Asia Pacific Journal of Information Systems
Vol. 25 No. 4 (DECEMBER 2015), 686-711

ISSN 2288-5404 (Print) / ISSN 2288-6818 (Online)
http://dx.doi.org/10.14329/apjis.2015.25.4.686

Known-Item Retrieval Performance of a PICO-based Medical Question Answering Engine

Wan-Tze Vong[a,*], Patrick Hang Hui Then[b]

[a] *Ph.D. Candidate, Faculty of Engineering, Computing and Science, Swinburne University of Technology, Malaysia*
[b] *Associate Professor, Faculty of Engineering, Computing and Science, Swinburne University of Technology, Malaysia*

ABSTRACT

The performance of a novel medical question-answering engine called CliniCluster and existing search engines, such as CQA-1.0, Google, and Google Scholar, was evaluated using known-item searching. Known-item searching is a document that has been critically appraised to be highly relevant to a therapy question. Results show that, using CliniCluster, known-items were retrieved on average at rank 2 (*MRR@10* $\approx$ *0.50*), and most of the known-items could be identified from the top-10 document lists. In response to ill-defined questions, the known-items were ranked lower by CliniCluster and CQA-1.0, whereas for Google and Google Scholar, significant difference in ranking was not found between well- and ill-defined questions. Less than 40% of the known-items could be identified from the top-10 documents retrieved by CQA-1.0, Google, and Google Scholar. An analysis of the top-ranked documents by strength of evidence revealed that CliniCluster outperformed other search engines by providing a higher number of recent publications with the highest study design. In conclusion, the overall results support the use of CliniCluster in answering therapy questions by ranking highly relevant documents in the top positions of the search results.

Keywords: Known-Item Search, Mean Reciprocal Rank, Pico Elements, Question-Answering Engine, Strength Of Evidence

Ⅰ. Introduction

Known-item search is the task where users look for a particular document from a result set in response to a query. In the current study, a set of question-document pairs was collected from a database of filtered, synopsized, evidence-based information for clinical decisions. Each pair contains a therapy question and a document that contains the most valid and relevant clinical information to answer the question. The paired documents, which serve as the known-items, were assumed to be more valuable than marginally relevant documents and were more likely to be ranked in higher positions of the search results. Besides, it was assumed that a user stops going through a ranked list of documents after finding

*Corresponding Author. E-mail: wvong@swinburne.edu.my Tel: 6082416353

one highly relevant document. Therefore, the higher the ranking of a known-item, the better the retrieval performance of a search engine.

A known-item search task was performed by submitting a question to a search engine and the paired document is identified from the resulting list. Both well- and ill-defined questions and five different structural patterns of questions were submitted to a prototype clinical question-answering engine called Clini Cluster and three existing search engines (CQA-1.0, Google and Google Scholar). Users generally look for the first 10 or 20 documents retrieved by a search engine only. Therefore, the known-items were identified from the top-10 and top-20 documents. The performance of the search engines was compared by determining the ranked positions of known-items, the percentage of known-items identified and the quality of evidence provided by the top-ranked documents.

The remainder of the paper is organized as follows. Section II introduces the search strategies that have been used to determine the quality of clinical evidence. The approaches and resources that have been used to develop medical question-answering (MedQA) systems are briefly reviewed in Section III. Sections IV and V discussed the search engines and the methods used for the evaluation of known-item search task. The results obtained and the limitations of the study were discussed, respectively, in Sections VI and VII.

Ⅱ. Background

Physicians seek clinical information to answer patient-specific questions, to stay current with new medical developments, to review previously learned information and to keep up with specific area of interest (Shaughnessy et al., 1994). To ensure that the best care is delivered to patients, a new paradigm for medical practice, called Evidence-Based Medicine (EBM), has been developed. The process of EBM involves four steps. The steps aim to ensure that the best available evidence from research studies, integrated with clinical expertise and values, is used to make and support clinical decision-making (Sackett et al., 1996).

Numerous barriers have been identified that cause the uptake of clinical evidence by physicians slow and reluctant. The barriers include lack of time, limited literature searching skills, the tendency to formulate unanswerable questions and a lack of awareness of the information needs (Lappa, 2005; Davies, 2007; Ely et al., 2007; Zwolsman et al., 2012). Therefore, to better serve the information needs of physicians practicing EBM, MedQA systems such as CQA-1.0 (Demner-Fushman and Lin, 2007) and AskHERMES (Cao et al., 2011) have emerged as the next generation search engines. The systems, different to common literature search engines such as PubMed and Google Scholar, aim to provide the most relevant and valid information that can be assessed quickly as answers for clinical practice.

2.1. Question Formulation

The first step of EBM is to convert an information need from practice into a focus and well-structured question.

2.1.1. The PICO Framework

To formulate an answerable question, physicians are recommended to change their search strategies by rephrasing their questions (Ely et al., 2007) or to use question/query frameworks. The PICO frame-

work has been widely accepted for the formulation of well-defined and answerable clinical questions (Schardt et al., 2007; Staunton, 2007). For instance, the question "*In children with acute asthma exacerbations, is oral or injected dexamethasone as effective as predisone or prednisolone*?" is broken down into four parts:

P: *children with acute asthma exacerbations*
I : *oral or injected dexamethasone*
C: *predisone or prednisolone*
O: -

P stands for population or problem that gives information about a group of patients and the primary problem, disease or co-existing condition that requires physicians' care. I stands for intervention that describes the treatment of interest. C stands for comparison and is an alternative to the intervention of interest. O stands for outcome that gives information about the results of an intervention.

Other question frameworks that have been introduced recently include PESICO (Schlosser et al., 2007), PICOS (Atkins et al. 2011), PICOT (Rios et al. 2010) and SPIDER (Cooke et al. 2012). Despite of these different frameworks, 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 answer or the relevancy of search results. In this regard, it seems worthwhile to continue to use PICO for the formulation of clinical questions.

2.1.2. Structural Patterns of Therapy Questions

One of the physicians' greatest information needs is for information about treatment and drugs (Davies, 2007; Schwartz et al., 2003; Smith, 1996; Yu and Cao, 2008). A study by Huang et al. (2006), who explored the clinical questions posed by physicians, concluded that the PICO framework is particularly useful for representing therapy questions. The authors identified five common structural patterns of therapy questions (<Table 1>): Patterns I and II are the most common, and Patterns III to V are less common. A question mark indicates the element that serves as the answer to a question. For example, [O?] indicates that the "outcome" of an intervention is the desired answer for Pattern I. The five patterns show that not all therapy questions have all four PICO elements present. For examples, Pattern II contains a [P] element and the [I?] element serves as the answer of interest, and among the five patterns, only Pattern V contains the [C] element.

<Table 1> Five Structural Patterns of Therapy Questions

Pattern	PICO Structure	An Example
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 duloxetine effective in reducing pain from chemotherapy-induced peripheral neuropathy in adult cancer survivors?
V	[P] [I] [C] [O?]	What is the comparative effectiveness of ondansetron and metoclopramide for the treatment of hyperemesis gravidarum?

A poorly formulated question can lead to the discovery of irrelevant documents. In the present study, therapy questions framed with inadequate number of PICO elements (i.e., ill-defined questions) were evaluated in terms of their performance in retrieving highly relevant documents.

2.2. Document Appraisal

The second and third steps of EBM involve a comprehensive search of literature and critical appraisal of the validity and applicability of research evidence. Physicians are advised to look for the most useful information by finding patient-oriented evidence (POEs) and determining the design of a study.

2.2.1. Patient-Oriented Evidence

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 that contain POEs 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 are found by McKibbon et al. (2004) who identified the "number of articles needed to be read" (NNR) in 170 primary healthcare journals. Both studies concluded that POEMs are concentrated in a small subset of journals. On the other hand, MEDLINE provides the "Core Clinical Journals" filter to restrict literature search to 119 journals particularly relevant to practicing physicians (US National Library of Medicine, 2014). In this regard, previous studies suggest that the search of the most useful evidence for clinical practice can be improved by focusing on journals that publish POEMs.

2.2.2. Type of Clinical Study

For clinical recommendations regarding treatment, prevention or screening, the quality of POEs from a clinical study can be determined as indicated in <Table 2>. Recommendations should be made based on the highest quality evidence available. As reported in a paper by Ebell (2005), vitamin E was found in some case-control studies (Level 2 study quality) to slow functional decline for patient with Alzheimer's disease, but good quality randomized control trials (Level 1 study quality) have not confirmed this benefit. Therefore, recommendations should be made based on the Level 1 studies. The example explains the importance of considering the study design when determining the quality of evidence provided by a clinical study.

The clinical query filters in PubMed are intended to retrieve citations related to specific clinical research areas and to avoid information overload. The filters

<Table 2> Level of Evidence by Study Design

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

retrieve five categories of studies (etiology, diagnosis, therapy, prognosis and clinical predication guides) with two options (a broad or a narrow search) (Haynes et al., 2005; Haynes and Wilczynski, 2004; Montori et al., 2005; Wilczynski et al., 2003; Wilczynski and Haynes, 2004; Wong et al., 2003). A broad search for "therapy" studies returns a higher number of randomized controlled trials (RCTs) than a narrow search. On the other hand, the "systematic reviews" clinical query filter allows the search of the highest quality studies such as meta-analyses and reviews of clinical trials (Montori et al., 2005).

The last step of EBM is the implementation of useful findings in clinical practice. To ensure that the most useful information can be identified rapidly by physicians, the most recent studies published in 119 core clinical journals were assigned a higher weight in the present study. The purpose is to rank studies that published the most up-to-date and the highest quality POEs in higher positions of the search result lists.

Ⅲ. MedQA 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. A brief review of current MedQA systems by three processing phases is described as follows.

3.1. Question Processing

In this phase, a question, generally in natural language, is input to a QA system. Current MedQA systems are limited by their ability to process only certain types and formats of questions (Athenikos and Han, 2010). The Demner-Fushman et al.'s InfoBot system (2008) accepts only structured PICO 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), the CQA-1.0 system (the later version of the InfoBot system) requires users to clearly identifies each component of PICO as the input query. Users will need to have a clear understanding of the PICO framework and the terminology of a specialized domain in order to pose a question to the systems.

A question is transformed into a search query in canonical form, which is then served as the input to a document retrieval engine. The Delbecque et al. (2005)'s, Demner-Fushman et al. (2006a)'s, Niu et al. (2006)'s, and Weiming et al. (2007)'s QA systems extract UMLS semantic concepts and relations from the input natural language question or PICO query as search query terms. Much effort has been put on identifying and expanding query terms for the search of relevant documents. Previous study demonstrated a lack of key medical concepts that comprise a well-formed query in questions posed by physicians (Huang et al., 2006; Thabane et al., 2009). More research needs to be done to enable more complicated analysis of ill-defined questions. For example, can a QA system returns information that best meets the needs of users when a question is formulated with only one of the four PICO elements?

3.2. Document Processing

The query generated from question processing phase is submitted to a Web-based or a Corpus-based search engine to retrieve relevant documents in the

document processing phase. Delbecque et al. (2005)'s and Niu et al. (2006)'s use Google and the XML document database respectively to retrieve relevant documents. Demner-Fushman et al. (2006) use domain-specific search engine, PubMed, to retrieve medical literature from MEDLINE database. Weiming et al. (2007)'s use Lucene, a standard information retrieval engine, to retrieve documents from the Web and from the MEDLINE database. Yu and Kaufman (2007) recommended the use of both Web-based and Corpus-based search machines 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 the MeSH terms and the Clinical Query Filters) for users to perform a more efficient search (Anders and Evans, 2010; Bramer et al., 2013; Henderson, 2005). Besides, 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 seems worthwhile to continue to use PubMed for the retrieval of relevant documents.

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 identify 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. A review of four MedQA 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 (Delbecque et al., 2005; Demner-Fushman et al., 2006; Niu et al., 2006; Weiming et al., 2007).

3.3. Answer Processing

In the answer processing phase, answers are generated by matching query from 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 (Demner-Fushman and Lin, 2007; Niu et al., 2006; Weiming et al., 2007). Delbecque et al. (2005), on the other hand, quantifies the co-occurrence of semantic types in candidate documents and selects tagged clauses as answers. In current semantic MedQA system, 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 handle conflicting evidence and for appropriate presentation of answers.

Ⅳ. Known-Item Search

Four search engines were evaluated in this study for their performance in retrieving and ranking known-items.

4.1. Search Engines

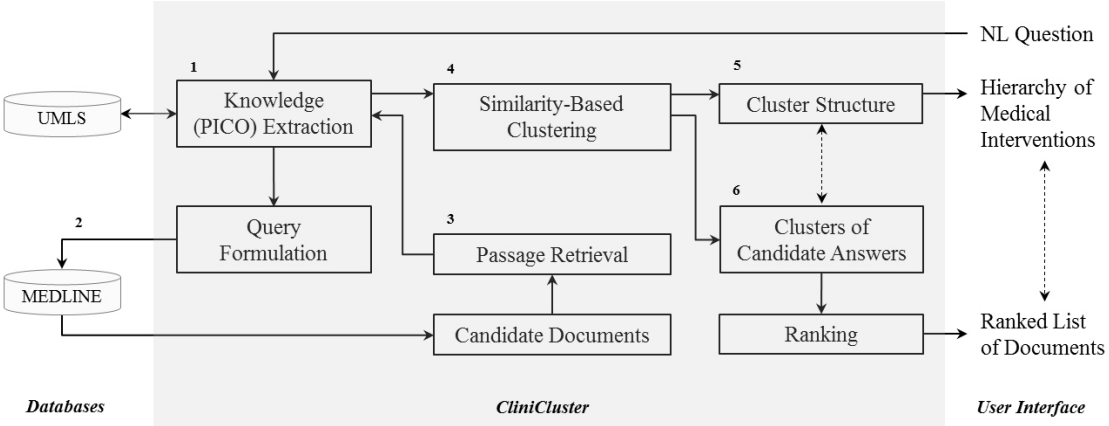
4.1.1. CliniCluster

It was proposed in our previous study (Vong and Then, 2014) that a hierarchical structure of medical interventions has the potential to assist users in explor-

ing a problem domain and in understanding their information needs. The hierarchy was implemented into a prototype engine called "CliniCluster". The architecture of the engine is demonstrated in <Figure 1>:

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 (Aronson, 2001). 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 <Appendix A> are recognized as the PICO elements.
2. The PICO elements are used as the search terms to retrieve relevant documents from the MEDLINE database. The concepts are automatically expanded in PubMed and clinical query filters are used to improve the search of therapy studies, particularly RCTs, and systematic reviews and meta-analyses.
3. 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.
4. The I and C elements are called jointly the "intervention". Each document is represented by a bag of interventions. The similarities between bags of interventions are calculated using Yule2 metric. Candidate documents are grouped into a tree of clusters using Ward-link clustering algorithm based on the calculated similarities.
5. A hierarchy of medical interventions is constructed and displayed to the users as an information seeking feature. Each cluster of the hierarchy contains documents with similar interventions and is labelled with therapy topic that appear the most frequent among the documents.
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 so that the most recent studies published in 119 core clinical journals and with the highest quality study design appear in the top positions of the result lists

The user interface of CliniCluster is shown in

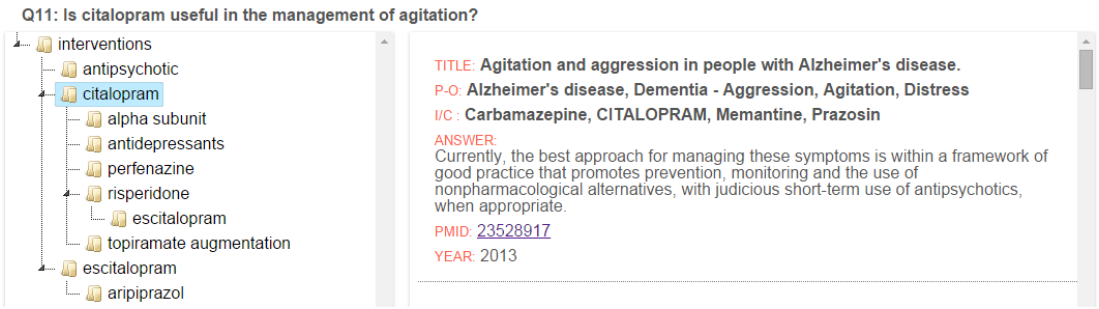


<Figure 1> Architecture of the Proposed CliniCluster Engine

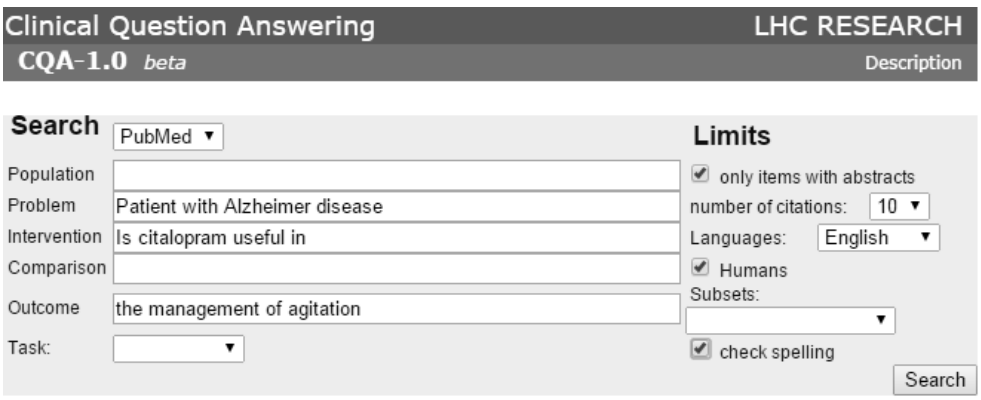
<Figure 2>. By posing a natural language question, two information seeking features are returned to the users. Feature 1: A hierarchy of medical interventions is displayed at the left side of the interface. It is expected that, by browsing through or exploring the hierarchy, users can gain a better understanding of the medical terminology related to the question posed. Feature 2: A ranked list of answers tagged with the 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.

4.1.2. CQA-1.0

CQA-1.0 is a clinical question-answering system developed for physicians practicing EBM. The homepage of CQA-1.0 (<Figure 3>) 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



<Figure 2> The User Interface of CliniCluster



<Figure 3> An Example of Broad Search Using CQA-1.0

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). The system utilizes the PICO framework to capture the information needs of users. 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. This may in turn lead to poor search results.

4.1.3. 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 (Cook et al., 2013; Giustini, 2005). 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).

4.2. Known-Item Search

The search tasks involved three key steps: construct question-document pairs, pose question to search engines and search for known-items.

4.2.1. Question-Document Pairs

70 POEMs were collected from the Essential Evidence Plus database (2015). Each POEM, as shown in the Appendix-B, 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. It was selected 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, and the article serves as the "known-item".

4.2.2. Test Questions

A total of 70 question-document pairs were collected. 30 ill-defined questions were created by removing one or two of the PICO elements from the original questions, as shown in <Figure 4>. The ill-defined questions were matched with the known-items from the original question-document pairs. This allows a comparison of search results obtained using original questions to those obtained using ill-defined questions. Besides, 40 questions categorized into five structural patterns were evaluated (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.

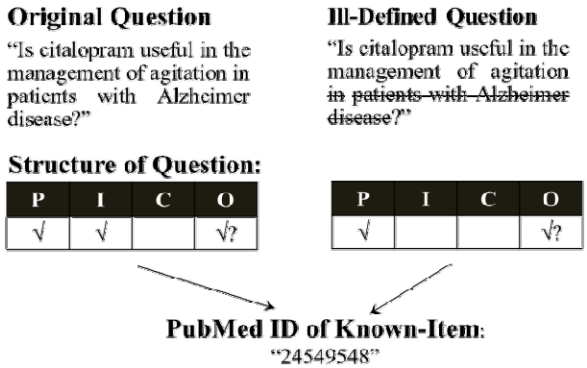
4.2.3. Document Retrieval

The test questions were posed respectively to the 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 3>) 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.

4.2.4. Interactive and Non-interactive Searches

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 5 (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.



<Figure 4> An Example of How an Ill-Defined Question is Created

Therefore, C_{1b} labelled with the most relevant topic "citalopram" is selected. The ranking of the known-item (k) increased 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 5 (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 increased from 6 to 1, and
2. By assuming that the "wrong" child cluster is chosen, a ranking of 0 is given to k.

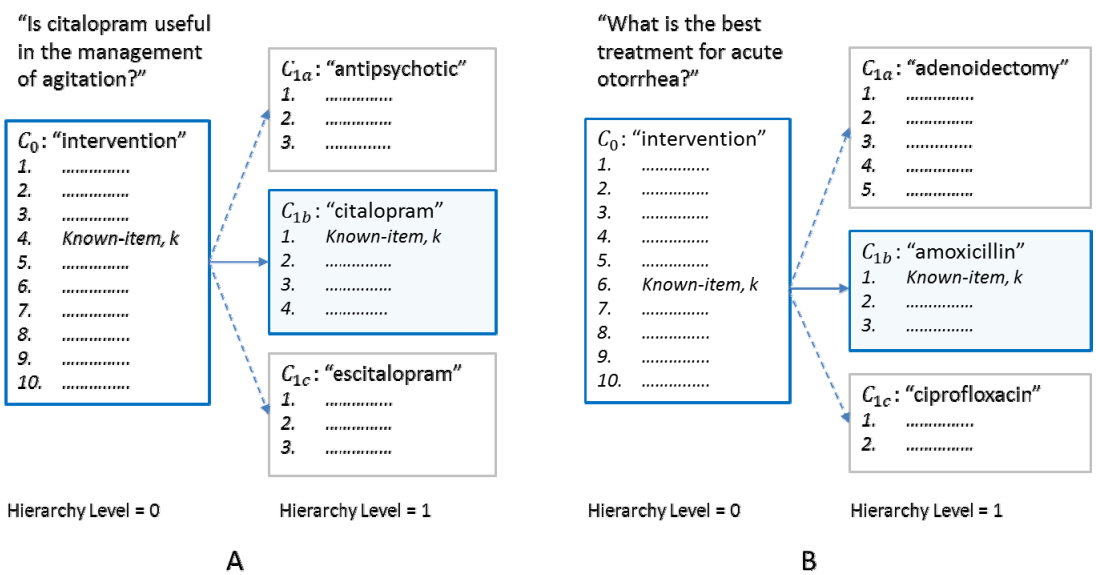
Non-interactive search: The known-items were identified from the ranked lists of top documents returned by Google, Google Scholar and CQA-1.0. The items were searched without exploiting the hierarchy returned by CliniCluster. This was achieved by retrieving all the relevant documents appear in the root nodes, which are called the "interventions" in the hierarchy of medical interventions (<Figure 2>).

Ⅴ. Performance Measures

Mean reciprocal rank, percentage gain and strength of evidence were calculated to compare the performance of the search engines.

5.1. Mean Reciprocal Rank

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. 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.



<Figure 5> Interactive Search of a Known-Item

$$MRR = \frac{1}{n}\sum_{i-1}^{n} = \frac{1}{rank_i} \quad (1)$$

As shown in (1), 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 $1/10 = 1.00$. The effectiveness of a search engine increases as the *MRR* approaches 1.00. A search engine that receives a *MRR* of 0.75 would mean that on average the engine finds the known-items between rank 1 and rank 2. A search engine 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.

5.2. 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 (2):

$$PG = \frac{N_c}{N_t} \times 100\% \quad (2)$$

where N_c is the number of questions correctly answered and N_t is the total number of questions in a test set. The measure indicates the percentage of known-items ranked as the top-10 documents.

5.3. Strength of Evidence

The strength of evidence (S_{SOE}) score, as shown in (3) and introduced by Demner-Fushman and Lin (2007), was used to indicate how well a document provides valid and reliable clinical evidence. A S_{SOE} score was assigned to each of the top-10 documents. The S_{Date} measures the recency of a document using (4). The S_{Study} is measured based on the design of a study. According to the hierarchy of evidence recommended by Evans (2003), systematic reviews and meta-analyses receive a score of 0.5; 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 journals listed in the Abridged Index Medicus (US National Library of Medicine, 2014) receive a score of 0.5, and 0 otherwise. For examples, a double-blind randomized controlled trial published in *N Engl J Med* on year 2013 obtains a score of 0.40 + 0.50 - 0.02 = 0.88.

$$S_{SOE} = S_{Date} + S_{Study} + S_{Journal} \quad (3)$$

$$S_{Date} = \frac{(Year_{publication} - Year_{current})}{100} \quad (4)$$

VI. Results and Discussion

6.1. Mean Reciprocal Rank

6.1.1. Original versus 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 3>. The results show that:

1. CliniCluster performed remarkably better than other search engines. A *MRR@10* of 0.54 indicates that the know-items appear on average in rank 2 of the result lists. Besides, CliniCluster is more likely to rank known-items in higher positions, followed by CQA-1.0 (narrow). The performance of Google and Google Scholar were the weakest, with MRR@10 scores less than 0.30.
2. There is no or only a slight difference between the *MRR@10* and *MRR@20* scores for each search engine, and the lowest score, 0.20 was given by Google. The results suggest that the majority of the known-items can be identified using the top-10 lists.
3. 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, when parts of the PICO elements appeared in original questions were removed. However, the MRR@10 scores for Google and Google Scholar remain similar when the same analyses were performed. The results indicate that the known-items were ranked lower when ill-defined questions were submitted to CliniCluster and CQA-1.0.

6.1.2. Five Structural Patterns of Questions

Similar analysis was carried out using five structural patterns of therapy questions. The results were compared using *MRR@10*. As shown in <Table 4>, 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 appear on average between rank 1 and rank 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 rank 2 (average *MRR* = 0.49). Besides, a narrow search using CQA-1.0 is better than a broad search (average *MRR* = 0.20 and 0.16, respectively), whereas similar results were obtained for both Google Scholar and Google (average *MRR* = 0.11 and 0.12, respectively).

Examples of the five patterns of questions are given in Appendix-C. 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

<Table 3> MRR@10 and MRR@20 of 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

<Table 4> MRR@10 and Average Rank Position of the Five Patterns of Therapy Questions

Search Engine	*MRR@10* *(Average 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)

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 in the present study that the known-items were more likely to be ranked higher, in response to questions that contain an [I] and an [O] element.

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 in 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 tends to rank known-

items in higher positions than other search engines.

6.1.3. 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 there is no [I] or [C] element that appears in 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 can be identified between rank 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 (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. The CliniCluster engine 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 described their information needs.

6.2. Percentage Gain

6.2.1. Original versus Ill-Defined Questions

A question is assumed to be correctly answered if the paired known-item is in the top-10 list. The percentage of questions correctly answered was interpreted using percentage gain in <Table 5>. 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 5> Percentage Gain (*PG*) of Original and Ill-Defined Questions

Search Engine	*PG* (%)	
	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

6.2.2. 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>, 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.

6.3. 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 7> 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

<Table 6> Percentage Gain (*PG*) of Five Patterns of Therapy Questions

Search Engine	*PG* (%)					
	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

than half of the top 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.

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 using CQA-1.0 by selecting the "systematic reviews" subset from the user interface. 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.

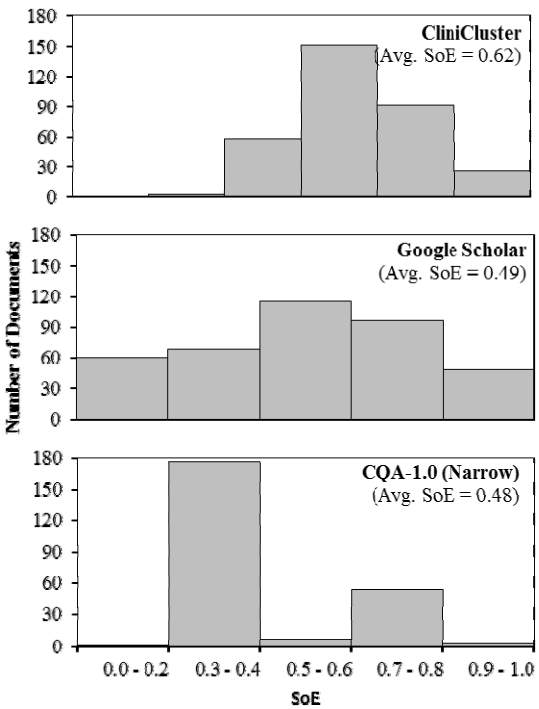
Besides, it was shown in the previous sections that, using the top-10 lists, CliniCluster returned a greater number of known-items than CQA-1.0 (narrow). CQA-1.0 uses a more complicated algorithm in ranking relevant documents (Demner-Fushman and Lin, 2007). Documents are weighted by matching a question 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 categorized relevant documents into different clusters using similarity-based clustering method and the documents in each cluster are ranked based on their strength of evidence, as described in Steps 4 and 6 of <Figure 1>. A comparison of CliniCluster and CQA-1.0 using the quality indicators described in <Table 7> suggests that the ranking and retrieval of known-items rely more heavily on the study design and the year of publication of clinical studies.

As reported by Beel and Gipp (2009), the highest weighed factor in Google Scholar ranking algorithm is the citation counts. The higher the citation count, the more likely that a document is being ranked in the top position of 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 percentage gain, CliniCluster was found to perform much better than Google Scholar in known-item retrieval. Using the quality indicators presented in <Table 7>, the results showed that CliniCluster returned a higher number of recent publications and systematic reviews or meta-analyses than Google Scholar. The results also support the previous study by Guistini (2013) that the use of

<Table 7> An Analysis of Top-10 Documents Using Three Quality Indicators for Clinical Studies

Search Engine	Percentage of Top-10 Document (%)		
	$S_{Date} \geq -0.04$	$S_{Study} = 0.5$	$S_{Journal} = 0.5$
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

Google Scholar alone is inadequate for the search of systematic reviews.



<Figure 6> Distributions of S_{SOE} Scores by Histograms

The distributions of S_{SOE} scores of the top-10 documents were visualized using histograms. As shown in <Figure 6>, the distribution of S_{SOE} scores for Clini Cluster 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 return clinical studies with higher quality of evidence, when compared to those retrieved by Google Scholar and CQA-1.0 (narrow).

Ⅶ. CONCLUSION

The study compared the known-item retrieval performance of a medical QA engine called 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, Clini Cluster outperformed other search engines with the known-items ranked higher in the results lists and 75% of the known-items can 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 are located on average between rank 2 and rank 3 in the result lists retrieved by CliniCluster.
3. An analysis of the strength of evidence provided by the top-10 documents, 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 and EpoCare systems that require users to transform their information needs into PICO query, CliniCluster is designed to accept both well- and ill-defined questions in natural

language. The user interface of CliniCluster provides information seeking features that aim to support users during the search of clinical information. These include a hierarchy of medical interventions to capture and narrow down users' search interest and a ranked list of answers tagged with the relevant PICO elements to assist users in recognizing their information needs. It is expected to be particularly useful for users who have a vague understanding of their search targets and who are unfamiliar with a problem domain.

The study was limited by a number of factors. First, the study focused on answering therapy questions using four search engines; the performance of the search engines in answering diagnosis, prognosis and epidemiology questions were not evaluated. Second, a rough interactive search of known-items was carried out using the hierarchy of medical interventions displayed by CliniCluster. To further evaluate the engine, a survey was conducted among health care providers to assess the usability of the hierarchy in supporting information seeking. A satisfactory result was obtained from the survey and will be published in the following paper. Third, the effectiveness of CliniCluster in answering a question was evaluated by identifying the paired document (i.e., the known-item) from a result list. Other documents, which are highly relevant to the question, are not included for the evaluation of an engine's information retrieval performance. Fourth, instead of identifying the most relevant sentences from the abstracts, the conclusions of abstracts are extracted and displayed as the answers to a question. Besides, similar to current MedQA systems, CliniCluster has limitation in terms of the ability to indicate whether the multiple answers displayed to users agree with each other on a particular query. Despite of these limitations, the results of the present study support the use of CliniCluster to answer therapy questions by ranking known-items, which have been judged by medical experts to be highly relevant, in the top positions of the search results. In order to be adopted in daily practice, the performance of the engine needs to be further optimized to process other types of clinical questions and to generate highly informative answers that can be utilized quickly for decision making and medical education by physicians.

<References>

[1] Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In Proceedings of the AMIA Symposium, American Medical Informatics Association, 17.

[2] Athenikos, S. J., and Han, H. (2010). Biomedical question answering: A survey. *Computer methods and programs in biomedicine, 99*(1), 1-24.

[3] Beel, J., and Gipp, B. (2009). 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, 439-446.

[4] Bergus, G. R., Randall, C. S., Sinift, S. D., and Rosenthal, D. M. (2000). Does the structure of clinical questions affect the outcome of curbside consultations with specialty colleagues? *Archives of Family Medicine, 9*(6), 541.

[5] Brunetti, L., and Hermes-Desantis, E. (2010). The Internet as a drug information resource. *US Pharmacist, 35*(1).

[6] Cao, Y., Liu, F., Simpson, P., Antieau, L., Bennett, A., Cimino, J. J., Ely, J., and Yu, H. (2011). AskHERMES: An online question answering system for complex clinical questions. *Journal of biomedical*

informatics, 44(2), 277-288.

[7] Cook, D. A., Sorensen, K. J., Hersh, W., Berger, R. A., and Wilkinson, J. M. (2013). Features of effective medical knowledge resources to support point of care learning: a focus group study. *PloS one, 8*(11), e80318.

[8] Davies, K. (2007). The information seeking behaviour of doctors: a review of the evidence. *Health Information & Libraries Journal, 24*(2), 78-94.

[9] Delbecque, T., Jacquemart, P., and Zweigenbaum, P. (2005). Indexing UMLS semantic types for medical question-answering. *Studies in Health Technology and Informatics, 116*, 805-810.

[10] Demner-Fushman, D., Few, B., Hauser, S. E., and Thoma, G. (2006a). Automatically identifying health outcome information in MEDLINE records. *Journal of the American Medical Informatics Association, 13*(1), 52-60.

[11] Demner-Fushman, D., Hauser, S. E., Humphrey, S. M., Ford, G. M., Jacobs, J. L., and Thoma, G. R. (2006b). *MEDLINE as a source of just-in-time answers to clinical questions. In AMIA Annual Symposium Proceedings.* American Medical Informatics Association, 190.

[12] Demner-Fushman, D., and Lin, J. (2007). Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics, 33*(1), 63-103.

[13] Demner-Fushman, D., Seckman, C., Fisher, C., Hauser, S. E., Clayton, J., and Thoma, G. R. (2008). *A prototype system to support evidence-based practice. In AMIA Annual Symposium Proceedings.* American Medical Informatics Association, 151.

[14] Duran-Nelson, A., Gladding, S., Beattie, J., and Nixon, L. J. (2013). Should we Google it? Resource use by internal medicine residents for point-of-care clinical decision making. *Academic Medicine, 88*(6), 788-794.

[15] Ebell, M. H. (2005). The vitamin E saga: lessons in patient-oriented evidence. *American Family Physician, 71*(11), 2052, 2054.

[16] Ebell, M. H., Barry, H. C., Slawson, D. C., and Shaughnessy, A. F. (1999). Finding POEMs in the medical literature. *The Journal of Family Practice, 48*(5), 350-355.

[17] Ely, J. W., Osheroff, J. A., Maviglia, S. M., and Rosenbaum, M. E. (2007). Patient-care questions that physicians are unable to answer. *Journal of the American Medical Informatics Association, 14*(4), 407-414.

[18] Essential Evidence Plus. (2014). "Browse Databases and Tools" [Online]. John Wiley and Sons. Available: http://www.essentialevidenceplus.com/content/poems.

[19] Evans, D. (2003). Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions. *Journal of Clinical Nursing. 12*(1), 77-84.

[20] Giustini, D. (2005). How Google is changing medicine. *Bmj, 331*(7531), 1487-1488.

[21] Giustini, D., and Barsky, E. (2005). 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, 26*(3), 85-89.

[22] Giustini, D., and Boulos, M. N. K. (2013). Google Scholar is not enough to be used alone for systematic reviews. *Online Journal of Public Health Informatics, 5*(2), 214.

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

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

[25] Huang, X., Lin, J., and Demner-Fushman, D. (2006). *Evaluation of PICO as a knowledge representation for clinical questions.* In AMIA Annual Symposium Proceedings, American Medical Informatics Association, 359.

[26] Hughes, B., Joshi, I., Lemonde, H., and Wareham, J. (2009). Junior physician's use of Web 2.0 for

information seeking and medical education: a qualitative study. *International Journal of Medical Informatics, 78*(10), 645-655.

[27] Krause, R., Moscati, R., Halpern, S., Schwartz, D. G., and Abbas, J. (2011). Can emergency medicine residents reliably use the internet to answer clinical questions? *Western Journal of Emergency Medicine, 12*(4), 442.

[28] Lappa, E. (2005). 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, 22*(2), 124-132.

[29] Mckibbon, K.A., Wilczynski, N.L., and Haynes, R.B. (2004). What do evidence-based secondary journals tell us about the publication of clinically important articles in primary healthcare journals? *BMC Medicine, 2*(1), 33.

[30] Methley, A. M., Campbell, S., Chew-Graham, C., Mcnally, R., and Cheraghi-Sohi, S. (2014). PICO, PICOS and SPIDER: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. *BMC health services Research, 14*(1), 579.

[31] Montori, V. M., Wilczynski, N. L., Morgan, D., and Haynes, R. B. (2005). Optimal search strategies for retrieving systematic reviews from Medline: analytical survey. *Bmj, 330*(7482), 68.

[32] Niu, Y., and Hirst, G. (2004). *Analysis of semantic classes in medical text for question answering.* In Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains, 54-61.

[33] Niu, Y., Hirst, G., Mcarthur, G., and Rodriguez-Gianolli, P. (2003). *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, 73-80.

[34] Niu, Y., Zhu, X., and Hirst, G. (2006). *Using outcome polarity in sentence extraction for medical question-answering.* In AMIA Annual Symposium Proceedings, American Medical Informatics Association, 599.

[35] Nixon, J., Wolpaw, T., Schwartz, A., Duffy, B., Menk, J., and Bordage, G. (2014). SNAPPS-Plus: An Educational Prescription for Students to Facilitate Formulating and Answering Clinical Questions. *Academic Medicine, 89*(8), 1174-1179.

[36] Sackett, D. L., Rosenberg, W., Gray, J., Haynes, R. B., and Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *Bmj, 312*(7023), 71-72.

[37] Schardt, C., Adams, M. B., Owens, T., Keitz, S., and Fontelo, P. (2007). Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Medical Informatics and Decision Making, 7*(1), 16.

[38] Schwartz, K., Northrup, J., Israel, N., Crowell, K., Lauder, N., and Neale, A. V. (2003). Use of on-line evidence-based resources at the point of care. *FAMILY MEDICINE-KANSAS CITY-, 35*(4), 251-256.

[39] Shariff, S. Z., Bejaimal, S. A., Sontrop, J. M., Iansavichus, A. V., Haynes, R. B., Weir, M. A., and Garg, A. X. (2013). Retrieving clinical evidence: a comparison of PubMed and Google Scholar for quick clinical searches. *Journal of Medical Internet Research, 15*(8),

[40] Shaughnessy, A. F., Slawson, D. C., and Bennett, J. H. (1994). Becoming an information master: a guidebook to the medical information jungle. *The Journal of Family Practice, 39*(5), 489-499.

[41] Smith, R. (1996). What clinical information do doctors need? *Bmj, 313*(7064), 1062-1068.

[42] Staunton, M. (2007). Evidence-based Radiology: Steps 1 and 2—Asking Answerable Questions and Searching for Evidence 1. *Radiology,* 242(1), 23-31.

[43] Thabane, L., Thomas, T., Ye, C., and Paul, J. (2009). Posing the research question: not so simple. *Canadian Journal of Anesthesia/Journal Canadien D'anesthésie, 56*(1), 71-79.

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

[45] Vong, W.-T., and Then, P. H. H. (2014). Visualization

of PICO Elements for Information Needs Clarification and Query Refinement. *Advances in Knowledge Discovery and Data Mining*. Springer.

[46] Weiming, W., Hu, D., Feng, M., and Wenyin, L. (2007). *Automatic clinical question answering based on UMLS relations*. In Third International Conference on Semantics, Knowledge and Grid, Citeseer, 495-498.

[47] Wilczynski, N. L., and Haynes, R. B. (2004). Developing optimal search strategies for detecting clinically sound prognostic studies in MEDLINE: an analytic survey. *BMC Medicine, 2*(1), 23.

[48] Wilczynski, N. L., Haynes, R. B., and Team, H. (2003). *Developing optimal search strategies for detecting clinically sound causation studies in MEDLINE*. In AMIA annual symposium proceedings, American Medical Informatics Association, 719.

[49] Wong, S. S.-L., Wilczynski, N. L., Haynes, R. B., Ramkissoonsingh, R., and Team, H. (2003). *Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE*. In AMIA Annual Symposium Proceedings, American Medical Informatics Association, 728.

[50] Yu, H., and Cao, Y.-G. (2008). *Automatically extracting information needs from ad hoc clinical questions*. In AMIA annual symposium proceedings, American Medical Informatics Association, 96.

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

[52] Zwolsman, S., Te Pas, E., Hooft, L., Wieringa-De Waard, M., and Van Dijk, N. (2012). Barriers to GPs' use of evidence-based medicine: a systematic review. *British Journal of General Practice, 62*(600), e511-e521.

<Appendix A> Generation of PICO Elements

Medical concepts with semantic types listed in the <Table A> were recognized as PICO elements, whereas those with other semantic types were excluded.

<Table A> Identification of PICO Elements by Semantic Types

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.

<Appendix B> Generation of Question-Document Pair

An example of how a question-document pair is generated from a POEM is shown in <Figure A>. The POEM is retrieved from (http://www.essentialevidence plus.com/content/poems). The paired document serves as the known-item.

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 A> An example of POEM

<Appendix C> Five Structural Patterns of Therapy Questions

<Table B> 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

<Table B> Five Structural Patterns of Therapy Questions

Pattern	Examples
[P][I][O?]	Is enoxaparin [I] useful for moderate renal impairment [P]? Does niacin plus laropiprant [I] useful for patients with vascular disease [P]?
[P][I?]	What is the best treatment for acute otorrhea [P]? What is the best way to treat menorrhagia [P]?
[I][O?]	Is zanamivir [I] effective in relieving flu symptoms [O]? Is gabapentin [I] useful in decreasing cough [O]?
[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]?
[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]?

◆ About the Authors ◆

Wan-Tze Vong

Wan-Tze Vong is a PhD candidate at Faculty of Engineering, Computing and Science, Swinburne University of Technology (Sarawak Campus). She holds a MRes degree in Medical and Molecular Biosciences from University of Newcastle Upon Tyne, United Kingdom. Her research interests include clinical question answering, knowledge representation, document visualization, natural language processing, medical informatics and biostatistics.

Patrick Hang Hui Then

Patrick Hang Hui Then is an associate professor at Faculty of Engineering, Computing and Science, Swinburne University of Technology (Sarawak Campus). He is an active researcher with strong collaboration with industries. He has been actively publishing papers for journal, conference proceedings and book chapters. Dr. Then research interests include knowledge discovery, data mining, information security, privacy preserving, health economics, biostatistics and microbiology.

Submitted: April 22, 2015; 1st Revision: June 17, 2015; 2nd Revision: August 12, 2015; Accepted: August 31, 2015