metric			
		Cosine	**Correlation**	**Yule**	**Yule2**
P@5	AL	0.25	0.24	0.29	0.29
	CL	0.26	0.25	0.31	0.31
	WL	0.46	0.48	**0.50**	**0.51**
P@10	AL	0.17	0.15	0.21	0.21
	CL	0.19	0.18	0.23	0.23
	WL	0.32	0.34	**0.35**	**0.35**
ARP	AL	0.73	0.69	0.76	0.78
	CL	0.75	0.74	0.78	0.78
	WL	0.89	0.90	**0.97**	**0.98**

The overall results suggest that the combination of Yule/Yule2 similarity metric and WL clustering algorithm provides the best hierarchical structure for visualizing a hierarchy of medical interventions.

5.2.5. Poorly- *vs.* Well-Formulated Questions

The documents resulting from 50 well-formulated questions were collected. For each of the question, 3 hierarchies were computed using Yule2 similarity metrics and three clustering methods (AL, CL and WL). A total of 150 hierarchies were evaluated. The number of hierarchy levels in each of the hierarchies was identified. The clusters that best represent 250 test topics were identified from the hierarchies. The percentage of relevant documents in each cluster was calculated by expanding the hierarchies to a maximum depth of 10 levels. The analysis showed that:

1. Similar number of hierarchy levels were found in AL, CL and WL clusterings ($\overline{HLvl}$ = 6 ± 4.19, 6 ± 3.89 and 6 ± 3.69, respectively),
2. The best clusters appeared on average at Level 2-3 of the three types of clusterings, and
3. Approximately 65% of the relevant documents were found in the best clusters by expanding the hierarchies to a maximum of depth 5 levels.

The results were compared with those generated using poorly-formulated questions. It was found that:

1. The processing of well-formulated questions, which contain higher number of PICO elements, resulted in hierarchies with lower number of clusters and lower number of levels. This also suggests that the fewer the number of PICO elements in a question, the greater the number of hierarchy levels that a user will need to discover (**Table 5-16**).
2. The best clusters were located on average at Level 2-3 and Level 4-5 of the Yule2-WL clusterings generated using well-formulated and poorly-formulated questions, respectively (**Table 5-17**). The results suggest that in response to both types of questions, the best clusters can be identified from the top five levels of Yule2-WL clusterings.
3. The left plot of **Figure 5-9** showed that WL clusterings performed better than AL and CL clusterings in response to poorly-formulated questions, whereas the right plot of the figure showed that the three types of clusterings generated using well-formulated questions performed similarly. About 65% of the relevant documents were found from the best clusters by expanding the Yule2-WL clusterings for both types of questions to a maximum depth of 5 levels.

Table 5-16. Average number of hierarchy levels in Yule2-based clusterings.

Question	Clustering algorithm		
	AL	CL	WL
Poorly-formulated	20 ± 5.89	14 ± 4.10	11 ± 4.15
Well-formulated	6 ± 4.19	6 ± 3.89	6 ± 3.69

Table 5-17. Average location of best clusters in Yule2-based clusterings by hierarchy level.

Question	Clustering algorithm		
	AL	CL	WL
Poorly-formulated	9 ± 4.55	7 ± 3.25	4 ± 2.34
Well-formulated	3 ± 2.30	2 ± 2.55	2 ± 2.20

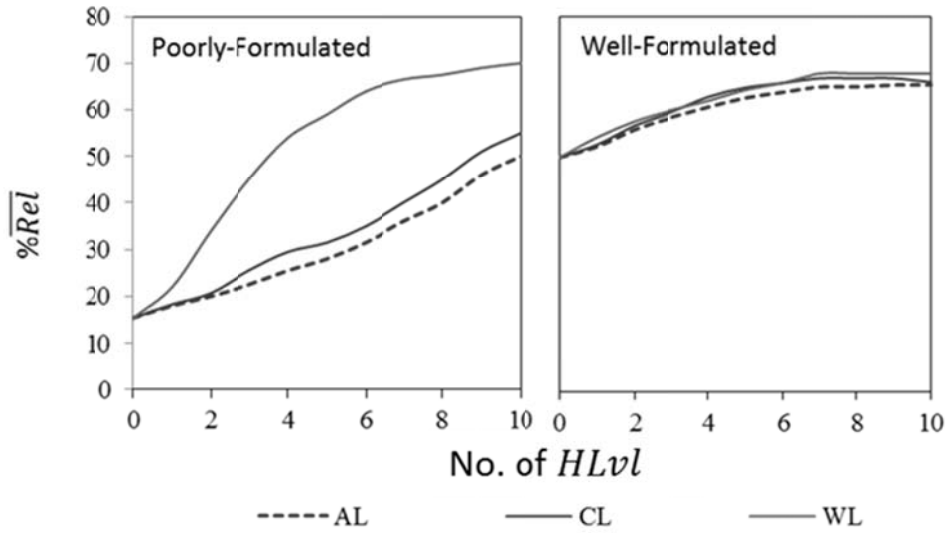


Figure 5-9. Percentage of relevant documents (%$\overline{Rel}$) by hierarchy level ($HLvl$).

The findings indicate that, in response to a question with low or high number of PICO elements, a Yule2-WL clustering that has been expanded to a maximum depth of 5 levels can facilitate the search of clusters with high number of relevant documents (i.e. the best clusters).

5.3. CONCLUSION

In this chapter, documents relevant to a search query were grouped into different clusters using concept similarity agglomerative hierarchical clustering technique. To identify the most appropriate structure for the visualization of a collection of documents, the performance of 12 types of hierarchies were evaluated. The key finding

was that Yule2-WL clusterings performed better than other types of clusterings. A previous study by Aljaber et al. (2010) reported that AL algorithm is more efficient than CL algorithm for clustering scientific documents. An early study by Leuski (2001) showed that both AL and WL algorithms are effective for interactive retrieval of relevant documents. In this chapter, WL algorithm performed better than AL algorithm for interactive search of documents relevant to a given therapy topic. To conclude this chapter, two examples of how a hierarchy of medical interventions was constructed, in response to a poorly-formulated and a well-formulated questions, are shown in **Figures 5-10** and **5-11** respectively.

An analysis of the hierarchies for poorly-formulated questions found that the best clusters were located on average at Level 4-5 and more than 90% of them can be obtained from the top 10 levels of Yule2-WL clusterings. These findings were used to construct a hierarchy with a maximum of 5 levels. As shown in the left panel of **Figure 5-10**, a Yule2-WL clustering was constructed and its height was adjusted so that a distance of 1 unit represents one hierarchy level. The deeper the hierarchy level, the higher the similarity between clusters. The hierarchy was cut at 5 units (or Level 5). The upper part of the hierarchy was removed, whereas the lower part was expanded to a maximum of 5 levels. Each cluster was labelled with the highest frequency terms (i.e. the extracted [I]/[C] elements or therapy topics) in the underlying documents. If any child nodes exist with same name as their parent node, the child nodes will be skipped (the shaded area of the middle figure). The right panel of the figure shows how medical interventions are presented as a hierarchy in the proposed clinical question answering engine.

In response to well-formulated questions, the best clusters were located on average at Level 2-3 of Yule2-WL clusterings. No significant difference was found between Yule2-based AL, CL and WL clusterings, as measured by the percentage of relevant documents in the best clusters ($\%Rel$). By expanding the Yule2-WL clusterings for both well- and poorly-formulated questions to a maximum of 5 levels, similar percentages (about 65%) of relevant documents were identified from the best clusters. Therefore, a hierarchy with equal to or less than 5 levels was cut at the deepest level. An example of how a hierarchy of medical intervention was constructed in response to a well-formulated question is shown in **Figure 5-11**.

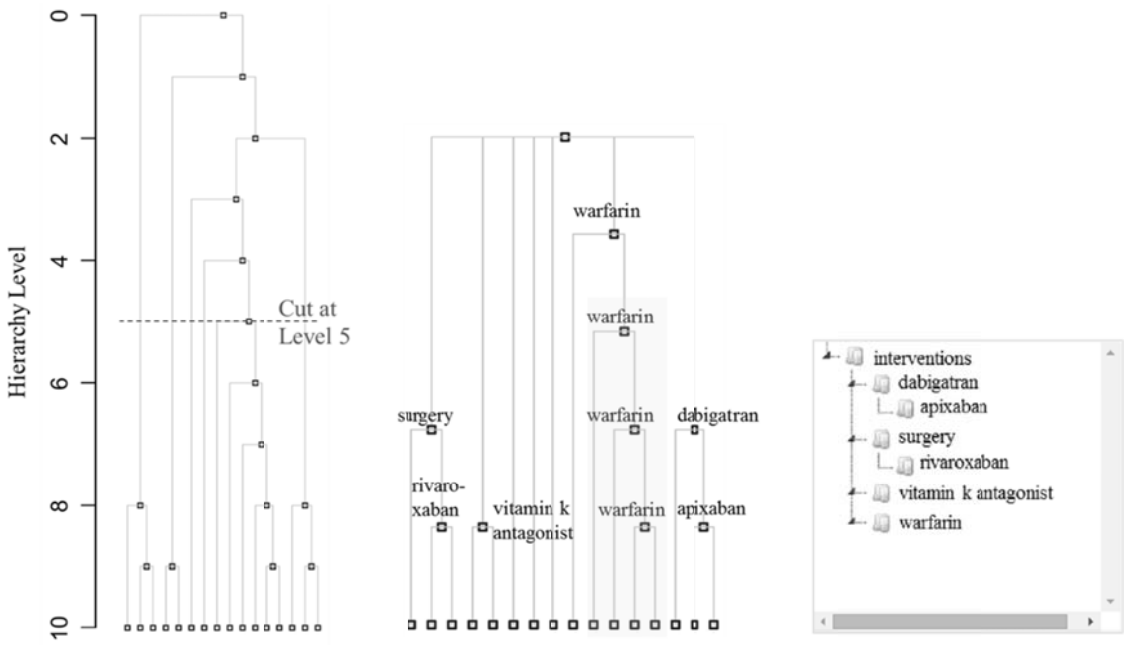


Figure 5-10. A hierarchy of medical interventions displayed by CliniCluster in response to a poorly-formulated question.

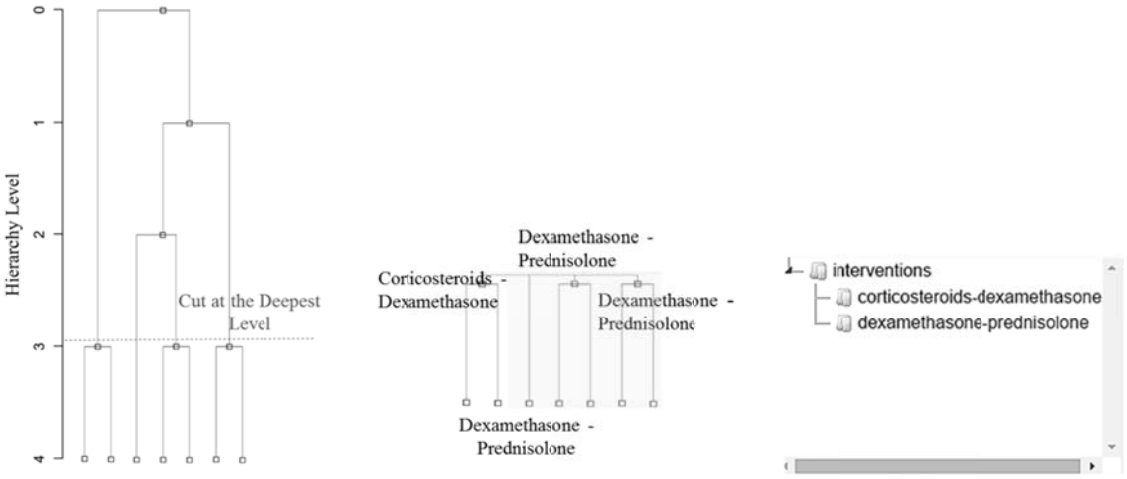


Figure 5-11. A hierarchy of medical interventions displayed by CliniCluster in response to a well-formulated question

The two figures show how a hierarchy of medical interventions was constructed to visualize a collection of documents. The results presented in this chapter were based on the analysis of (600 + 150) = 750 similarity-based clusterings generated using 50 well-formulated and 50 poorly-formulated questions. The purpose is to identify the most appropriate hierarchical structure for browsing, exploring and searching of a collection of documents. In the following chapters, two approaches, known-item searching and a pilot survey, were employed respectively to evaluate the performance of CliniCluster in retrieving highly relevant documents and to validate the usability and user satisfaction with CliniCluster.

6. CHAPTER VI: Known-Item Search

The typical interaction between a user and a question answering system involves the user submitting a query to the system and the system returning a ranked list of relevant documents as answers. The answers are generally sorted by relevance, with the most relevant documents positioned at the top of the ranked list. The user goes through the answers manually and tries to find the information that best corresponds to his or her needs. **Chapters 4** and **5** of this thesis describe respectively how PICO elements are extracted from a collection of documents, and how a hierarchy of medical intervention is constructed. Using the methodologies and findings described in these two chapters, a semi-automated clinical question-answering engine was developed during this thesis, with the aims to support and assist users in meeting their search request and in finding documents that best describe their information needs during the course of information seeking.

In this chapter, the performance of the prototype, called CliniCluster, was evaluated using known-item search method. Known-item search is an information seeking task where users look for a particular document from the result set of a query. The following assumptions were made for the study:

(i) Highly relevant documents are more valuable than marginally relevant documents,
(ii) Highly relevant documents are more likely to be ranked higher in a search result list, and
(iii) A user stops going through a ranked list of document after finding one highly relevant document.

The study was conducted by collecting a set of question-document pairs from a database of filtered, synopsized, evidence-based information for clinical decisions. Each question-document pair contains a clinical question and a document that contains the most valid and relevant clinical information to answer the question. The paired document is defined as the "**known item**". A known-item search task was then performed by submitting a question to a search engine, and the known-item was then identified from the resulting list. Both ill-defined questions and questions of different structural patterns were submitted to CliniCluster and three existing search engines (CQA-1.0, Google and Google Scholar). The performance of the search engines was compared by determining the ranked positions of known-items and the percentage of

known-items identified. Besides, the strength of evidence score was calculated for each of the top-ranked documents to evaluate the effectiveness of the search engines in retrieving evidence-based clinical information.

The remainder of the chapter is organized as follows. **Section 6.1** describes how question-documents pairs were collected and the search engines used for known-item searches. **Section 6.2** gives details of how the known-item search task was carried out. The evaluation metrics used in this chapter are defined in **Section 6.3**. The results are presented and discussed in **Section 6.4**.

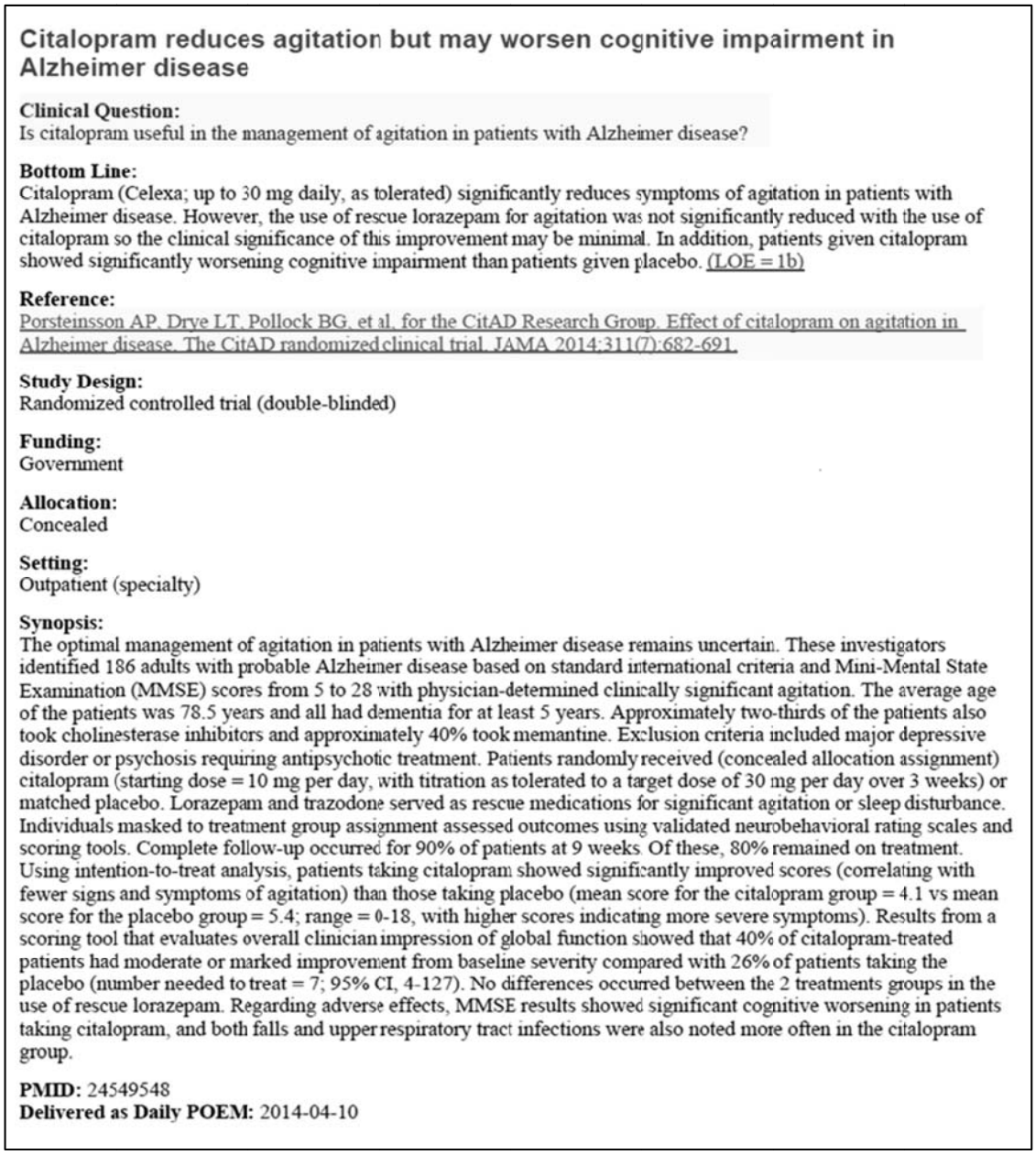
6.1. RESOURCES

6.1.1. Question-Document Pairs

POEMs (Patient-Oriented Evidence that Matters) are articles containing information that have the potential to change clinical practice (**Section 2.1.2**). 70 POEMs concerning the effectiveness of a treatment or preventive measure were collected from the Essential Evidence Plus (EEP) database available at: http://www.essentialevidence plus.com/content/poems. Each POEM, as shown in **Figure 6-1**, contains a clinical question, a bottom-line answer labelled with a level of evidence (LoE) from the Oxford Centre for EBM, a synopsis that indicates the validity and summarizes the most important details of a study, a description of study design and financial support, and the article citation. The article was selected after critically appraising original studies and systematic reviews from more than 100 journals by the Evidence-Based Medicine Working Group. In this chapter, the article was treated as the most valid and relevant study to answer the clinical question posed in the POEM. The clinical question in each POEM was paired with the corresponding article, which is defined as the "known-item". Using the POEM given in **Figure 6-1**, the question-document pair is identified as follows:

Question: "Is citalopram useful in the management of agitation in patients with Alzheimer disease?"

Paired Document: Porsteinsson, A.P., et al. "Effect of citalopram on agitation in Alzheimer disease: the CitAD randomized clinical trial". *JAMA*, Vol. *311*, No. 7, 2014, pp. 682-691.



Citalopram reduces agitation but may worsen cognitive impairment in Alzheimer disease

Clinical Question:
Is citalopram useful in the management of agitation in patients with Alzheimer disease?

Bottom Line:
Citalopram (Celexa; up to 30 mg daily, as tolerated) significantly reduces symptoms of agitation in patients with Alzheimer disease. However, the use of rescue lorazepam for agitation was not significantly reduced with the use of citalopram so the clinical significance of this improvement may be minimal. In addition, patients given citalopram showed significantly worsening cognitive impairment than patients given placebo. (LOE = 1b)

Reference:
Porsteinsson AP, Drye LT, Pollock BG, et al, for the CitAD Research Group. Effect of citalopram on agitation in Alzheimer disease. The CitAD randomized clinical trial. JAMA 2014;311(7):682-691.

Study Design:
Randomized controlled trial (double-blinded)

Funding:
Government

Allocation:
Concealed

Setting:
Outpatient (specialty)

Synopsis:
The optimal management of agitation in patients with Alzheimer disease remains uncertain. These investigators identified 186 adults with probable Alzheimer disease based on standard international criteria and Mini-Mental State Examination (MMSE) scores from 5 to 28 with physician-determined clinically significant agitation. The average age of the patients was 78.5 years and all had dementia for at least 5 years. Approximately two-thirds of the patients also took cholinesterase inhibitors and approximately 40% took memantine. Exclusion criteria included major depressive disorder or psychosis requiring antipsychotic treatment. Patients randomly received (concealed allocation assignment) citalopram (starting dose = 10 mg per day, with titration as tolerated to a target dose of 30 mg per day over 3 weeks) or matched placebo. Lorazepam and trazodone served as rescue medications for significant agitation or sleep disturbance. Individuals masked to treatment group assignment assessed outcomes using validated neurobehavioral rating scales and scoring tools. Complete follow-up occurred for 90% of patients at 9 weeks. Of these, 80% remained on treatment. Using intention-to-treat analysis, patients taking citalopram showed significantly improved scores (correlating with fewer signs and symptoms of agitation) than those taking placebo (mean score for the citalopram group = 4.1 vs mean score for the placebo group = 5.4; range = 0-18, with higher scores indicating more severe symptoms). Results from a scoring tool that evaluates overall clinician impression of global function showed that 40% of citalopram-treated patients had moderate or marked improvement from baseline severity compared with 26% of patients taking the placebo (number needed to treat = 7; 95% CI, 4-127). No differences occurred between the 2 treatments groups in the use of rescue lorazepam. Regarding adverse effects, MMSE results showed significant cognitive worsening in patients taking citalopram, and both falls and upper respiratory tract infections were also noted more often in the citalopram group.

PMID: 24549548
Delivered as Daily POEM: 2014-04-10

Figure 6-1. An example of POEM retrieved from Essential Evidence Plus database.

6.1.2. Search Engines

Four different search engines were used to retrieve relevant documents. CliniCluster and CQA-1.0 are evidence-based clinical tools designed to assist physicians in searching useful information for clinical practice. Google and Google Scholar, on the other hand, are two commonly used search engines by physicians for medical information. The four search engines were evaluated in this chapter for their performance in retrieving and ranking known-items. The following discusses the use of the four search engines for clinical question answering.

CliniCluster.

The current version of CliniCluster is a semi-automated search engine for answering therapy questions. The general architecture of CliniCluster is presented in **Figure 6-2** and is described as follows:

Step 1. A natural language question submitted to the engine is processed to identify medical concepts that represent the four elements of the PICO framework. This is achieved using the MetaMap Transfer (MMTx) program. The program tokenizes an input question into separate phrases and returns relevant UMLS concepts along with their semantic types. Concepts associated with 37 semantic types (**Table 4-4**) are recognized as the PICO elements.

Step 2. The PICO elements are used as the search terms to retrieve relevant documents from the MEDLINE database. The search terms are automatically expanded in PubMed and clinical query filters are applied to improve the search of therapy studies, particularly RCTs, and systematic reviews and meta-analyses. A description of the search strategies is given in **Table 4-1**.

Step 3. It was found in **Chapter 4** that the titles and abstracts of MEDLINE documents provide the most useful medical interventions for the similarity measurement between documents. Therefore, the titles and abstracts of the relevant documents are extracted as the candidate passages. The passages are processed by the MMTx program to identify PICO elements, as described in Step 1.

Step 4. Each relevant document is represented by a bag of medical interventions consists of the [I] and [C] elements. Based on the results from **Chapter 5**, the similarities between documents are calculated using Yule2 similarity metric, and the candidate documents are grouped into a tree of clusters using Ward-link clustering algorithm.

Step 5. A hierarchy of medical interventions with a maximum depth of 5 levels is constructed to represent a collection of documents. Each cluster of the hierarchy contains documents with similar interventions and is labelled with the most frequent therapy topic. The labelling of each cluster is achieved by weighting the [I] and [C] elements that appear among the documents in the same cluster using term occurrence (TO) (**Table 4-6**).

Step 6. By selecting a cluster of interest from the hierarchy, a ranked list of candidate answers is returned to the users along with their associated PICO elements. The candidate answers are extracted from the conclusions of the abstracts and are ranked by strength of evidence score (S_{SOE}) described in **Section 6.3.3**. The purpose is to rank the most recent studies published in 119 core clinical journals and with the highest quality study design on the top position of the result list.

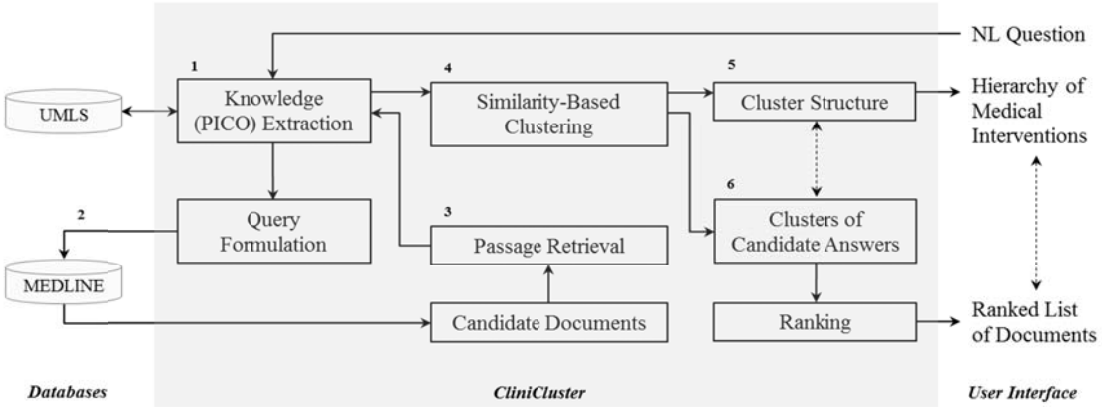


Figure 6-2. Architecture of the proposed CliniCluster engine.

The user interface of CliniCluster is designed to provide an interactive environment to support the search of medical literature for evidence-based clinical practice. As shown in **Figure 6-3**, by posing a natural language question, a hierarchy of medical interventions is displayed at the left side of the interface. It is expected that, by browsing through or exploring the hierarchy, users can gain a better understanding of the medical terminology related to the question posed. A ranked list of answers presented along with the relevant [P-O] and [I/C] elements are shown on the right side of the interface. The elements are extracted from the relevant documents with the intention to support users in searching the documents that best described their information needs. A more detailed description of the features and their usability are included in **Chapter 7**.

Figure 6-3. User Interface of CliniCluster.

CQA-1.0.

CQA-1.0 is a clinical question-answering system developed for physicians practicing EBM. The homepage of CQA-1.0 (**Figure 6-4**) provides an interface that requires users to break down their information needs into four components of the PICO framework. Two search engines, Essie and PubMed are provided by the system. The search results can be limited to human studies, articles with abstracts and those published in English. Besides, a more focused search can be achieved by selecting a specific clinical task (such as treatment, prevention or prognosis), or by retrieving articles from one of the following subsets: core clinical journals, nursing journals, systematic reviews, toxicology and Cochrane reviews. A maximum of 20 top-ranked answers are returned by the system in response to an input query. Each of the answers is supplemented with the relevant PICO elements and the strength of recommendation of A to C. The system is particularly useful for physicians looking for the best available evidence to answer complex clinical questions (Demner-Fushman and Lin, 2007).

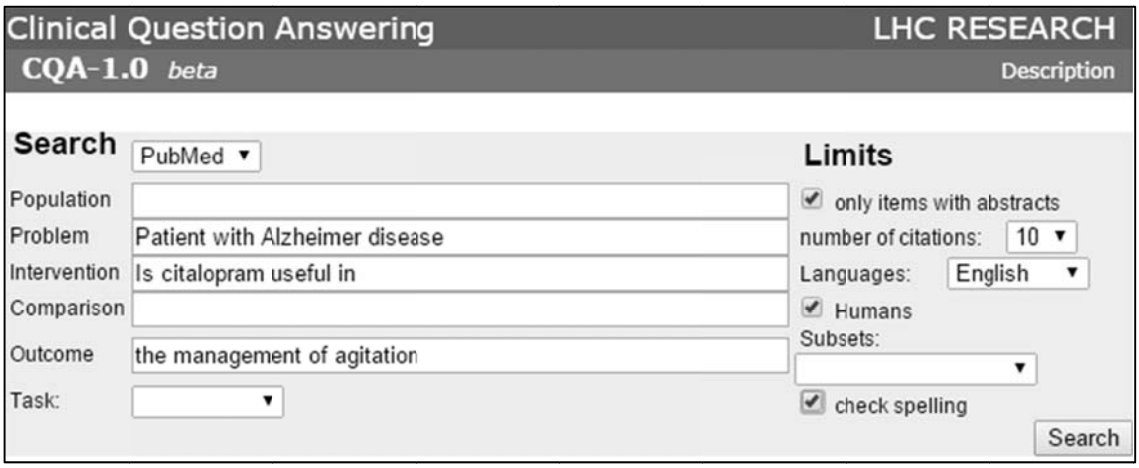


Figure 6-4. An example of broad search using CQA-1.0.

Google and Google Scholar.

Although not specially designed for clinical practice, a study by Hughes (2009) found that 80% of junior physicians used Google for clinical decision making and medical education. A recent study by Duran-Nelson (2013) reported that Google was used by internal medicine residents primarily to locate Web sites and general information about diseases, whereas Google Scholar, was used to locate journal articles and for treatment and management decisions. The advantages of Google include its ease and speed of use, simplicity, and access to images and other knowledge resources such as UpToDate and MD Consult (Giustini and Barsky, 2005; Cook et al., 2013). Google Scholar, as reported by Giustini and Barsky (2005), provides quick and simple browsing, known-item searching, "cited by" feature that links to articles that have cited a given article, and "related articles" feature that presents a list of articles that are closely related to an article selected. However, Google and Google Scholar rank web sites based on keyword relevance and popularity, not on quality for clinical practice and how current are the web pages. Furthermore, Krause et al. (2011) reported that the ability of emergency medicine residents to answer clinical questions correctly using Google was poor, indicating that Google may not be a reliable tool for clinical decision making and medical education. Google Scholar, on the other hand, emphasizes pages that are highly cited, resulting in bias towards older literature. Besides, Google Scholar offers less accurate and less frequently updated medical literature compared to PubMed and does not offer Google's "did you mean" feature to assist with misspellings of search terms (Brunetti and Hermes-DeSantis, 2010; Giustini and Barsky, 2005).

6.2. KNOWN-ITEM SEARCH

The known-item search task involved three key steps: collect test questions, collect relevant documents and search for known-items.

6.2.1. Collect Test Questions

A total of 70 question-document pairs were collected, as described in **Section 6.1.1**. Two sets of therapy questions were generated from the question-document pairs. The first set contains 30 original and 30 ill-defined questions. The ill-defined questions were created by removing one or two of the PICO elements from the original questions, and were matched with the known-items from the original question-document pairs (**Figure 6-5**). This allows a comparison of search results obtained using original

questions to those obtained using ill-defined questions. The second set contains 40 questions of five structural patterns, as described in **Chapter 2 Table 2-1** (20 of Pattern I and 5 of each of Patterns II-V). The purpose is to compare the search results from different search engines using therapy questions formulated with different combinations of PICO elements.

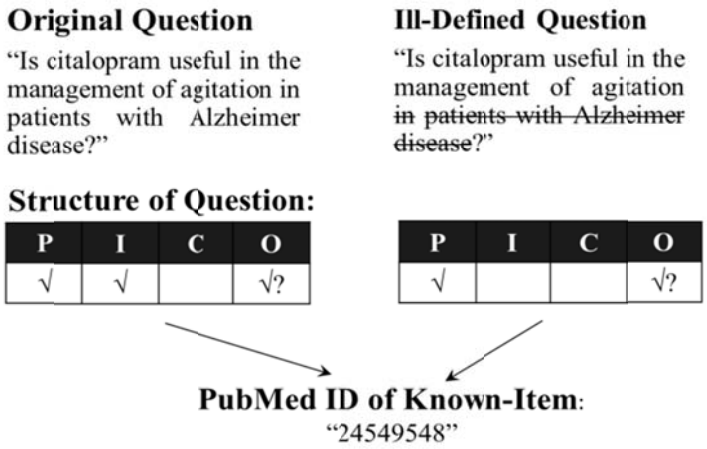


Figure 6-5. An example of how an ill-defined question is created.

6.2.2. Collect Relevant Documents

The two sets of test questions were posed respectively to the four search engines. The test questions were submitted directly to CliniCluster, Google and Google Scholar without applying any of the available search tools. The test questions were broken down into PICO format and entered into CQA-1.0. Two different search strategies were performed in CQA-1.0 to retrieve relevant documents: A narrow search was performed by selecting "treatment" in the "task" option of the system's user interface, whereas a broad search (**Figure 6-4**) was performed without selecting any of the "task" options. Besides, the searches were limited to human studies with abstracts written in English. In response to a question, Google, Google Scholar and CQA-1.0 return respectively a ranked list of relevant documents. CliniCluster returns clusters of documents. Each cluster contains a ranked list of documents. The top-10 and top-20 documents retrieved by each of the search engines were collected.

6.2.3. Search for Known-Items

Non-interactive search. The known-items were identified from the ranked lists of top documents returned by Google, Google Scholar and CQA-1.0. The items were searched without exploiting the hierarchy returned by CliniCluster. In response to a

question, CliniCluster returned automatically a ranked list of relevant documents in the answer field. This was achieved by clicking on the root node called “interventions” in the hierarchy of medical interventions (**Figure 6-3**).

Interactive search. This was performed by expanding the hierarchy returned by CliniCluster to a depth of one level. Two examples were given to describe the approaches to select the child cluster that best answers a question, from which the position of a known-item was identified. As illustrated in **Figure 6-6 (A)**, by clicking the root node (C_0), three child clusters labelled with different therapy topics are displayed. The question “Is citalopram useful in the management of agitation?” contains the [I] element. Therefore, C_{1b} labelled with the most relevant topic “citalopram” is selected. This causes the ranking of known-item (k) to increase from 4 to 1. In case that the most relevant cluster could not be identified by label, or a question does not contain an [I] or [C] element, two assumptions were made to identify the known-items. As shown in **Figure 6-6 (B)**, the question “What is the best treatment for acute otorrhea?” contains only the [P] element,

1. By assuming that the “correct” child cluster is chosen, the ranking of k increases from 6 to 1, and
2. By assuming that the “wrong” child cluster is chosen, a ranking of 0 is given to k.

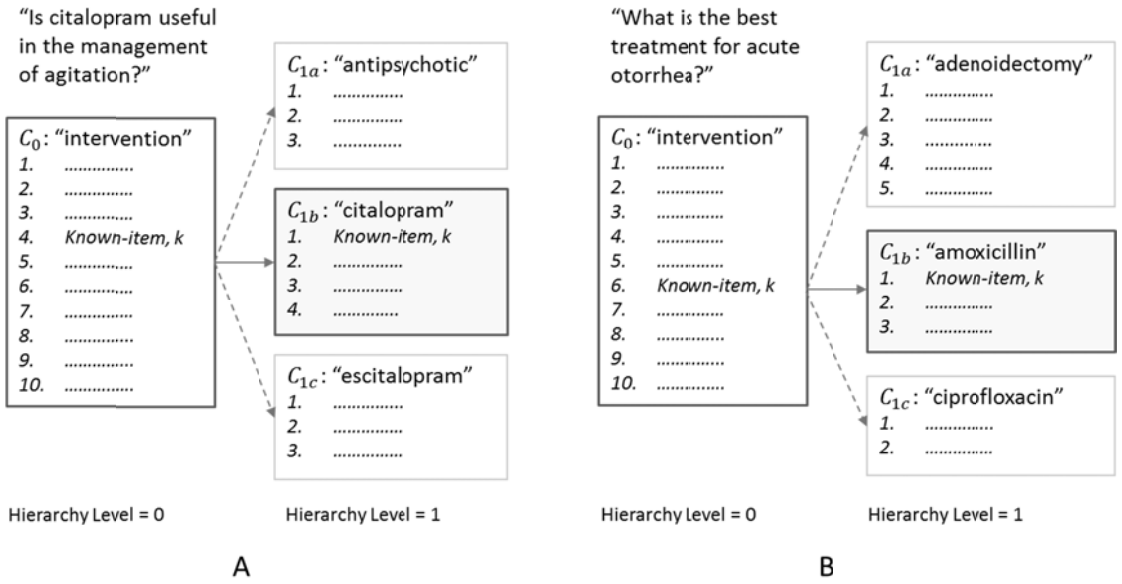


Figure 6-6. Interactive search of a known-item.

6.3. PERFORMANCE MEASURES

The goal of a known-item search is to retrieve a single, specific item. Therefore, evaluation metrics such as precision and recall, that require the search of all the highly relevant documents, were not used to indicate the search performance. Three performance metrics were calculated to compare the performance of the search engines in retrieving known-items: mean reciprocal rank, percentage gain and strength of evidence

6.3.1. Mean Reciprocal Rank

The performance of a search engine over a set of questions was measured using mean reciprocal rank (*MRR*). The measure indicates the average ranking of known items and is defined by the equation below:

$$MRR = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{rank_i}$$

where n is the number of questions and $rank_i$ is the rank of known-item for the i-th question. If a known-item is at rank 1, the reciprocal rank is $1/1 = 1.00$, and if it is at rank 2, the reciprocal rank is $1/2 = 0.50$. If a known-item does not appear in a top-10 result list, the reciprocal rank is 0.00, and if it is at rank 10 of the list, the reciprocal rank is $1/10 = 0.10$. The effectiveness of a search engine increases as the *MRR* approaches 1.00. A system that receives a *MRR* of 0.75 would mean that on average the system finds the known-items between rank 1 and rank 2. A system that obtains a *MRR* of $1/4 = 0.25$ would be finding the known-items on average in position 4 of the result list. *MRR@10* and *MRR@20* indicate that the known-items were searched from the top-10 and top-20 lists, respectively.

6.3.2. Percentage Gain

It was assumed that a question is answered correctly if the known-item appears in the top-10 list, and if it does not, the question is answered incorrectly. The percentage gain (PG) was calculated using:

$$PG = \frac{N_c}{N_t} \times 100\%$$

where N_c is the number of questions correctly answered and N_t is the total number of questions in a test set. *PG@10* tells the percentage of known-items ranked as the top-10 documents.

6.3.3. Strength of Evidence

The strength of evidence (S_{SOE}) score, calculated using the equation below, was first introduced by Demner-Fushman and Lin (2007) to indicate how well a document provides valid and reliable clinical evidence.

$$S_{SOE} = S_{Date} + S_{Study} + S_{Journal}$$

The S_{Date} measures the recency of a document using:

$$S_{Date} = \frac{(Year_{publication} - Year_{current})}{100}$$

The S_{Study} is measured based on the design of a study. **Figure 6-7** presents a hierarchy of evidence for ranking research studies evaluating health care interventions (Evans, 2003). Using the hierarchy of evidence, systematic reviews and meta-analyses receive a score of 0.5; randomized controlled trials (RCTs) 0.4; non-RCTs such as case-control and cohort studies 0.2; and 0 for other non-clinical trials.

The $S_{Journal}$ is determined by the strength of a journal in providing POEs. Documents published in 119 core clinical journals listed in the Abridged Index Medicus, such as *American Family Physician*, *JAMA* and *Lancet*, receive a score of 0.5, and 0 otherwise.

A S_{SOE} score was assigned to each of the top-10 documents. For examples, a double-blind randomized controlled trial ($S_{Study} = 0.4$) published in *N Engl J Med* ($S_{Journal} = 0.5$) on year 2013 ($S_{Date} = -0.02$) obtains a S_{SOE} score of 0.88.

	Effectiveness	Appropriateness	Feasibility
Excellent	• Systematic review • Multi-centre studies	• Systematic review • Multi-centre studies	• Systematic review • Multi-centre studies
Good	• RCT • Observational studies	• RCT • Observational studies • Interpretive studies	• RCT • Observational studies • Interpretive studies
Fair	• Uncontrolled trials with dramatic results • Before and after studies • Non-randomized controlled trials	• Descriptive studies • Focus groups	• Descriptive studies • Action research • Before and after studies • Focus groups
Poor	• Descriptive studies • Case studies • Expert opinion • Studies of poor methodological quality	• Expert opinion • Case studies • Studies of poor methodological quality	• Expert opinion • Case studies • Studies of poor methodological quality

Figure 6-7. Hierarchy of Evidence (adapted from Evans, 2003).

6.4. RESULTS and DISCUSSION

6.4.1. Mean Reciprocal Rank

Original vs. Ill-Defined Questions.

Both the original and ill-defined questions were submitted respectively to each of the search engines. The *MRR@10* and *MRR@20* achieved by each of the search engines are presented in **Table 6-1**. The results showed that:

1. CliniCluster performed remarkably better than other search engines. A *MRR@10* of 0.54 indicates that the know-items could be found on average in rank 2 of the result lists. Besides, the results suggest that CliniCluster is more likely to rank known-items in higher positions, followed by CQA-1.0 (narrow). The performance of Google and Google Scholar was the weakest (*MRR@10* scores < 0.30).
2. There was no or only a slight difference between the *MRR@10* and *MRR@20* scores for each of the search engines. The lowest *MRR@10* score was achieved by Google with a score of 0.20, indicating that the majority of the known-items could be identified from the top-10 lists.
3. By removing one or two of the PICO elements from the original questions, the *MRR@10* score for CliniCluster reduced from 0.54 to 0.45, CQA-1.0 (narrow) from 0.42 to 0.38 and CQA-1.0 (broad) from 0.36 to 0.25. The *MRR@10* scores for original and ill-defined questions submitted to Google and Google Scholar were similar. The results indicate that the known-items were ranked lower when ill-defined questions were submitted to CliniCluster and CQA-1.0.

The same type of analysis was performed using AskHERMES, giving MRR@10 and MRR@20 scores close to 0. Due to the poor results, AskHERMES was not used as one of the benchmarking tools in this chapter.

Table 6-1. *MRR@10* and *MRR@20* for original and ill-defined questions.

Search Engine	Original Questions		Ill-Defined Questions	
	MRR@10	***MRR@20***	***MRR@10***	***MRR@20***
CliniCluster	0.54	0.54	0.45	0.45
CQA-1.0 (narrow)	0.42	0.42	0.38	0.38
CQA-1.0 (broad)	0.36	0.36	0.25	0.26
Google Scholar	0.28	0.28	0.28	0.28
Google	0.20	0.21	0.23	0.23

Five Structural Patterns of Questions

Similar analysis was carried out using five structural patterns of therapy questions. The results were compared using *MRR@10* scores. As shown in **Table 6-2**, CQA-1.0 (narrow) performed the best in retrieving known-items for questions categorized under Patterns I and II, followed by CliniCluster. However, the search performance of CliniCluster for Patterns III to V was significantly better than other search engines, with the known-items ranked on average between positions 1 and 2. By averaging the MRR@10 scores for the five patterns of questions, CliniCluster outperformed other search engines by ranking known-items on average at position 2 (average *MRR@10* = 0.49). Besides, CQA-1.0 (narrow) performed better than CQA-1.0 (broad) (average *MRR@10* = 0.20 and 0.16, respectively). Again, similar results were achieved by Google Scholar and Google (average *MRR@10* = 0.11 and 0.12, respectively).

Table 6-2. *MRR@10* and average rank position for five patterns of therapy questions.

Search Engine	***MRR@10 (Rank Position)***					
	Structural Pattern					***Average***
	I	***II***	***III***	***IV***	***V***	
CliniCluster	0.46 (2~3)	0.15 (6-7)	0.70 (1-2)	0.52 (~2)	0.60 (1-2)	0.49 (~2)
CQA-1.0 (Narrow)	0.48 (~2)	0.35 (2-3)	0.00 (>10)	0.00 (>10)	0.20 (~5)	0.20 (~5)
CQA-1.0 (Broad)	0.33 (~3)	0.17 (~6)	0.02 (>10)	0.07 (>10)	0.20 (~5)	0.16 (~6)
Google Scholar	0.38 (2-3)	0.00 (>10)	0.13 (~8)	0.00 (>10)	0.02 (>10)	0.11 (~9)
Google	0.33 (~3)	0.00 (>10)	0.07 (>10)	0.00 (>10)	0.23 (4~5)	0.12 (8-9)

Examples of the five patterns of questions are given in **Appendix-A**. Using CliniCluster as the search engine, a comparison of the five patterns using *MRR@10* revealed that:

1. Compared to Patterns I and II, Patterns III and IV contain both [I] and [O] elements in the questions. The *MRR@10* scores for Patterns III and IV were higher than those for Patterns I and II. An early study by Bergus et al. (2000) reported that questions formulated with a proposed intervention and a relevant

outcome were unlikely to be unanswered. This is further supported by the present finding that, in response to questions that contain an [I] and an [O] element, the known-items were more likely to be ranked in higher positions.

2. Pattern II contains one, Patterns I and III contain two, and Patterns IV and V contain three PICO elements. Except Pattern II, other patterns yielded a *MRR@10* score close to or greater than 0.50, indicating that the known-items were ranked on average as the top-3 documents. A study by Staunton (2007) reported that a question should include at least three of the four PICO elements in order to be answerable. The results of the present study suggest that at least two of the PICO elements are needed to rank known-items at higher positions in search results.
3. Questions under Patterns I, III and V were posed to return [O?] as the desired answers. The similarities and differences between the three patterns are that:
    a. All patterns contain an [I] element,
    b. Only Pattern III contains an [O] elements,
    c. Only Pattern V contains a [C] element, and
    d. Patterns I and V contain a [P] and an [I] element.

    The results showed that an addition of [C] element to the questions increased the *MRR@10* from 0.46 (Pattern I) to 0.60 (Pattern V). Pattern III yielded the highest *MRR@10*, suggesting that questions that contain both the [I] and [C] elements performed the best in retrieving known-items.
4. Questions under Patterns II and IV were posed to return [I?] as the desired answers. The two patterns differ in that Pattern IV contains an addition [O] element. The *MRR@10* increased from 0.15 for Pattern II to 0.52 for Pattern IV. Once again, the results showed that the presence of [I] and [O] elements in the questions greatly improved the ranking of known-items.

The results presented in this section demonstrate that, in response to different structural patterns of therapy questions, CliniCluster tended to rank known-items in higher positions than other search engines.

Interactive Search of Known-Items

A measure of *MRR@20* was carried out using 5 of each of the five patterns of therapy questions. Each of the questions was submitted to CliniCluster, and the resulting hierarchy was expanded to a depth of one level. The known-item was searched by exploring the root node and the child clusters in the hierarchy. Out of the 25

"correct" child clusters, 20 were identified by matching the [I] and [C] elements in the input questions to those displayed by the hierarchies, and the remaining 5 were assumed to be correctly selected. The average *MRR@20* of the five patterns of questions increased from 0.54 to 0.63, indicating an increase in the ranking of known-items. The deeper the hierarchy level, the higher the similarity of documents in a cluster. This in turn ranks known-items higher in a result list. However, this is true only if the "correct" clusters are selected. By assuming that the "wrong" child clusters were selected (when no [I]/ [C] element that that could be identified from the input questions or when no matching topic that could be identified from the hierarchy), the average *MRR@20* decreased from 0.54 to 0.48. Although a decrease in average *MRR@20* was found, the results indicate that most of the known-items could be identified between rank positions 2 and rank 3. A further analysis revealed that, except for questions categorized under Pattern II, other patterns of questions contain an identified [I] and/or [C] element, which enable the search of "correct" clusters.

A study by Zhu (2008) reported that categorized (or clusters of) results are better than ranked lists of results in information retrieval for very good queries. However, the performance of classification-based system is worse than ranking-based system when human or machine error occurs. The authors introduced a hybrid-based search strategy that a category-based strategy is reverted to a ranked list strategy if the target document is not presented in the first category selected. CliniCluster differs in that, when a question is posed, a ranked list of answers is provided by the root node (i.e. a non-interactive search). The search results can then be narrowed down by selecting the cluster that best described the information need (i.e. an interactive search). It is expected that when a well-structured question is submitted to the engine, a user would not have to perform an interactive search and the most relevant documents can be obtained directly from the ranked list of answers included in the root node. In contrast, when an ill-defined question is submitted, an interactive search can assist them in finding the documents that best match their information needs.

6.4.2. Percentage Gain

Original vs. Ill-Defined Questions.

A question is assumed to be correctly answered if the paired known-item is in a top-10 list. The percentage of questions correctly answered was interpreted using

percentage gain in **Table 6-3**. Up to 90% (27 out of 30) of the original questions were correctly answered by CliniCluster. The percentage gain of CQA-1.0 increased from 53.3% to 60.0% by narrowing down the search to treatment-based studies. Again, ill-defined questions peformed weaker than original questions. Surprisingly, an increase in percentage gain was obtained when ill-defined questions were submitted to Google. The overall results however showed that, using the top-10 lists, CliniCluster is superior to other search engines in answering ill-defined questions.

Table 6-3. Percentage gain (*PG*) for original and ill-defined questions.

Search Engine	***PG@10* (%)**	
	Original Question	***Ill-Defined Question***
CliniCluster	90.0	86.7
CQA-1.0 (narrow)	60.0	50.0
CQA-1.0 (broad)	53.3	53.3
Google Scholar	33.3	26.7
Google	33.3	46.7

Five Structural Patterns of Questions.

Similar to the results obtained using *MRR@10*, CliniCluster performed better than other search engines in answering five structural patterns of questions. As shown in **Table 6-4**, using the top-10 lists retrieved by CliniCluster, more than or up to 80% of questions categorized under Patterns I, III and IV, and up to 60% of questions categorized under Patterns II and V were answered correctly. Using CQA-1.0 as the search engine, a broad search of known-items returned a higher percentage gain than a narrow search (average PG = 36% and 29%, respectively). The lowest percentage gain was achieved by Google (average PG = 19%). Regardless of the pattern of questions, about 75% of the questions were answered correctly by CliniCluster, whereas for other search engines, less than 40% were answered correctly. The results suggest that a higher number of known-items can be identified using the top-10 documents retrieved by CliniCluster, when compared to other search engines.

Table 6-4. Percentage gain (*PG*) for five patterns of therapy questions.

Search Engine	*PG@10* (%)					
	Structural Pattern					Average
	I	*II*	*III*	*IV*	*V*	
CliniCluster	95	60	80	80	60	75
CQA-1.0 (Narrow)	65	60	0	0	20	29
CQA-1.0 (Broad)	60	40	20	40	20	36
Google Scholar	60	0	40	0	20	24
Google	35	0	20	0	40	19

6.4.3. Strength of Evidence

The quality of clinical evidence provided by CliniCluster, Google Scholar and CQA-1.0 (narrow) was evaluated by calculating the S_{SOE} score of each of the top-10 documents. **Table 6-5** shows the percentage of top documents that were published on the past five years ($S_{Date} \geq -0.04$), that were systematic reviews or meta-analyses ($S_{Study} = 0.5$) and that were published in core journals ($S_{Journal} = 0.5$). The table revealed that:

1. CQA-1.0 (narrow) returned a higher percentage of recent publications (from year 2010 to 2014), followed by CliniCluster,
2. More than half of the top-10 documents retrieved by CliniCluster were of the highest quality study design (i.e. systematic reviews or meta-analyses), and
3. Google Scholar outperformed CliniCluster and CQA-1.0 (narrow) with a higher percentage of top documents published in core journals.

Table 6-5. An analysis of top-10 documents using three clinical study quality indicators.

Search Engine	Percentage of Top-10 Documents (%)		
	$S_{Date} \geq -0.04$	$S_{Study} = 0.5$	$S_{Journal} = 0.5$
CliniCluster	77.7	53.1	36.9
CQA-1.0 (narrow)	85.5	3.3	24.1
Google Scholar	17.4	33.0	51.7

A further analysis of the top documents found that a narrow search using CQA-1.0 returned up to 96% of RCTs, whereas 45% of those retrieved by CliniCluster were

RCTs and another 53% were systematic reviews or meta-analyses. The results indicate that CQA-1.0 (narrow) is particularly useful for the search of RCTs. However, an alternative search of review studies can be performed by selecting the "systematic reviews" subset from the user interface of CQA-1.0. An understanding of the search filters provided by CQA-1.0 is needed to conduct a successful search. CliniCluster is different to CQA-1.0 in that a single search returns a ranked list of both review studies and RCTs. Multiple searches are not required to look for the needed information.

It was shown in the previous sections that, using the top-10 lists, CliniCluster returned a greater number of known-items than CQA-1.0 (narrow). CQA-1.0 uses a more complicated algorithm in ranking relevant documents (Demner-Fushman and Lin, 2007). Documents are weighted by matching a question to the candidate documents with the PICO frame, by determining the type of clinical task using MeSH terms, and by discovering the strength of evidence presented by a study. Compared to CQA-1.0, CliniCluster categorizes relevant documents into different clusters using similarity-based clustering method and the documents in each cluster are ranked based on their strength of evidence. A comparison of CliniCluster and CQA-1.0 using the quality indicators described in **Table 6-5** suggests that the ranking and retrieval of known-items rely more heavily on the year of publication of clinical studies. This can be explained by the finding that up to 77% and 85% of top documents retrieved by CliniCluster and CQA-1.0 (narrow), respectively were published on the past 5 years ($S_{Date} \geq -0.04$).

As reported by Beel and Gipp (2009), citation counts is the highest weighted factor in Google Scholar ranking algorithm. The higher the citation count, the more likely that a document is being ranked at the top position in a result list. An analysis of the top documents retrieved by Google Scholar found that about 52% of the documents were published in core journals. The result indicates that the majority of the highly cited documents were published in core journals that are particularly relevant to practicing physicians. On the other hand, as measured using *MRR@10* and *PG@10*, CliniCluster was found to perform much better than Google Scholar in known-item retrieval. An analysis of the top-10 documents revealed that CliniCluster returned a higher number of recent publications and systematic reviews or meta-analyses than Google Scholar. This finding supports the previous study by Guistini (2013) that the use of Google Scholar alone is not enough to search for systematic reviews.

The distributions of S_{SOE} scores of the top-10 documents were presented using histograms. As shown in **Figure 6-8**, the distribution of S_{SOE} scores for CliniCluster skewed to the right (high score region) with an average score of 0.62. For Google Scholar, the histogram was normally distributed whereas for CQA-1.0 (narrow), the scores were distributed mostly between 0.30 and 0.40. Both Google Scholar and CliniCluster obtained an average score close to 0.50. The average S_{SOE} score for each of the search engines indicates that the top-10 documents retrieved by CliniCluster are related to clinical studies with higher quality of evidence, when compared to those retrieved by Google Scholar and CQA-1.0 (narrow).

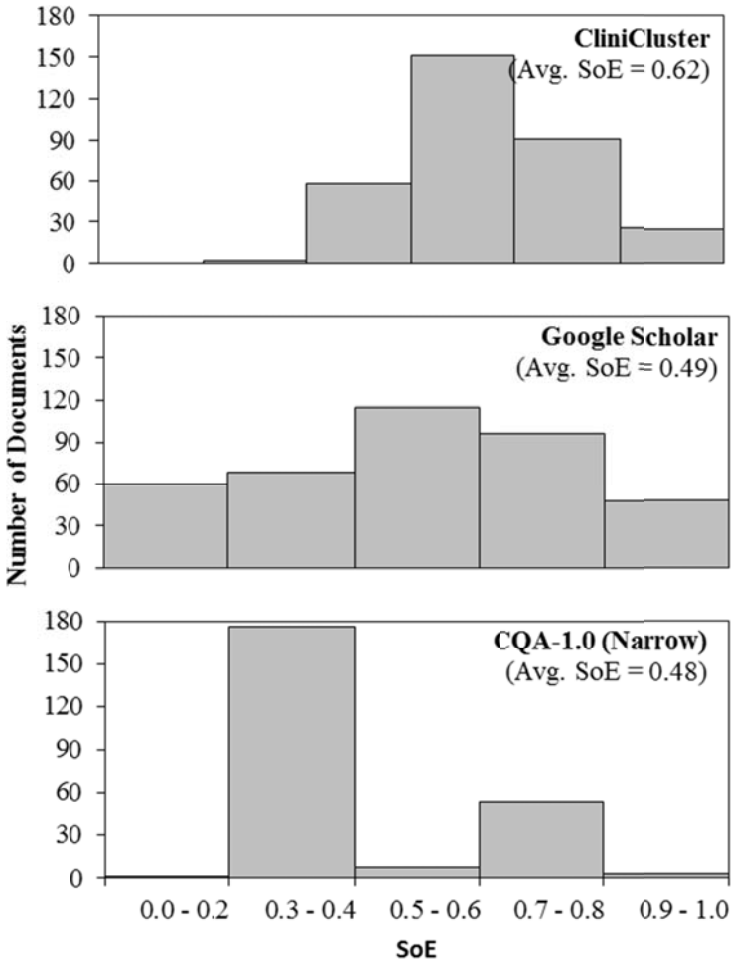


Figure 6-8. Distributions of S_{SOE} scores by histograms.

6.5. CONCLUSION

The study compared the known-item retrieval performance of CliniCluster with three existing search engines. Known-items were identified from the top-ranked documents. The key results are summarized as follows:

1. In terms of *MRR@10* and percentage gain, CliniCluster outperformed other search engines with the known-items ranked higher in the results lists and about 79% of the known-items could be identified from the top-10 lists.
2. In response to therapy questions formulated with different number and combinations of PICO elements, the known-items were ranked on average between position 2 and position 3 in the result lists returned by CliniCluster.
3. An analysis of the strength of evidence provided by the top-10 documents revealed that CliniCluster is superior to other search engines in providing higher number of recent studies of the highest study design.

The overall results concluded that CliniCluster is superior to CQA-1.0, Google and Google Scholar in retrieving and ranking known-items. As described earlier, the known-items were selected critically from a large number of journals and were judged by medical experts to be highly relevant to a therapy question. Although only one item was searched from a result list, the item is highly relevant to a test question and can be identified easily from the top-ranked documents retrieved by CliniCluster.

An ideal QA system is expected to be capable of accepting a variety of natural language question. Compared to CQA-1.0 that require users to transform their information needs into PICO query, CliniCluster is designed to accept both well-defined and ill-defined questions in natural language. Besides, CliniCluster supports and assists users during the information search process by offering a hierarchy of medical interventions and a ranked list of answers presented along with the relevant PICO elements to assist users in finding documents that best match their information needs.

7. CHAPTER VII: A Usability and User Satisfaction Survey

As discussed in the literature review, the majority of MedQA system studies focused on generating a ranked list of documents for a given search query. Less effort has been put on assisting a user in clarifying and meeting his/her information need. Existing MedQA systems assume that users have a clear understanding of their search targets and are aware of their knowledge deficit when formulating a question. There is a lack of interaction between the users and the systems during the information search process. Users may fail to clearly define and express their information needs and have difficulty in formulating a well-focused question. CliniCluster was developed to assist and support users during the information search process. In this chapter, a pilot survey was performed to obtain subjective evidence to support the objective results presented in Chapter 5 and 6. The survey was conducted among 20 health care providers with the purposes to assess:

- The usability of CliniCluster in improving the information search process, and
- The satisfaction of users in completing a search task using CliniCluster.

This chapter starts with a description of the support provided by CliniCluster (**Section 7.1**), followed by the methodology used to conduct and analyze the survey (**Section 7.2**). The results are presented and discussed in **Section 7.3**.

7.1. INFORMATION SEARCH SUPPORT

The current version of CliniCluster is semi-automated and is designed for the interactive search of clinical literature to answer therapy questions. In response to a question in natural language, the support provided by CliniCluster includes:

- A hierarchy of medical interventions is displayed at the left side of the interface, which is constructed based on the similarity of the [I] and [C] elements in a collection of documents (**Figure 7-1**). For instance, the question "*Is citalopram useful in the management of agitation?*" returns a hierarchy with a depth of 3 levels. By clicking on the root node ("interventions"), three child clusters ("antipsychotic", "citalopram" and "escitalopram") are presented to the users at Level 1. These are the therapy topics that appear the most frequent in each of the three clusters of documents. A more narrow search can be performed by selecting the clusters at deeper levels. For example, documents related to both "citalopram"

and “perfenazine” can be found at Level 2. It is expected that, by browsing through or exploring the hierarchy, users can recognize their information needs and gain a better understanding of the medical terminology related to the question posed.

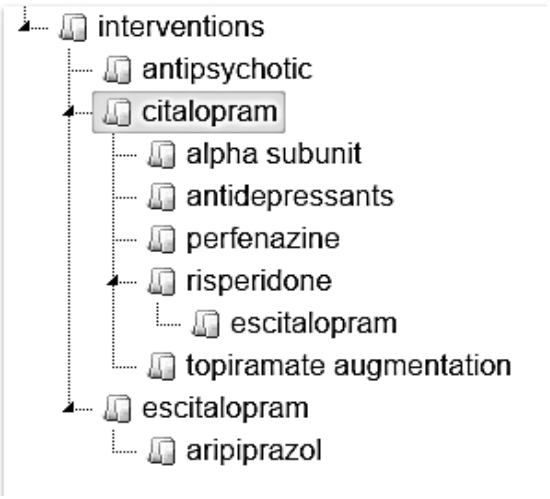


Figure 7-1. Hierarchy of Medical Interventions (Feature 1).

- A ranked list of answers presented along with the relevant [P-O] and [I/C] elements are shown on the right side of the interface. Using the first answer as an example (**Figure 7-2**), [P-O] gives information about the health conditions of a group of patients (“Alzheimer’s disease, Dementia”) and the treatment outcomes of interest (“Aggression, Agitation, Distress”). [I/C] gives information about the treatments or interventions that the study participants received. Other information provided in the result field includes the title, the PMID and the year of publication of an article, and an answer extract from the conclusion in the abstract of an article. Both [P-O] and [I/C] elements are presented with the intention to support users in finding the documents that meet their needs.

TITLE: **Agitation and aggression in people with Alzheimer's disease.**
P-O: **Alzheimer's disease, Dementia - Aggression, Agitation, Distress**
I/C : **Carbamazepine, CITALOPRAM, Memantine, Prazosin**
ANSWER:
Currently, the best approach for managing these symptoms is within a framework of good practice that promotes prevention, monitoring and the use of nonpharmacological alternatives, with judicious short-term use of antipsychotics, when appropriate.
PMID: 23528917
YEAR: 2013

Figure 7-2. PICO elements in the answer field (Feature 2).

7.2. METHODOLOGY

The usability and user satisfaction with CliniCluster were investigated using a questionnaire survey, which is discussed in the following sections.

7.2.1. How was the survey conducted?

A survey was conducted among 20 health care providers in January to April 2015. 10 of the respondents were contacted via email to complete an online questionnaire. Another 10 respondents were contacted by visiting general hospitals (in Kuching, Malaysia) in person. Both groups of respondents were given the same questionnaire (**Appendix-B**). They were first instructed to go through the 25 therapy questions (**Appendix-C**) included in CliniCluster and then complete the questionnaire. The questionnaire consisted of 16 items. The items were designed to collect the information described in **Table 7-1**. The questionnaire was completed by 4 medical specialists, 3 general practitioners, 4 clinical research associates (with medical background), 5 pharmacists and 4 junior doctors (with ≤ 2 years of clinical experience).

Table 7-1. The purpose of the 16-items.

Item	Format	Purpose
1-4	Multiple-Choice	To collect demographic information about the respondents.
5-6	Multiple Choice	To investigate the information seeking behavior of the respondents.
7-8	Five-point Likert scale	To investigate how familiar and how difficult are the 25 therapy topics to the respondents.
9-12	Five-point Likert scale	To evaluate the usability of the hierarchy of medical interventions and the [P-O] and [I/C] elements.
13-16	Five-point Likert scale	To explore the satisfaction of users in completing a search task using CliniCluster.

7.2.2. Statistical Analyses

Two nonparametric tests were used to test the differences in responses to item 7 or item 8 (two ordinal variables that rate the familiarity and the difficulty of the 25 therapy questions respectively) between different demographic groups of respondents. A Mann-Whitney U test was used for comparison of two groups and a Kruskal-Wallis for comparing three or more groups of respondents.

Kendall's tau-b test (for ordinal by ordinal variables) was performed to estimate the correlation between item 7 and item 8 and between item 8 and item 13. Item 13 indicates the respondents' previous knowledge on the topics of the 25 therapy questions.

The five-point responses to items 9-16 were collapsed into two categories: positive ("strongly agree" and "agree") and negative ("neutral", "disagree" and "strongly disagree") responses. Fisher's exact tests were used to compare the two types of responses among two groups of respondents.

All statistical analyses, including descriptive statistics, were performed using SPSS (version 20.0.0, IBM Corporation, New York, USA).

7.3. RESULTS and DISCUSSION

7.3.1. The Respondents

A large number of studies have been conducted to determine the barriers to the uptake of research evidence by clinical decision makers. Most of the studies were conducted on general practitioners and some studies involved medical specialists, surgeons, pharmacists and nurses (Davies, 2011; Wallace et al., 2012; Zwolsman et al., 2012). In the present study, items 1-4 were used to collect demographic information about the respondents. 20 respondents with different medical specialties were included. **Table 7-2** gives a summary of the demographic characteristics.

Table 7-2. Demographic characteristics of respondents.

Characteristic		No. of Respondents (%)	
Gender	Male	10	(50)
	Female	10	(50)
Age	< 30 years old	9	(45)
	≥ 30 years old	11	(55)
Years of Clinical	≤ 5 years	10	(50)
Experience	> 5 years	10	(50)
Medical Specialty	Medical Specialist	4	(20)
	General Practitioner	3	(15)
	Clinical Research Associate	4	(20)
	Junior Doctor	4	(20)
	Pharmacist	5	(25)

7.3.2. Topic Familiarity and Difficulty

The respondents were instructed to complete the search tasks using a list of predefined questions on different therapy topics. Puspitasari and Qu et al. (2015; 2010) showed that users' familiarity with health topics influences their information seeking behaviors. Kim (2006; 2008), on the other hand, reported that pre-task difficulty was related to how much the participants knew about the topics. To identify how familiar and how difficult are the 25 therapy topics to the respondents, they were asked to rate the two items below on a five-point Likert scale.

- **Item 7**: "I was familiar with the topics of the 25 therapy questions."
- **Item 8**: "The topics of the 25 therapy questions were easy for me."

As shown in **Figure 7-3**, only a small number of respondents indicated that they were familiar with the questions (20%) and the questions were easy for them (15%). The two items were rated "medium" by most of the respondents.

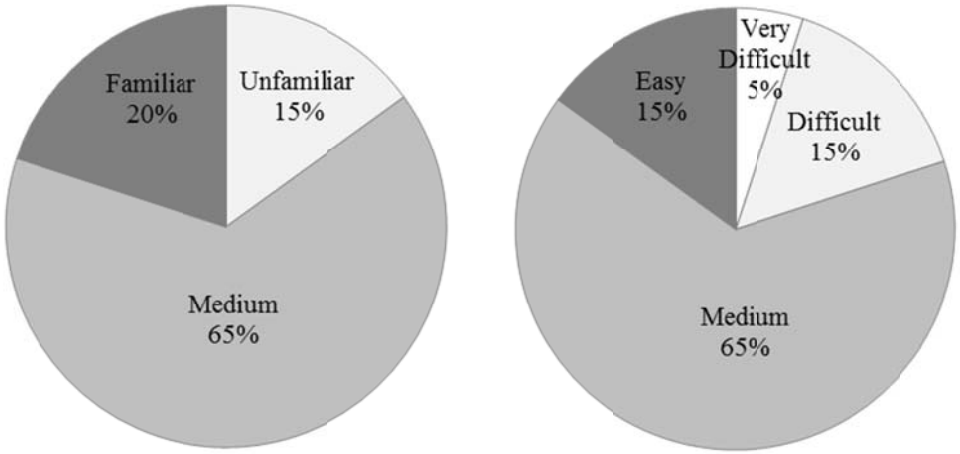


Figure 7-3. Responses to item 7 and item 8.

As the responses were not equally distributed among the respondents, Mann-Whitney U and Kruskal-Wallis H tests were performed with the goal to stratify the respondents into comparable groups. The null hypothesis is: no difference in responses (if $p \geq 0.05$), whereas the alternative hypothesis is: a difference in responses (if $p < 0.05$) to a test item (i.e. item 7 or item 8) between different demographic groups. The results obtained are as follows:

1. A statistical significant difference in responses to item 8 (difficulty) was found between respondents aged < 30 and aged ≥ 30 ($p < 0.05$).
2. There was no difference in responses ($p \geq 0.05$) to item 7 (familiarity) and item 8 (difficulty) when the respondents were grouped by gender, by years of clinical experience and by their medical specialty.
3. As shown in **Table 7-3**, a high mean rank indicates a high familiarity or a high difficulty level. The respondents who aged < 30 found the questions more difficult and more unfamiliar than those aged ≥ 30 years.

A further correlation analysis revealed no statistical significant relationship ($p \geq 0.05$) between item 7 and item 8. The respondents were categorized by age and the responses to items 7 and 8 were illustrated using boxplots in **Figure 7-4**. Based on the

figure, a median of 3 ("median") was achieved by both age groups, as indicated by the thick black lines in the boxplots. Compared to the responses to item 7, item 8 that indicates the difficulty of the therapy questions to the respondents was better represented by the two age groups. This can be explained using the upper and lower whiskers of the boxplots. The responses ranged from "medium" to "very difficult" for those aged < 30 and from "medium" to "easy" for those aged ≥ 30.

Table 7-3. Difference in responses to item 7 and item 8 by two age groups

Age	Mean Rank		No. of Respondents (%)
	Item 7 (familiarity)	*Item 8 (difficulty)*	
< 30 years old	11.78	12.78	9 (45)
≥ 30 years old	8.40	7.50	11 (55)
Significant Testing[t]	$p > 0.05$	$p < 0.01$	

[t] Mann-Whitney U test was used.

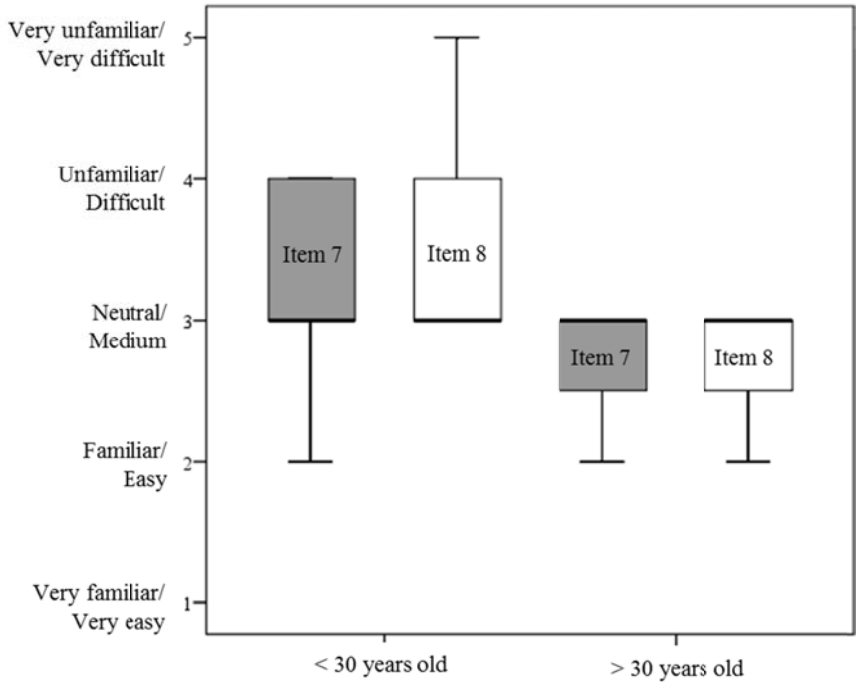


Figure 7-4. Boxplots showing the responses to item 7 and item 8.

A demographic analysis of the age groups, as presented in **Table 7-4**, revealed that 88% of the respondents aged < 30 had ≤ 5 years of clinical experience and more than 70% of them are junior doctors and clinical research associates. In contrast, up to 80% of the respondents aged ≥ 30 had > 5 years of clinical experience and 63% of them

are specialists or general practitioners. This helps explain the findings that those aged < 30 found the questions more difficult and more unfamiliar than those aged ≥ 30 in this study.

Table 7-4. Years of clinical experience and medical specialty of respondents by two age groups

Characteristic		No. of Respondents (%)	
		Aged < 30	*Aged ≥ 30*
Years of Clinical Experience	≤ 5 years	8 (88)	2 (18)
	> 5 years	1 (12)	9 (82)
Medical Specialty	Specialist	-	4 (36)
	General Practitioner	-	3 (27)
	Clinical Research Associate	3 (33)	1 (9)
	Junior Doctor	4 (44)	-
	Pharmacist	2 (22)	3 (27)

7.3.3. Usability of CliniCluster

Previous studies by Kim (2006; 2008) found that the reasons of post-task difficulty focused mainly on the problems encountered during the search process. These are such as searching a specific phrase or terms in a page and assessing a certain site. In the present study, the respondents were not required to report the problems they encountered during the search process. In turn, the four items below were rated by the respondents to measure the usability of CliniCluster:

- **Item 9**: "The hierarchy allowed me to narrow down the search results effectively"
- **Item 10**: "The hierarchy allowed me to explore the relationship between medical interventions in a collection of documents."
- **Item 11**: "I found that it was easy to find relevant documents by checking the [P-O] elements."
- **Item 12**: "I found that it was easy to find relevant documents by checking the [I/C] elements."

As presented in **Figure 7-5**, no difference in responses to items 9 and 10 were observed between the two age groups. 70% of the respondents agreed and 20% were neutral that the hierarchy of medical interventions allowed them to narrow down the

search results effectively (item 9) and to explore the relationship between documents (item 10). This was disagreed by 10% (2 out of 20) of the respondents. On the other hand, the results for items 11 and 12 showed that 80% of respondents aged ≥ 30 agreed that the [P-O] and [I/C] elements allowed them to find relevant documents more easily. Surprisingly, this was agreed by only 60% of those aged < 30, and another 30% and 40% of them rated "neutral" respectively for items 11 and 12. There is no disagreement from both age groups for item 12, suggesting that the [I/C] elements are more useful than the [P-O] elements for the purpose of searching and identifying relevant documents.

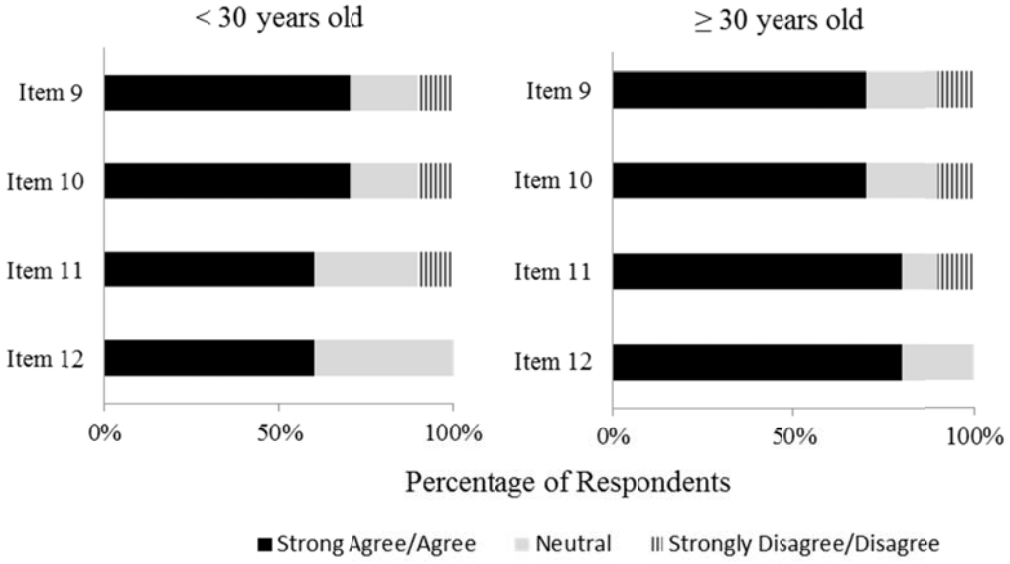


Figure 7-5. Responses to items 9-12.

The median of 4 ("agree") in **Table 7-5** indicates that the majority of the respondents gave a positive response regarding the usability of CliniCluster (items 9-12). The five-point responses (ranged from "strongly agree" to "strongly disagree) were grouped into two categories ("positive" and "negative" responses). Fisher's exact tests were performed to test the null hypothesis: no difference in responses (if $p \geq 0.05$) and the alternative hypothesis: a difference in responses (if $p < 0.05$) to a test item between the two age groups. Although the respondents aged < 30 found the therapy questions more difficult than those aged ≥ 30, there is no significant difference between the two age groups ($p > 0.05$ or $p = 1.00$) in rating the usability of CliniCluster (items 9-12). The overall results support the usability of the hierarchy of interventions and the [P-O] and [I/C] elements in the answer field of the user interface.

Table 7-5. Medians of responses to items 9-12 and significant tests for difference between two age groups

Item	Median of Responses		Significant Testing[t]
	< 30 years old	*≥ 30 years old*	
9	4 ("Agree")	4 ("Agree")	$p > 0.05$
10	4 ("Agree")	4 ("Agree")	$p > 0.05$
11	4 ("Agree")	4 ("Agree")	$p = 1.00$
12	4 ("Agree")	4 ("Agree")	$p = 1.00$

[t] Two-tailed Fisher's exact test was used.

7.3.4. Level of Satisfaction

A few number of studies revealed that, through the search and the critical appraisal of medical literature, physicians gained an improved knowledge and an increased level of confidence in clinical decisions (Scott et al., 2000; Lucas et al., 2004; Straus et al., 2005). Items 13-16 were designed to measure the satisfaction of the respondents to complete a search task using CliniCluster. The items were described as follows:

- **Item 13**: "My previous knowledge on these 25 topics helped me with the search task."
- **Item 14**: "I gained a better understanding of some of the topics during the search task."
- **Item 15**: "I learned new knowledge from some of the topics during the search task."
- **Item 16**: "Overall, I am satisfied with the ease of completing the search task in this scenario."

The result of Kendall's tau-b test revealed that there is no relationship between the respondents' previous knowledge (item 13) and the difficulty of the therapy topics rated by them (item 8) ($p \geq 0.05$). Besides, **Figure 7-6** shows that 60% of the respondents agreed, 35% were neutral and 5% disagreed that their previous knowledge helped them in performing the search task (item 13). As the item was rated "strongly agree" or "disagree" by only 10% (2 out of 20) of the respondents, the respondents were grouped broadly into "high knowledge" and "moderate knowledge" groups. The purpose is to

identify whether the respondents' previous knowledge on the 25 topics affect the satisfaction gained by them during the search tasks.

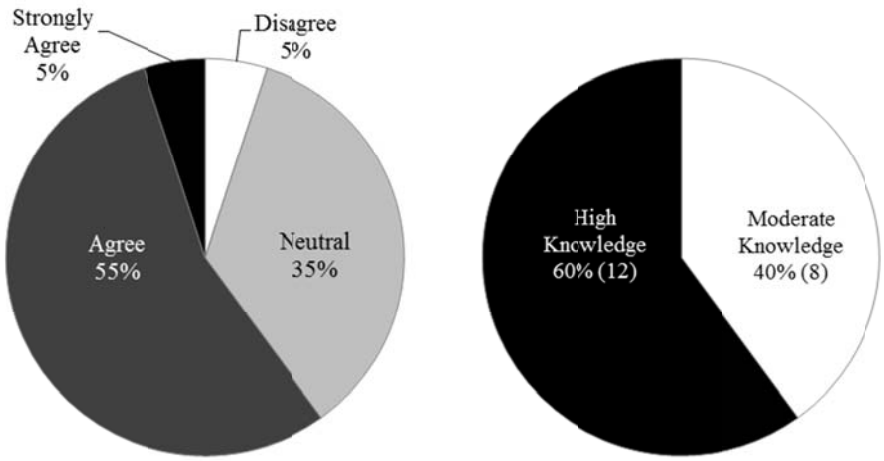


Figure 7-6. Responses to item 13.

As illustrated in **Figure 7-7**, a mark difference in responses was observed between the two knowledge groups. 62.5% of the respondents from the moderate knowledge group agreed, 25% were neutral and 12.5% disagreed that they gained a better understanding of some of the topics during the search task (item 14). However, this was 100% agreed by the 12 respondents from high knowledge group. Besides, 75% of the respondents from the moderate group were neutral, and only 25% of them agreed that they gained new knowledge from some of the topics (item 15). Again, this was strongly agreed by the high knowledge group (100% for item 15).

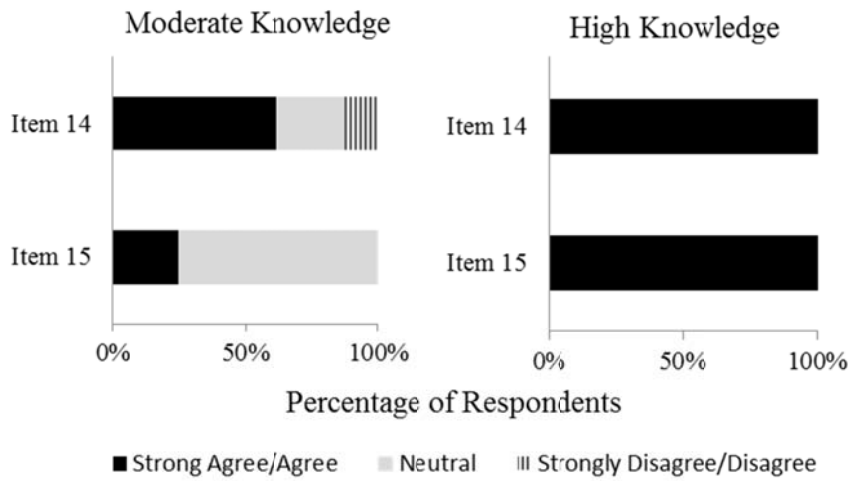


Figure 7-7. Responses to items 14-15.

Similar to the previous section, the five-point responses were collapsed into positive and negative responses and Fisher's exact tests were performed. The null

hypothesis is: there is no difference in responses (if $p \geq 0.05$), whereas the alternative hypothesis is: there is a difference in responses (if $p < 0.05$) to the test item between the moderate and high knowledge groups. It can be seen from the medians and p-values listed in **Table 7-6** that there is no significant difference in responses to item 14 between the two knowledge groups (median = "agree", $p = 0.005$). Both groups agreed that they gained a better understanding of some of the topics during the search tasks. On the other hand, a significant difference in responses was found between the two knowledge groups for item 15 ($p = 0.001$), with a median of 3 ("neutral") for the moderate knowledge group and a median of 4 ("agree") for the high knowledge group. The results suggest that, compared to the moderate knowledge group, respondents with high previous knowledge obtained a higher satisfaction during the search tasks. They gained new knowledge and better understandings of the topics provided. Overall, 60% (12 out of 20) of the respondents involved in this survey agreed that they were satisfied with the ease of completing the search tasks using the prototype system (item 16), whereas the other 40% (8 out of 20) of the respondents gave a neutral response.

Table 7-6. Medians of responses to items 14-15 and significant tests for difference between two knowledge groups

Item	Median of Responses		Significant Testing[t]
	Moderate Knowledge (n = 8)	*High Knowledge (n = 12)*	
14	4 ("Agree")	4 ("Agree")	$p = 0.05$
15	3 ("Neutral")	4 ("Agree")	$p = 0.001$

[t] Two-tailed Fisher's exact test was used.

7.3.5. Information Seeking Behavior

Items 5 and 6 were designed to explore how often the respondents search the Internet and how they search for health-related information. **Table 7-4** shows that the majority of respondents aged < 30 are junior doctors and clinical research associates with ≤ 5 years of clinical experience, whereas the majority of those aged ≥ 30 are specialists and general practitioners with > 5 years of clinical experience. As shown in **Figure 7-8**, respondents aged ≥ 30 search the Internet more often (5-7 times a week) than those aged < 30 (2-4 times a week) for health-related information. This also indicates that respondents with greater clinical experience search more frequently for health-related information from the Internet than those with less clinical experience.

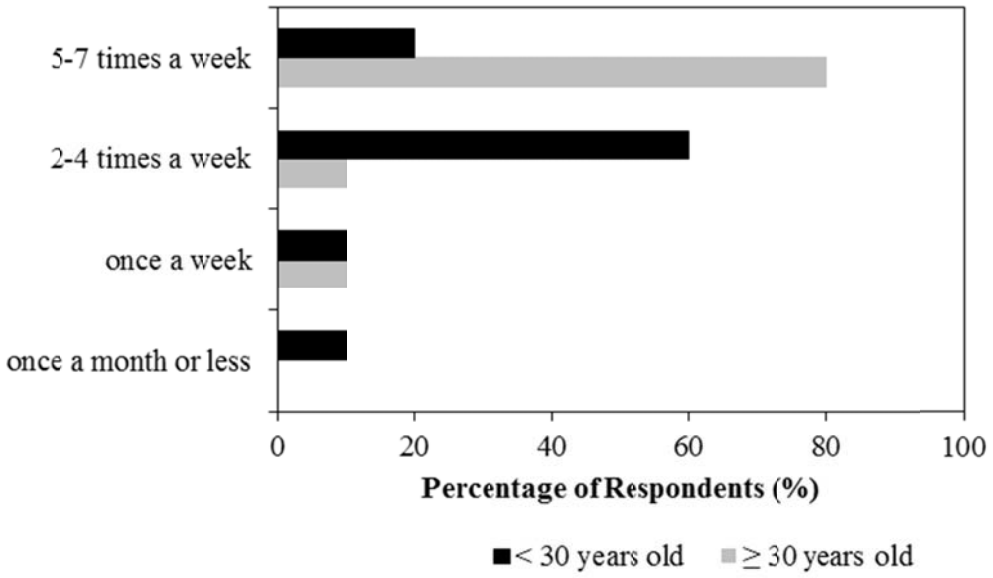


Figure 7-8. Responses to item 5.

A few number of studies reported that the most commonly used electronic resources by health care providers is MEDLINE (Schilling et al., 2005; Cullen et al., 2011; Davies, 2011). A recent review by Kosteniuk et al. (2013) reported that family physicians (FPs) used different information sources for different purposes. The most popular source used by FPs for clinical decision purposes is medical textbooks and for the update of general medical knowledge is medical journal. In the present study, the respondents were allowed to select up to three of the resources that they use for searching health-related information (as listed in **Figure 7-9**). The most frequently used resource by both groups is Web-based search engines (such as Google), followed by textbooks or colleagues and corpus-based search engines (such as MEDLINE database though PubMed). The least frequently used resource by both groups is MedQA systems (such as the AskHERMES). Those aged < 30 assess web-based search engines and consult textbooks or colleagues the most for health information. The most frequently used resources by respondents aged ≥ 30 are web-based and corpus-based search engines. Only half of the 20 respondents use evidence-based medicine databases (such as the Cochrane Library). The reasons that limit the use of the Cochrane Library have been reported previously by De Vito et al. and Wallace et al. (2009; 2012). Only 4 out of the 20 respondents (2 specialist and 2 junior doctors) use MedQA systems for health-related information searching. Although not widely used by health care providers, the studies by Athenikos et al. and Bauer et al. (2010; 2012) demonstrated that MedQA

systems are improving and are close to becoming valuable tools for the search of quick and reliable clinical evidence.

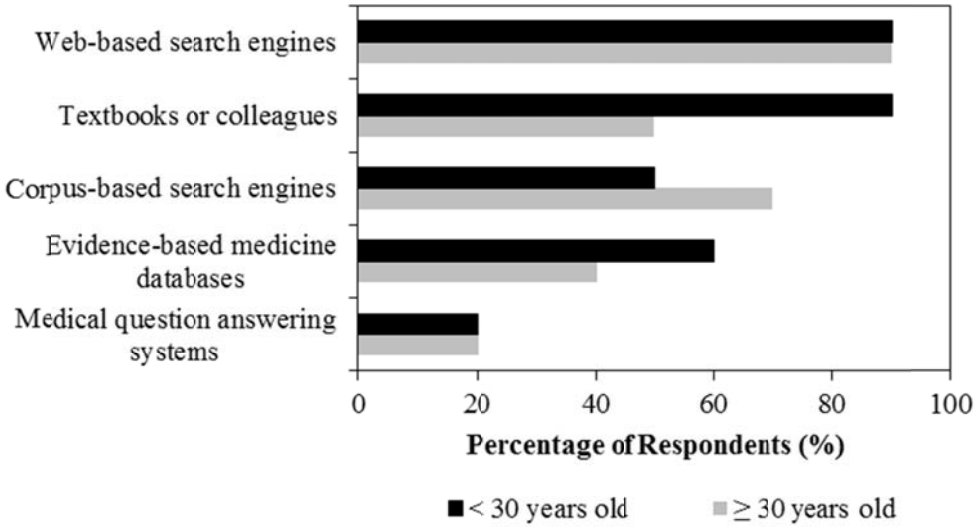


Figure 7-9. Responses to item 6.

7.4. CONCLUSION

This chapter reports the results of a pilot survey conducted among 20 health care providers with different medical backgrounds. They were asked to rate the usability of support provided by CliniCluster and the satisfaction that they experienced from completing a search task using CliniCluster. Although only a few of them had experience in using MedQA systems, the majority of them agreed that CliniCluster assists them in narrowing down the search results and in identifying relevant documents. Besides, most of them gained new knowledge and better understanding of the therapy topics included in the prototype engine. The findings of this chapter are limited by low number of respondents and the use of predefined questions as the information needs of respondents. The usability of CliniCluster can be better evaluated if the respondents can pose a question that they are familiar with. However, the overall results of this chapter support the use of CliniCluster for clinical question answering.

8. CHAPTER VIII: Thesis Conclusion

The purpose of this final chapter is to summarize the contents of the previous chapters and to draw a conclusion about this thesis. The chapter also highlights the limitations of the thesis and provides recommendations for future research.

8.1. SUMMARY and CONTRIBUTIONS

As discussed in **Chapter 1**, physicians are encouraged to find evidence-based clinical evidence to answer clinical questions that arise in daily practice. However, physicians are often time-constrained and have difficulties in searching and evaluating literature for the best available clinical evidence. Therefore, MedQA systems are developed for physicians to find direct and precise answers to patient-care questions.

Chapter 2 provides the strategies to find high quality clinical evidence. The key strategies include:

- Converting an information need into a searchable and answerable question using the PICO question framework,
- Focusing on studies that address questions and outcomes that matter to patients such as quality of life and mortality, and
- Concentrating on studies that deliver high quality research evidence such as randomized controlled trials and systematic reviews.

Besides, a literature review of MedQA systems is presented in **Chapter 2**. Most of the current MedQA systems assume that users have a good idea of what they are searching for and have the ability to convert their information needs into searchable queries. The problems that users may experience when using the current MedQA systems include:

- Inability to formulate well-focused questions due to a lack of terminology or a lack of knowledge of a new or specialized domain.
- Inability to break down their information needs to fit the question framework used by a system,
- Inability to use specific search syntax such as Boolean operators, and
- A misplaced expectation that a system is aware of their information needs.

There is a lack of studies that focus on assisting users in clarifying and refining their information needs by promoting the interaction between the users and the systems.

CliniCluster is a prototype clinical question answering engine consists of two stages: the exploratory stage and the concept stage. As proposed in **Chapter 3**, by submitting a question to CliniCluster, users are allowed to capture, explore and narrow down their information needs by interacting with a hierarchy of medical interventions in the exploratory stage. The concept stage on the other hand aims to assist users in quickly locating their information needs. This is achieved by selecting a cluster of interest from the hierarchy and a ranked list of documents is presented to the users along with their associated PICO elements.

Chapter 4 contributes by investigating the most appropriate field of MEDLINE documents for the extraction of PICO elements, and by identifying the most optimal similarity/distance metrics for the measurement of concept-based similarity between documents. **Chapter 5** contributes by identifying the most effective agglomerative hierarchical clustering algorithm to organize documents into meaningful clusters, and by determining the most appropriate hierarchical structure for the visualization of therapy topics related to a given search request. The results from these two chapters demonstrate that:

- The titles and abstracts of MEDLINE documents provide the most useful PICO elements for the similarity measurement between documents. The extraction of PICO elements is important for the subsequent similarity-based clustering of documents and the visualization of key medical concepts as a feature to support the information seeking process.
- Yule2-WL clusterings provide the most appropriate hierarchical structure for the clustering of relevant documents. Most of the clusters with high recall and high precision can be obtained from Level 5 to Level 10 of Yule2-WL clusterings. This suggests that a hierarchy should be cut at Level 5, and be further expanded to a maximum depth of 5 levels.
- A poorly-formulated question is more likely to result in a hierarchy with a higher number of levels than a well-formulated question. Therefore, if a hierarchy has equal to or less than 5 levels, the clusters appear at the deepest level are selected, whereas those appear between Level 5 and Level 10 are selected if a hierarchy has greater than 5 levels.

Using the findings in **Chapter 4** and **Chapter 5**, CliniCluster was developed with the capability to support the information search process by offering an expandable

hierarchical structure of medical interventions, and a ranked list of documents supplemented with the [P-O] and [I/C] elements.

Two approaches were adopted to evaluate the performance of CliniCluster and are presented respectively in the next two chapters. In **Chapter 6**, the performance of CliniCluster and three existing search engines (CQA-1.0. Google and Google Scholar) was compared by known-item searching. The known-items are a collection of evidence-based documents that have been critically appraised by medical experts to be highly relevant to a set of test questions. It was found that:

- In response to well- and poorly-formulated questions, CliniCluster is more likely to rank known-items higher than other search engines and most of them can be obtained using the top-10 documents.
- Similarly, in response to therapy questions of five structural patterns, known items are ranked higher by CliniCluster, followed by CQA-1.0, and then Google and Google Scholar.
- In terms of the strength of evidence provided by the top-ranked documents, CliniCluster is superior to other search engines in providing higher number of recently published systematic reviews and meta-analyses.

Chapter 7 reported a pilot survey conducted among a group of 20 health care providers. The respondents were asked to rate the usability and their satisfaction with the support provided by CliniCluster. The survey results indicate that:

- The majority of the respondents agreed that the hierarchy of medical interventions and the [P-O] and [I/C] elements in the answer field assisted them in narrowing down the search results and in identifying relevant documents.
- The majority of the respondents agreed that they gained new knowledge and better understanding of the searched topics. Besides, they were satisfied with the ease of completing the search tasks using CliniCluster.

The results presented in **Chapter 6** indicate that CliniCluster is effective in retrieving highly relevant and evidence-based documents for both well- and poorly-formulated questions and questions of different structural patterns. The pilot study described in **Chapter 7** further supports the usability of CliniCluster for searching and recognizing relevant documents.

To conclude, the results obtained using the concept-similarity clustering approach support the hypothesis that a hierarchical structure of medical interventions can assist user in narrowing down and better understanding their search intent. Besides, the visualization of PICO elements can facilitate the search of relevant documents and therefore, improve the information retrieval performance of users.

8.2. LIMITATIONS and FUTURE DIRECTIONS

There are a number of limitations of this thesis that must be discussed to increase the effectiveness of CliniCluster in retrieving high-quality clinical evidence, and to improve the applicability of CliniCluster in daily clinical practice. The general architecture of a QA system consists of a question processing phase, a document retrieval phase and an answer extraction phases.

Question Processing Phase. The current version of CliniCluster was designed to be capable of processing and answering therapy questions in natural language. Therapy questions are the most commonly asked questions at the point of care. However, in order to be suitable for a wide range of applications in future, the system needs to be further developed to process other types of clinical questions such as diagnosis, prognosis and epidemiology questions.

Document Processing Phase. Documents relevant to the test questions were retrieved only from the MEDLINE database. The reason is that the MEDLINE database is the most widely used electronic resource by physicians for systematic reviews and primary studies. The usability of the system can be further improved by allowing users to search multiple databases such as the Cochrane Library, CINAHL and EMBASE at once, or to select the database that they prefer.

Answer Processing Phase. The current version of CliniCluster has limitation in its ability to indicate whether the multiple answers displayed to users agree with each other on a particular query. Although not the focus of the current research, an important direction for future research is to develop a novel search support feature that can point out the similarities and differences of findings from multiple studies. Besides, the pilot survey showed that the majority of health care providers agreed that the features in CliniCluster support the information search process. The study however was limited by the small number of participants.

In summary, in order to be adopted in daily practice, CliniCluster needs to be further studied and optimized to process other types of clinical question and to generate highly informative answers that can be utilized quickly without further effort.

APPENDIX

A. Five Structural Patterns of Therapy Questions

The table below gives two examples for each of the five structural patterns of therapy questions. The examples illustrate how the PICO elements were identified from the questions

Pattern		Examples
I.	[P][I][O?]	• Is enoxaparin [I] useful for moderate renal impairment [P]? • Does niacin plus laropiprant [I] useful for patients with vascular disease [P]?
II.	[P][I?]	• What is the best treatment for acute otorrhea [P]? • What is the best way to treat menorrhagia [P]?
III.	[I][O?]	• Is zanamivir [I] effective in relieving flu symptoms [O]? • Is gabapentin [I] useful in decreasing cough [O]?
IV.	[P][I?][O]	• Is duloxetine [I] effective in reducing pain [O] from chemotherapy-induced peripheral neuropathy in adult cancer survivors [P]? • Are epidural corticosteroid injections [I] effective in decreasing pain and improving function [O] in patients with sciatica [P]?
V.	[P][I][C][O?]	• What is the comparative effectiveness of ondansetron [I] and metoclopramide [C] for treatment of hyperemesis gravidarum [P]? • Is aspirin [I] as effective as dalteparin [C] for extended venous thromboembolism prophylaxis in patients who have undergone total hip arthroplasty [P]?

B. 16-Item Questionnaire

1. What is your gender?
   - □ Male
   - □ Female
2. What is your age?
   - □ 20-29
   - □ 30-39
   - □ 40-49
   - □ 50 and over
3. What is your medical specialty?
   - □ Medical specialist
   - □ General Practitioner
   - □ Clinical Research Associate
   - □ Pharmacist
   - □ Others: _____
4. How many years of clinical practical experience do you have?
   - □ Less than 2 years
   - □ 2-3 years
   - □ 4-5 years
   - □ Over 5 years
5. How often do you search the Internet for health-related information?
   - □ once a week
   - □ 2-4 times a week
   - □ 5-7 times a week
   - □ once a month or less
6. How do you search for health-related information?

   (You may select more than one option)
   - □ Evidence-based medicine databases (e.g. Cochrane Library)
   - □ Medical question answering systems (e.g. AskHermes system)
   - □ Corpus-based search engines (e.g. PubMed)
   - □ Web-based search engines (e.g. Google)
   - □ Textbooks or Colleagues

7. The topics of the 25 therapy questions were easy for me.

- □ Very Easy
- □ Easy
- □ Medium
- □ Difficult
- □ Very Difficult

8. I was familiar with the topics of the 25 therapy questions.

- □ Very Unfamiliar
- □ Unfamiliar
- □ Medium
- □ Familiar
- □ Very Familiar

9. The hierarchy of medical interventions allowed me to narrow down the search results effectively.

- □ Strongly Disagree
- □ Disagree
- □ Neutral
- □ Agree
- □ Strongly Agree

10. The hierarchy of medical interventions allowed me to explore the relationship between medical interventions in a collection of documents.

- □ Strongly Disagree
- □ Disagree
- □ Neutral
- □ Agree
- □ Strongly Agree

11. I found that it was easy to find relevant documents by checking the P-O elements in the answer field.

- □ Strongly Disagree
- □ Disagree
- □ Neutral
- □ Agree
- □ Strongly Agree

12. I found that it was easy to find relevant documents by checking the I/C elements in the answer field.

- Strongly Disagree
- Disagree
- Neutral
- Agree
- Strongly Agree

13. My previous knowledge on these 25 topics helped me with the search task.

- Strongly Disagree
- Disagree
- Neutral
- Agree
- Strongly Agree

14. I gained a better understanding of some of the topics during the search task.

- Strongly Disagree
- Disagree
- Neutral
- Agree
- Strongly Agree

15. I learned new knowledge from some of the topics during the search task.

- Strongly Disagree
- Disagree
- Neutral
- Agree
- Strongly Agree

16. Overall, I am satisfied with the ease of completing the search task in this scenario.

- Strongly Disagree
- Disagree
- Neutral
- Agree
- Strongly Agree

C. 25 Predefined Therapy Questions

1. Does stopping antibiotic treatment after cholecystectomy for mild to moderate acute calculous cholecystitis affect outcomes?
2. Do epidural glucocorticoid injections improve the symptoms of spinal stenosis?
3. In patients with obstructive sleep apnoea, is continuous positive airway pressure or nocturnal oxygen therapy better for reducing blood pressure than usual care alone?
4. Does Ramipril improve symptoms and quality of life in patients with intermittent claudication?
5. Does metformin affect cardiovascular events in patients with type 2 diabetes?
6. Is nortriptyline effective in the treatment of adults with idiopathic gastroparesis?
7. Is the measles-mumps-rubella booster vaccine safe and effective for children with juvenile idiopathic arthritis?
8. Is dabigatran a safe and effective anticoagulant for patients with mechanical heart valves?
9. Is cognitive behavioral therapy combined with amitriptyline superior to amitriptyline alone for the treatment of chronic migraine in children and adolescents?
10. In adults with nontraumatic supraspinatus tears, is physical therapy alone as effective as physical therapy plus surgery after 1 year?
11. Is citalopram useful in the management of agitation in patients with Alzheimer disease?
12. In children presenting to the emergency department with mouth ulcers, does lidocaine treatment improve fluid intake?
13. Does supplemental vitamin D increase bone mineral density?
14. Is zanamivir effective in relieving symptoms in patients with confirmed or suspected influenza?
15. Is gabapentin effective in treating patients with refractory chronic cough?
16. How should anaemia and iron deficiency be treated in adults with heart disease?
17. Does magnesium supplementation reduce the symptoms of nocturnal leg cramps?
18. Is acupuncture effective in relieving pain in patients with chronic low-back pain?
19. Are epidural corticosteroid injections effective in decreasing pain and improving function in patients with sciatica?
20. Is duloxetine effective in reducing pain from chemotherapy-induced peripheral neuropathy in adult cancer survivors?
21. Is an angiotensin-converting enzyme inhibitor plus an angiotensin-receptor blocker (ARB) better than an ARB alone for patients with type 2 diabetes mellitus and impaired renal function?
22. Is aspirin as effective as dalteparin for extended venous thromboembolism prophylaxis in patients who have undergone total hip arthroplasty?
23. Is high-dose oseltamivir more effective than the standard dose in patients admitted to the hospital with confirmed severe influenza?
24. What is the comparative effectiveness of ondansetron and metoclopramide for treatment of hyperemesis gravidarum?
25. In children with acute asthma exacerbations, is oral or injected dexamethasone as effective as predisone or prednisolone?

D. List of Publications

Chapter 2: Vong, W. T. & Then, P. H. H. (2014). Visualization of PICO Elements for Information Needs Clarification and Query Refinement. In *Advances in Knowledge Discovery and Data Mining* (pp. 360-372). Springer International Publishing.

Chapter 4: Vong, W. T. & Then, P. H. H. (2016). A Comparison of Similarity Metrics for Hierarchical Clustering of Medical Interventions. *JP Journal of Biostatistics,* 13(1), 1-27.

Chapter 5: Vong, W. T. & Then, P. H. H. (2016). An Evaluation of Hierarchical Methods for Visualization of Medical Interventions. *JP Journal of Biostatistics,* 13(1), 29-63.

Chapter 6: Vong, W. T. & Then, P. H. H. (2015). Known-Item Retrieval Performance of a PICO-based Medical Question Answering System. *Asia Pacific Journal of Information System*, 25(4), 686-711.

Chapter 7: Vong, W. T. & Then, P. H. H. (2015). Information Seeking Features of a PICO-based Medical Question-Answering System: A Usability and Satisfaction Survey. *9th International Conference on IT in Asia*. (pp. 1-7). IEEE.

REFERENCES

Agoritsas, T., Merglen, A., Courvoisier, D.S., Combescure, C., Garin, N., Perrier, A. & Perneger, T.V., "Sensitivity and predictive value of 15 PubMed search strategies to answer clinical questions rated against full systematic reviews", *Journal of medical Internet research,* Vol. 14, No. 3, 2012, pp.

Ahluwalia, S., Murray, E., Stevenson, F., Kerr, C. & Burns, J., "'A heartbeat moment': qualitative study of GP views of patients bringing health information from the internet to a consultation", *British Journal of General Practice,* Vol. 60, No. 571, 2010, pp. 88-94.

Aljaber, B., Stokes, N., Bailey, J. & Pei, J., "Document clustering of scientific texts using citation contexts", *Information Retrieval,* Vol. 13, No. 2, 2010, pp. 101-131.

Anders, M.E. & Evans, D.P., "Comparison of PubMed and Google Scholar literature searches", *Respiratory care,* Vol. 55, No. 5, 2010, pp. 578-583.

Andrews, J.E., Pearce, K.A., Ireson, C. & Love, M.M., "Information-seeking behaviors of practitioners in a primary care practice-based research network (PBRN)", *Journal of the Medical Library Association,* Vol. 93, No. 2, 2005, pp. 206.

Aronson, A.R., "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program". In Proceedings of the AMIA Symposium, American Medical Informatics Association, 2001, pp. 17.

Athenikos, S.J. & Han, H., "Biomedical question answering: A survey", *Computer methods and programs in biomedicine,* Vol. 99, No. 1, 2010, pp. 1-24.

Attaran, A., Barnes, K.I., Curtis, C., D'alessandro, U., Fanello, C.I., Galinski, M.R., Kokwaro, G., Looareesuwan, S., Makanga, M. & Mutabingwa, T.K., "WHO, the Global Fund, and medical malpractice in malaria treatment", *The Lancet,* Vol. 363, No. 9404, 2004, pp. 237-240.

Bauer, M.A. & Berleant, D., "Usability survey of biomedical question answering systems", *Human genomics,* Vol. 6, No. 1, 2012, pp. 17.

Beel, J. & Gipp, B., "Google Scholar's ranking algorithm: the impact of citation counts (an empirical study)". In Research Challenges in Information Science, 2009. RCIS 2009. Third International Conference on, IEEE, 2009, pp. 439-446.

Bergus, G.R., Randall, C.S., Sinift, S.D. & Rosenthal, D.M., "Does the structure of clinical questions affect the outcome of curbside consultations with specialty colleagues?", *Archives of Family Medicine,* Vol. 9, No. 6, 2000, pp. 541.

Booth, A., O'rourke, A.J. & Ford, N.J., "Structuring the pre-search reference interview: a useful technique for handling clinical questions", *Bulletin of the Medical Library Association,* Vol. 88, No. 3, 2000, pp. 239.

Boudin, F., Nie, J.-Y., Bartlett, J.C., Grad, R., Pluye, P. & Dawes, M., "Combining classifiers for robust PICO element detection", *BMC medical informatics and decision making,* Vol. 10, No. 1, 2010, pp. 29.

Bramer, W.M., Giustini, D., Kramer, B.M. & Anderson, P., "The comparative recall of Google Scholar versus PubMed in identical searches for biomedical systematic reviews: a review of searches used in systematic reviews", *Systematic reviews,* Vol. 2, No. 1, 2013, pp. 1-9.

Brożek, J., Akl, E., Jaeschke, R., Lang, D., Bossuyt, P., Glasziou, P., Helfand, M., Ueffing, E., Alonso‐Coello, P. & Meerpohl, J., "Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies", *Allergy,* Vol. 64, No. 8, 2009, pp. 1109-1116.

Brunetti, L. & Hermes-Desantis, E., "The Internet as a drug information resource", *US Pharmacist,* Vol. 35, No. 1, 2010, pp.

Cao, Y., Liu, F., Simpson, P., Antieau, L., Bennett, A., Cimino, J.J., Ely, J. & Yu, H., "AskHERMES: An online question answering system for complex clinical questions", *Journal of biomedical informatics,* Vol. 44, No. 2, 2011, pp. 277-288.

Cook, D.A., Sorensen, K.J., Hersh, W., Berger, R.A. & Wilkinson, J.M., "Features of effective medical knowledge resources to support point of care learning: a focus group study", *PloS one,* Vol. 8, No. 11, 2013, pp. e80318.

Corcoran, A.T., Peele, P.B. & Benoit, R.M., "Cost comparison between watchful waiting with active surveillance and active treatment of clinically localized prostate cancer", *Urology,* Vol. 76, No. 3, 2010, pp. 703-707.

Couper, M.P., Singer, E., Levin, C.A., Fowler, F.J., Fagerlin, A. & Zikmund-Fisher, B.J., "Use of the Internet and ratings of information sources for medical decisions: results from the DECISIONS survey", *Medical Decision Making,* Vol. 30, No. 5 suppl, 2010, pp. 106S-114S.

Cullen, R., Clark, M. & Esson, R., "Evidence ‐ based information ‐ seeking skills of junior doctors entering the workforce: an evaluation of the impact of information literacy training during pre ‐ clinical years", *Health Information & Libraries Journal,* Vol. 28, No. 2, 2011, pp. 119-129.

Davies, K., "The information ‐ seeking behaviour of doctors: a review of the evidence", *Health Information & Libraries Journal,* Vol. 24, No. 2, 2007, pp. 78-94.

Davies, K., "UK doctors awareness and use of specified electronic evidence-based medicine resources", *Informatics for Health and Social Care,* Vol. 36, No. 1, 2011, pp. 1-19.

De Vito, C., Nobile, C.G., Furnari, G., Pavia, M., De Giusti, M., Angelillo, I.F. & Villari, P., "Physicians' knowledge, attitudes and professional use of RCTs and meta-analyses: a cross-sectional survey", *The European Journal of Public Health,* Vol. 19, No. 3, 2009, pp. 297-302.

Del Fiol, G., Workman, T.E. & Gorman, P.N., "Clinical questions raised by clinicians at the point of care: a systematic review", *JAMA internal medicine,* Vol. 174, No. 5, 2014, pp. 710-718.

Delbecque, T., Jacquemart, P. & Zweigenbaum, P., "Indexing UMLS semantic types for medical question-answering", *Studies in Health Technology and Informatics,* Vol. 116, No., 2005, pp. 805-810.

Demner-Fushman, D., Few, B., Hauser, S.E. & Thoma, G., "Automatically identifying health outcome information in MEDLINE records", *Journal of the American Medical Informatics Association,* Vol. 13, No. 1, 2006, pp. 52-60.

Demner-Fushman, D. & Lin, J., "Answer extraction, semantic clustering, and extractive summarization for clinical question answering". In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2006, pp. 841-848.

Demner-Fushman, D. & Lin, J., "Answering clinical questions with knowledge-based and statistical techniques", *Computational Linguistics,* Vol. 33, No. 1, 2007, pp. 63-103.

Demner-Fushman, D., Seckman, C., Fisher, C., Hauser, S.E., Clayton, J. & Thoma, G.R., "A prototype system to support evidence-based practice". In AMIA Annual Symposium Proceedings, American Medical Informatics Association, 2008, pp. 151.

Eaton, C. & Zhao, H. 2001. Visualizing Web Search Results.

Ebell, M.H., "The vitamin E saga: lessons in patient-oriented evidence", *American family physician,* Vol. 71, No. 11, 2005, pp. 2052, 2054.

Ebell, M.H., Siwek, J., Weiss, B.D., Woolf, S.H., Susman, J., Ewigman, B. & Bowman, M., "Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature", *The Journal of the American Board of Family Practice,* Vol. 17, No. 1, 2004, pp. 59-67.

Ely, J.W., Osheroff, J.A., Chambliss, M.L., Ebell, M.H. & Rosenbaum, M.E., "Answering physicians' clinical questions: obstacles and potential solutions", *Journal of the American Medical Informatics Association,* Vol. 12, No. 2, 2005, pp. 217-224.

Ely, J.W., Osheroff, J.A., Maviglia, S.M. & Rosenbaum, M.E., "Patient-care questions that physicians are unable to answer", *Journal of the American Medical Informatics Association,* Vol. 14, No. 4, 2007, pp. 407-414.

Ertek, G., Tapucu, D. & Arın, I., "Text mining with rapidminer", *RapidMiner: Data Mining Use Cases and Business Analytics Applications,* Vol., No., 2013, pp. 241.

Evans, D., "Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions", *Journal of clinical nursing,* Vol. 12, No. 1, 2003, pp. 77-84.

Fourie, I., "Learning from research on the information behaviour of healthcare professionals: a review of the literature 2004–2008 with a focus on emotion", *Health Information & Libraries Journal,* Vol. 26, No. 3, 2009, pp. 171-186.

Galili, T., "dendextend: an R package for scientific visualization of dendograms and hierarchical clustering", Vol., No., 2014, pp.

Giustini, D. & Barsky, E., "A look at Google Scholar, PubMed, and Scirus: comparisons and recommendations", *Journal of the Canadian Health Libraries Association/Journal de l'Association des bibliothèques de la santé du Canada,* Vol. 26, No. 3, 2005, pp. 85-89.

Giustini, D. & Boulos, M.N.K., "Google Scholar is not enough to be used alone for systematic reviews", *Online journal of public health informatics,* Vol. 5, No. 2, 2013, pp. 214.

Gordon, J., "Educating the patient: challenges and opportunities with current technology", *Nursing Clinics of North America,* Vol. 46, No. 3, 2011, pp. 341-350.

Grol, R., "Successes and failures in the implementation of evidence-based guidelines for clinical practice", *Medical care,* Vol. 39, No. 8, 2001, pp. II-46-II-54.

Hastings, C. & Fisher, C.A., "Searching for proof: Creating and using an actionable PICO question", *Nursing management,* Vol. 45, No. 8, 2014, pp. 9-12.

Haynes, R.B., Mckibbon, K.A., Wilczynski, N.L., Walter, S.D. & Werre, S.R., "Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey", *Bmj,* Vol. 330, No. 7501, 2005, pp. 1179.

Haynes, R.B. & Wilczynski, N.L., "Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey", *Bmj,* Vol. 328, No. 7447, 2004, pp. 1040.

Heighes, P.T. & Doig, G.S., "Intensive care specialists' knowledge, attitudes, and professional use of published research evidence: A mail-out questionnaire survey of appropriate use of research evidence in clinical practice", *Journal of critical care,* Vol. 29, No. 1, 2014, pp. 116-122.

Henderson, J., "Google Scholar: A source for clinicians?", *Canadian Medical Association Journal,* Vol. 172, No. 12, 2005, pp. 1549-1550.

Hoogendam, A., De Vries Robbe, P.F. & Overbeke, A.J.P., "Comparing patient characteristics, type of intervention, control, and outcome (PICO) queries with unguided searching: a randomized controlled crossover trial", *Journal of the Medical Library Association: JMLA,* Vol. 100, No. 2, 2012, pp. 121.

Hosmer, D.W. & Lemeshow, S.L. 2000. *Assessing the fit of the model. Applied logistic regression.*, John Wiley & Sons, 2nd edition, pp. 143-202

Huang, A., "Similarity measures for text document clustering". In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, 2008, pp. 49-56.

Huang, X., Lin, J. & Demner-Fushman, D., "Evaluation of PICO as a knowledge representation for clinical questions". In AMIA Annual Symposium Proceedings, American Medical Informatics Association, 2006, pp. 359.

Hutin, Y.J., Hauri, A.M. & Armstrong, G.L., "Use of injections in healthcare settings worldwide, 2000: literature review and regional estimates", *Bmj,* Vol. 327, No. 7423, 2003, pp. 1075.

Jones, G., Steketee, R.W., Black, R.E., Bhutta, Z.A., Morris, S.S. & Group, B.C.S.S., "How many child deaths can we prevent this year?", *The lancet,* Vol. 362, No. 9377, 2003, pp. 65-71.

Kim, J., "Task difficulty as a predictor and indicator of web searching interaction". In CHI'06 Extended Abstracts on Human Factors in Computing Systems, ACM, 2006, pp. 959-964.

Kim, J., "Perceived difficulty as a determinant of Web search performance", *Information Research,* Vol. 13, No. 4, 2008, pp.

Kosteniuk, J.G., Morgan, D.G. & D'arcy, C.K., "Use and perceptions of information among family physicians: sources considered accessible, relevant, and reliable", *Journal of the Medical Library Association: JMLA,* Vol. 101, No. 1, 2013, pp. 32.

Krause, R., Moscati, R., Halpern, S., Schwartz, D.G. & Abbas, J., "Can emergency medicine residents reliably use the internet to answer clinical questions?", *Western Journal of Emergency Medicine,* Vol. 12, No. 4, 2011, pp. 442.

Landis, J.R. & Koch, G.G., "The measurement of observer agreement for categorical data", *biometrics,* Vol., No., 1977, pp. 159-174.

Lappa, E., "Undertaking an information‐needs analysis of the emergency‐care physician to inform the role of the clinical librarian: a Greek perspective", *Health Information & Libraries Journal,* Vol. 22, No. 2, 2005, pp. 124-132.

Lechtenfeld, M. & Fuhr, N., "Result clustering supports users with vague information needs", Vol., No., 2012, pp.

Leuski, A., "Evaluating document clustering for interactive information retrieval". In Proceedings of the tenth international conference on Information and knowledge management, ACM, 2001, pp. 33-40.

Leuski, A. & Allan, J., "Interactive information retrieval using clustering and spatial proximity", *User Modeling and User-Adapted Interaction,* Vol. 14, No. 2-3, 2004, pp. 259-288.

Lin, Y.-S., Jiang, J.-Y. & Lee, S.-J., "A similarity measure for text classification and clustering", *IEEE Transactions on Knowledge and Data Engineering,* Vol., No., 2013, pp. 1-15.

Lu, Z., Kim, W. & Wilbur, W.J., "Evaluation of query expansion using MeSH in PubMed", *Information retrieval,* Vol. 12, No. 1, 2009, pp. 69-80.

Lucas, B.P., Evans, A.T., Reilly, B.M., Khodakov, Y.V., Perumal, K., Rohr, L.G., Akamah, J.A., Alausa, T.M., Smith, C.A. & Smith, J.P., "The impact of evidence on physicians' inpatient treatment decisions", *Journal of general internal medicine,* Vol. 19, No. 5p1, 2004, pp. 402-409.

Malik, H.H. & Kender, J.R., "High quality, efficient hierarchical document clustering using closed interesting itemsets". In Data Mining, 2006. ICDM'06. Sixth International Conference on, IEEE, 2006, pp. 991-996.

Markey, P. & Schattner, P., "Promoting evidence-based medicine in general practice—the impact of academic detailing", *Family Practice,* Vol. 18, No. 4, 2001, pp. 364-366.

Mckibbon, K.A., Wilczynski, N.L. & Haynes, R.B., "What do evidence-based secondary journals tell us about the publication of clinically important articles in primary healthcare journals?", *BMC medicine,* Vol. 2, No. 1, 2004, pp. 33.

Mcmullan, M., "Patients using the Internet to obtain health information: how this affects the patient–health professional relationship", *Patient education and counseling,* Vol. 63, No. 1, 2006, pp. 24-28.

Methley, A.M., Campbell, S., Chew-Graham, C., Mcnally, R. & Cheraghi-Sohi, S., "PICO, PICOS and SPIDER: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews", *BMC health services research,* Vol. 14, No. 1, 2014, pp. 579.

Meyer, D. & Buchta, C. "Package 'Proxy'" [Online]. The R project for statistical computing. Available: http://www.r-project.org/. Assessed: 2014.

Montori, V.M., Wilczynski, N.L., Morgan, D. & Haynes, R.B., "Optimal search strategies for retrieving systematic reviews from Medline: analytical survey", *Bmj,* Vol. 330, No. 7482, 2005, pp. 68.

Moyer, V. & Neuspiel, D.R., "PICO Questions: What Are They and Why Bother?", *AAP Grand Rounds,* Vol. 31, No. 5, 2014, pp. 50-50.

Murtagh, F. & Contreras, P., "Algorithms for hierarchical clustering: an overview", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* Vol. 2, No. 1, 2012, pp. 86-97.

Niu, Y. & Hirst, G., "Analysis of semantic classes in medical text for question answering". In Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains, 2004, pp. 54-61.

Niu, Y., Hirst, G., Mcarthur, G. & Rodriguez-Gianolli, P., "Answering clinical questions with role identification". In Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13, Association for Computational Linguistics, 2003, pp. 73-80.

Niu, Y., Zhu, X. & Hirst, G., "Using outcome polarity in sentence extraction for medical question-answering". In AMIA Annual Symposium Proceedings, American Medical Informatics Association, 2006, pp. 599.

Nixon, J., Wolpaw, T., Schwartz, A., Duffy, B., Menk, J. & Bordage, G., "SNAPPS-Plus: An Educational Prescription for Students to Facilitate Formulating and Answering Clinical Questions", *Academic Medicine,* Vol. 89, No. 8, 2014, pp. 1174-1179.

Owens, D.K., Lohr, K.N., Atkins, D., Treadwell, J.R., Reston, J.T., Bass, E.B., Chang, S. & Helfand, M., "AHRQ series paper 5: grading the strength of a body of evidence when comparing medical interventions—Agency for Healthcare Research and Quality and the Effective Health-Care Program", *Journal of clinical epidemiology,* Vol. 63, No. 5, 2010, pp. 513-523.

Punitha, S., Mugunthadevi, K. & Punithavalli, M., "Impact of ontology based approach on document clustering", *International Journal of Computer Applications (0975–8887) Volume,* Vol., No., 2011, pp.

Puspitasari, I., Moriyama, K., Fukui, K.I. & Numao, M., "Effects of Individual Health Topic Familiarity on Activity Patterns During Health Information Searches", *JMIR medical informatics,* Vol. 3, No. 1, 2015, pp.

Qu, P., Liu, C. & Lai, M., "The effect of task type and topic familiarity on information search behaviors". In Proceedings of the third symposium on Information interaction in context, ACM, 2010, pp. 371-376.

Rangrej, A., Kulkarni, S. & Tendulkar, A.V., "Comparative study of clustering techniques for short text documents". In Proceedings of the 20th international conference companion on World wide web, ACM, 2011, pp. 111-112.

Richardson, W.S., Wilson, M.C., Nishikawa, J. & Hayward, R.S., "The well-built clinical question: a key to evidence-based decisions", *ACP J Club,* Vol. 123, No. 3, 1995, pp. A12-3.

Rzany, B., "Formulating well-built clinical questions", *Evidence-based Dermatology,* Vol., No., 2009, pp. 35.

Saca-Hazboun, H., "Empowering patients with knowledge. An update on trends in patient education", *ONS connect,* Vol. 22, No. 5, 2007, pp. 8.

Sackett, D.L., Rosenberg, W., Gray, J., Haynes, R.B. & Richardson, W.S., "Evidence based medicine: what it is and what it isn't", *Bmj,* Vol. 312, No. 7023, 1996, pp. 71-72.

Sadeghi‐Bazargani, H., Tabrizi, J.S. & Azami‐Aghdash, S., "Barriers to evidence‐based medicine: a systematic review", *Journal of evaluation in clinical practice,* Vol. 20, No. 6, 2014, pp. 793-802.

Schardt, C., Adams, M.B., Owens, T., Keitz, S. & Fontelo, P., "Utilization of the PICO framework to improve searching PubMed for clinical questions", *BMC medical informatics and decision making,* Vol. 7, No. 1, 2007, pp. 16.

Schilling, L.M., Steiner, J.F., Lundahl, K. & Anderson, R.J., "Residents' patient-specific clinical questions: opportunities for evidence-based learning", *Academic Medicine,* Vol. 80, No. 1, 2005, pp. 51-56.

Schuster, M.A., Mcglynn, E.A. & Brook, R.H., "How good is the quality of health care in the United States?", *Milbank Quarterly,* Vol. 76, No. 4, 1998, pp. 517-563.

Schwartz, K., Northrup, J., Israel, N., Crowell, K., Lauder, N. & Neale, A.V., "Use of on-line evidence-based resources at the point of care", *FAMILY MEDICINE-KANSAS CITY-,* Vol. 35, No. 4, 2003, pp. 251-256.

Scott, I., Heyworth, R. & Fairweather, P., "The use of evidence‐based medicine in the practice of consultant physicians. Results of a questionnaire survey", *Australian and New Zealand journal of medicine,* Vol. 30, No. 3, 2000, pp. 319-326.

Shariff, S.Z., Bejaimal, S.A., Sontrop, J.M., Iansavichus, A.V., Haynes, R.B., Weir, M.A. & Garg, A.X., "Retrieving clinical evidence: a comparison of PubMed and Google Scholar for quick clinical searches", *Journal of medical Internet research,* Vol. 15, No. 8, 2013, pp.

Shuval, K., Shachak, A., Linn, S., Brezis, M., Feder-Bubis, P. & Reis, S., "The impact of an evidence-based medicine educational intervention on primary care physicians: a qualitative study", *Journal of general internal medicine,* Vol. 22, No. 3, 2007, pp. 327-331.

Smith, R., "What clinical information do doctors need?", *Bmj,* Vol. 313, No. 7064, 1996, pp. 1062-1068.

Staunton, M., "Evidence-based Radiology: Steps 1 and 2—Asking Answerable Questions and Searching for Evidence 1", *Radiology,* Vol. 242, No. 1, 2007, pp. 23-31.

Straus, S.E., Ball, C., Balcombe, N., Sheldon, J. & Mcalister, F.A., "Teaching Evidence‐based Medicine Skills Can Change Practice in a Community Hospital", *Journal of general internal medicine,* Vol. 20, No. 4, 2005, pp. 340-343.

Strehl, A., Ghosh, J. & Mooney, R., "Impact of similarity measures on web-page clustering". In Workshop on Artificial Intelligence for Web Search (AAAI 2000), 2000, pp. 58-64.

Subhashini, R. & Kumar, V.J.S., "Evaluating the performance of similarity measures used in document clustering and information retrieval". In Proceedings of 1st International Conference on Integrated Intelligent Computing (ICIIC): August, 2010, pp. 5-7.

Sultan, S., Falck–Ytter, Y. & Inadomi, J.M., "The AGA institute process for developing clinical practice guidelines part one: grading the evidence", *Clinical Gastroenterology and Hepatology,* Vol. 11, No. 4, 2013, pp. 329-332.

Ulvenes, L.V., Aasland, O., Nylenna, M. & Kristiansen, I.S., "Norwegian physicians' knowledge of and opinions about evidence-based medicine: cross-sectional study", *PLoS One,* Vol. 4, No. 11, 2009, pp. e7828.

Us National Library of Medicine. "Clinical Questions Collection" [Online]. Available: http://clinques.nlm.nih.gov/. Assessed: 2015.

Us National Library of Medicine. "Abridged index medicus list of journals indexed" [Online]. Abridged Index Medicus. Available: http://www.nlm.nih.gov/bsd/aim.html. Assessed: 2015.

Vong, W.-T. & Then, P.H.H. 2014. Visualization of PICO Elements for Information Needs Clarification and Query Refinement. *Advances in Knowledge Discovery and Data Mining.* Springer.

Wallace, J., Nwosu, B. & Clarke, M., "Barriers to the uptake of evidence from systematic reviews and meta-analyses: a systematic review of decision makers' perceptions", *BMJ open,* Vol. 2, No. 5, 2012, pp. e001220.

Wang, P., Bownas, J. & Berry, M.W. 2004. Trend and behavior detection from Web queries. *Survey of Text Mining.* Springer.

Weiming, W., Hu, D., Feng, M. & Wenyin, L., "Automatic clinical question answering based on UMLS relations". In Third International Conference on Semantics, Knowledge and Grid, Citeseer, 2007, pp. 495-498.

Wilczynski, N.L. & Haynes, R.B., "Developing optimal search strategies for detecting clinically sound prognostic studies in MEDLINE: an analytic survey", *BMC medicine,* Vol. 2, No. 1, 2004, pp. 23.

Wilczynski, N.L., Haynes, R.B. & Team, H., "Developing optimal search strategies for detecting clinically sound causation studies in MEDLINE". In AMIA annual symposium proceedings, American Medical Informatics Association, 2003, pp. 719.

Wong, S.S.-L., Wilczynski, N.L., Haynes, R.B., Ramkissoonsingh, R. & Team, H., "Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE". In AMIA Annual Symposium Proceedings, American Medical Informatics Association, 2003, pp. 728.

Wurst, M. & Mierswa, I., "The word vector tool and the rapidminer text plugin", *Dortmund, Germany,* Vol., No., 2007, pp.

Yu, H. & Cao, Y.-G., "Automatically extracting information needs from ad hoc clinical questions". In AMIA annual symposium proceedings, American Medical Informatics Association, 2008, pp. 96.

Yu, H. & Kaufman, D., "A cognitive evaluation of four online search engines for answering definitional questions posed by physicians". In Pacific Symposium on Biocomputing, 2007, pp. 328-339.

Zhao, Y. & Karypis, G., "Evaluation of hierarchical clustering algorithms for document datasets". In Proceedings of the eleventh international conference on Information and knowledge management, ACM, 2002, pp. 515-524.

Zhu, Z., Cox, I.J. & Levene, M. 2008. Ranked-listed or categorized results in IR: 2 is better than 1. *Natural Language and Information Systems.* Springer.

Zwolsman, S., Te Pas, E., Hooft, L., Wieringa-De Waard, M. & Van Dijk, N., "Barriers to GPs' use of evidence-based medicine: a systematic review", *British Journal of General Practice,* Vol. 62, No. 600, 2012, pp. e511-e521.