Official URL:

https://doi.org/10.1016/j.ipm.2016.11.004

To cite this version:

Tamine-Lechani, Lynda and Chouquet, Cécile *On the impact of domain expertise on query formulation, relevance assessment and retrieval performance in clinical settings.* (2017) Information Processing and Management, 53 (2). 332-350. ISSN 0306-4573.

On the impact of domain expertise on query formulation, relevance assessment and retrieval performance in clinical settings

Lynda Tamine*, Cecile Chouquet

Paul Sabatier University, 118 route de Narbonne, Toulouse 31062 Cedex 9, France

ABSTRACT

The large volumes of medical information available on the web may provide answers for a wide range of users attempting to solve health-related problems. While experts generally utilize reliable resources for diagnosis search and professional development, novices utilize different (social) web resources to obtain information that helps them manage their health or the health of people who they care for. A diverse number of related search topics address clinical diagnosis, advice searching, information sharing, connecting with experts, etc. This paper focuses on the extent to which expertise can impact clinical query formulation, document relevance assessment and retrieval performance in the context of tailoring retrieval models and systems to experts vs. non-experts. The results show that medical domain expertise 1) plays an important role in the lexical representations of information needs; 2) significantly influences the perception of relevance even among users with similar levels of expertise and 3) reinforces the idea that a single ground truth does not exist, thereby leading to the variability of system rankings with respect to the level of user's expertise. The findings of this study presents opportunities for the design of personalized health-related IR systems, but also for providing insights about the evaluation of such systems.

Keywords:
Medical information retrieval
Domain expertise
Query formulations
Relevance assessment
Retrieval performance

1. Introduction

Several studies (Fox & Duggan, 2013; Fox, 2011) have clearly shown that people, both experts (e.g., physicians and nurses) and novices (e.g., patients and their family), have strong desires for medical information. Regardless of the domain expertise of users seeking information, medical and health-search have been acknowledged as a complex search tasks leading to search failures or biases (Ely et al., 2005, 2007; Roberts, Simpson, Demner-Fushman, Voorhees, & Hersh, 2015; White & Horvitz, 2015). Even if it appears that the effectiveness of specialized search engines within the medical domain is not significantly higher than the effectiveness of general web search engines (Bin & Lun, 2001), several previous studies have revealed that significant differences between them in search intents may be linked to the information resources being used (Choudhury, Morris, & White, 2014; Natarajan, Stein, Jain, & Elhadad, 2010; Zhang & Fu, 2011):

- *General web resources*: This category of resources includes resources indexed by general web search tools and social platforms not particularly devoted to or certified for health concerns, thereby leading to general web searching (which is

* Corresponding author.
E-mail addresses: tamine@irit.fr, lechani@irit.fr (L. Tamine), cecile.chouquet@math.univ-toulouse.fr (C. Chouquet).

different from a vertical search). Searching the web for health-related information has been acknowledged as a frequent activity of a wide variety of users (Fox & Duggan, 2013; Spink et al., 2004; Zhang & Fu, 2011). The web is used for addressing a wide range of search topics, such as those concerning (Choudhury et al., 2014; Eysenbach, Powell, & Englesakis, 2004; Spink et al., 2004; White & Horvitz, 2009; Zhang & Fu, 2011): 1) general health, drug and dosing, and disease management (searching for rare diseases or updates on common diseases); 2) (differential) diagnosis or referral guidelines; 3) professional development; 4) personal and opinion-oriented goals (personalized healthy lifestyle information such as diet, nutrition, and sexual health information); 5) advice (e.g., advice after being dissatisfied with professional care); 6) information sharing (e.g., with doctors and/other patients) ; 7) people with similar conditions on social platforms; and 8) connecting with experts.

- *Clinical information resources*: This category of resources is used within a domain-specific or vertical search, including 1) electronic health records (EHRs) that are used by medical professionals and 2) medical scientific reviews or content from certified health and medical sites that are both used by experts (e.g., clinicians) and non-experts (novices) for different purposes. Expert clinical information searches are generally performed by clinicians under the Evidence-Based Medicine (EBM) approach (Sackett, 1997) as the basis for clinical decisions that better suit the patient under consideration. In contrast, non-expert clinical information searches are completed to help patients and their representatives to better understand their own health conditions or conditions of people they care for. Searching for clinical information is also a common pursuit. A previous study showed, for example, that 1.8 billion clinical searches were conducted on PubMed in 2011 (NLM, 2012); another previous study showed that one-third of PubMed users is not medical experts (Lacroix & Mehnert, 2002).

  Early studies (Ely et al., 2000; Pratt & Wasserman, 2000) proposed a general classification for search topics hidden behind clinical queries that are clearly less diversified than are health-related searches performed on general web resources. In Pratt and Wasserman (2000), the authors classified clinical queries that were addressed to MEDLINE into 10 category topics, including prevention, risk factors, diagnosis, symptoms, treatments and side effects.

In this paper, clinical information search is specifically investigated, the performance of which remains questionable and subject to numerous issues (Cohen, Stavri, & Hersh, 2004; Francke, Smit, & de Veer, 2008; Natarajan et al., 2010; Suominen et al., 2013; White & Horvitz, 2015). These issues mainly arise from the following: 1) the complexity of expressing precise, context-specific clinical queries that better facilitate the identification of the relevant evidence and 2) the lack of a higher level expertise that can be used to perform evidence appraisal. Thus, we argue that an ideal clinical search engine should exploit information nuggets from both the query and the domain expertise level of the user to accurately identify clinically relevant information. Achieving this requires a deep understanding of the key differences that exist between expert-based and non-expert-based clinical information searches. To the best of our knowledge, how expert-clinical queries differ from non-expert queries is not well established in the literature; furthermore, the differences in the relevance assessment provided by either experts or non-experts and their impact on system ranking stability have not thoroughly investigated. With this in mind, we attempt to investigate the differences, commonalities, and relationships between expert-based and novice-based clinical searches. We focus on: 1) the query formulation in terms of length, domain-specificity and difficulty attributes, acknowledged as being important factors that could contribute to search success/failure (Ely et al., 2005, 2007; Tamine, Chouquet, & Palmer, 2015); 2) the relevance assessment in terms of difficulty and related reasons, relevance agreement between assessors, time spent to assess relevance and 3) the relationship between user's expertise level and retrieval effectiveness with respect to his relevance assessment. We conducted our study by assigning search tasks to experts and novices via two distinct crowdsourcing platforms allowing to recruit the two categories of clinical information seekers (experts/novices). To design reliable simulated clinical search tasks, we used the medical cases provided within major medical IR evaluation tracks namely the TREC[1] Filtering (Robertson & Hull, 2000) and the CLEF[2] e-Health (Suominen et al., 2013) with related search contexts.

The remainder of this paper is structured as follows. In Section 2 we describe research work related to the effects of domain expertise on search and relevance assessment and those related to crowdsourced user studies. To put this work in context, the findings are reported for both cross-domain expertise and specific medical domain expertise. Section 3 announces the research questions and then describes the studies that we perform in order to identify the commonalities and the differences between expert-based search and novice-based search within the medical domain, including query formulation, relevance assessment and retrieval performance. In Section 5 we report the findings of our studies based on quantitative and qualitative analysis. Section 6 discusses the results and highlights the study implications. Section 7 concludes this article.

2. Related work

2.1. On the influence of domain expertise on information search: query formulation, search behavior and search difficulty

Based on intensive research work that has been performed in information science, researchers agree that information seeking and retrieval are perceived as cognitive activities constrained by several contextual factors used for reducing the

[1] Text REtrieval Conference.

[2] Conference and Labs of the Evaluation Forum.

complexity of the retrieval process (Ingwersen & Belkin, 2004). One of the major factors identified is knowledge, which can be divided into search knowledge and domain knowledge. While search knowledge concerns the knowledge of search processes, domain knowledge, which is also referred to as factual knowledge, concerns knowledge about the search topic. A large amount of research work has examined the effects of differences in domain expertise on both search processes (Bhavnani, 2001; White, Dumais, & Teevan, 2009; Wildemuth, 2004) and search difficulty (Liu, Liu, Cole, Belkin, & Zhang, 2012; Liu, Kim, & Creel, 2014). In Bhavnani (2001), the authors examined the impact of the cognitive components of domain-specific search knowledge on search behavior. In their study, five healthcare search experts and four online shopping experts were recruited. Their search behaviors were examined while searching for both health-related and shopping-related concerns. The main study finding was that searches within and outside domain expertise are significantly different. The key facet of the difference concerns the sequencing behavior adopted by experts vs. novices to resolve the information need. Using problem behavior graphs (PBG), the authors showed that while experts launched the search from key resources, novices started from general search tools. Moreover, expert-based searches are more focused and successful following, for instance, a review-comparison-discount search pattern for a camera search task. However, novice-based searches generally lead to unsuccessful searches ending with irrelevant information entries. Similarly, Wildemuth (2004) examined the search behavior, called a search tactic, of 77 students involved in search sessions to answer six clinical microbiology problems; the search sessions occurred at different timestamps before and in several sessions after the end of the course. These timestamps represented different levels of domain knowledge evolving over time. The analysis of the search tactics used at these different stages highlighted significant changes from adding/deleting concepts at the beginning of the period to adding multiple concepts or adding a small number of useful concepts at the end. White et al. (2009) investigated the effect of expertise on web search behavior. The authors performed a large-scale analysis that included 90 million search sessions related to real-life searches performed in four domains (medicine, finance, law and computer science). The results show that expert-based searches are significantly different than non-expert-based searches in terms of several features such as query formulation (length, used vocabulary), search behavior (visited sites, page dwell time, ratio of querying browsing), and search success.

More specifically, to gain a broad understanding of the research performed on the influence of medical expertise on information retrieval (IR) and seeking, we examined two lines of studies that have explicitly compared searches performed by experts and non-experts (Palotti, Hanbury, & Henning, 2014; White, Dumais, & Teevan, 2008) as well as other studies that focused on expert-based searches (Hersh et al., 2002, 2000; Lykke, Price, & Delcambre, 2012; Soldaini, Yates, Yom-Tov, Frieder, & Goharian, 2016; Tamine et al., 2015; Yang, Mei, K.Zheng, & Hanauer, 2011).

In the first line of work, White et al. (2008) investigated the effects of medical expertise on web search interaction using a range of query-based and session-based features (the query length, the percentage of technical terms, the number of queries per session, etc.), behavioral features (browsing, visiting, etc.) and source selection. Authors found that experts issued longer and more technical queries than did novices. Moreover, experts issued more queries per session and spent more time searching. Palotti et al. (2014) designed a classifier that was able to distinguish between medical professionals and novices. Groups of features such as semantic features and common term usage features (linking technical expressions to those used by novices) were first identified. The authors have shown that the top relevant features customarily related to query formulation. Recently, Soldaini et al. (2016) proposed a query clarification strategy based on medical resources mappings in order to bridge the gap between novice and expert vocabularies when formulating medical queries to web search engines.

In the second line of work, Lykke et al. (2012) examined doctor query behavior within a workplace search. The authors collected data about the search behavior of 30 family practice physicians through interviews, questionnaires and search logs. They found that doctors typically expressed well-structured queries; the most important reasons for search failures were related to technical term mismatches between the queries and the documents rather than to the query length. Yang et al. (2011) analyzed a collection of query logs from the EMERSE search engine, which facilitates access to electronic health records (EHRs). The collection includes 202,905 queries issued by 533 medical professionals recorded over 4 years. The main study finding was that queries were underspecified, included acronyms (18.9% of queries contain at least one acronym) and had little coverage within medical resources, including ontologies and dictionaries (coverage rate: 68%). In Tamine et al. (2015), the authors examined the differences in expert medical query formulations across various tasks. Exploratory analyses were performed using 11 TREC and CLEF medical test collections, including 374 queries related to different medical tasks such as gene retrieval and clinical IR. The authors showed that language specificity levels and search difficulty vary significantly across tasks; the best predictive factors are linked to query length and query clarity.

2.2. On document relevance assessment in medical IR

It is well known in IR evaluation area that relevance assessment provided by human judges or annotators is crucial Voorhees (2000). According to the Cranfield evaluation paradigm, human relevance assessments allow building the ground truth as the starting point for system performance measurement and beyond, allow making comparisons between IR systems. Two core questions are related to relevance assessment: 1) time and cognitive costs and 2) agreement between assessors. The first question which impacts the experiments cost has been addressed in the community through evaluation campaign initiatives such as TREC and CLEF or via crowdsourcing evaluation methods Lease and Yelmaz (2011). A recent work in the specific medical IR domain Koopman and Zuccon (2014) outlined that providing relevance assessments is a time-consuming and a cognitively expensive process; more specifically, using relevance assessments provided by four experts asked to judge documents taken from the TREC MedTrack Voorhees and Hersh (2012), the authors found that time spent to assess relevance is not obviously related to document length and that cognitive load is query-dependent. The second question related to assessor agreement which impacts in contrast system rankings and has been addressed in early IR works by Lesk and Salton (1968). Their study clearly showed that only a low level of agreement is achieved between assessors (0.31 and 0.33) but also showed that this difference does not significantly impact systems rankings considering ground truth built with one or another assessment. Other studies revealed similar findings Voorhees (2000) but unlikely, more recent studies Bailey et al. (2008) showed that variation in tasks and domain expertise significantly impact search engine rankings.

Relevance assessment agreement has also been studied in the medical domain based on expert judges. Previous works mostly related to TREC evaluation campaigns revealed that agreement level in the relevance of family physicians within ad hoc searches achieves relative low levels computed using the j-statistic measure. For instance in the TREC Genomics track, the overlap of relevance judgments was only able to achieve 0.51 in 2004 Hersh et al. (2004) and 0.59 in 2005 Hersh, Cohen, Yang, Bhupatiraju, and Roberts (2005), while the agreement does not exceed 0.44 in the TREC Medical track Roberts et al. (2015). In Koopman and Zuccon (2014), the authors stated that disagreement between the assessors particularly occurs for "interpretation queries" which require considerable consideration regarding different possible interpretations of document or query contents.

2.3. User studies using crowdsourcing platforms

Crowdsourcing has become a powerful tool for obtaining labels for IR system development and evaluation (Lease & Yelmaz, 2011). Crowd source platforms such as CrowdFlower and Mechanical Turck have been used in previous work to perform a wide range of cheap and reliable controlled studies including those related to human behavior Kittur, Chi, and Suh (2008), question generation Jeong, Morris, Teevan, and Liebling (2013) and relevance evaluation Alonso, Rose, and Stewart (2008). In the medical domain, crowdsourcing platforms have also been employed for different purposes including collecting belief ratings before, during and after engaging with a search engine White and Hassan (2014); White and Horvitz (2015) and for answering questions Soldaini et al. (2016).

3. Study

3.1. Research questions

As outlined in the literature review, only a few studies have examined the differences between expert-based and non-expert-based information searches in the medical domain (Palotti et al., 2014; White et al., 2009, 2008). Furthermore, previous research did not focus on understanding the differences within clinical information searches specifically. We identified the following gap in previous research:

- There is a lack of studies that thoroughly identify the query features that better distinguish between expert-based and non-expert based clinical information needs. Although multiple features were studied for classification purposes by Palotti et al. (2014), the study is preliminary and needs to be completed to draw firm conclusions. In contrast to the work presented by White et al. (2009, 2008) who studied the impact of domain expertise on user behavior, this research investigates whether expert vs. non-expert searches performed specifically on vertical repositories vary according to query formulation, relevance assessment and query performance.
- There is a crucial need of in-depth empirical research highlighting the differences in both the levels of agreement and the causes of difficulty in relevance assessment performed by experts and novices within clinical searches. The research presented in this paper significantly extends prior work on the cognitive and time costs of expert-based relevance assessment in the medical domain (Koopman & Zuccon, 2014) by examining the relationship between relevance assessment task, its difficulty and the time spent to achieve it and also comparing the levels of relevance agreement and the qualitative reasons for its difficulty between experts and novices.
- To the best of our knowledge, no prior studies have investigated the impact of the variation in medical domain expertise level on retrieval performance. We attempt to fill this gap by considering both quantitative and qualitative measurements of retrieval effectiveness using ground truth estimated from multiple settings based on levels of relevance agreement and levels of domain expertise.

This motivates the formulation of three research questions:

- **RQ1:** Are there significant differences in the clinical query formulations considering the domain expertise of users?
- **RQ2:** What is the relationship between domain expertise and the relevance assessment task in terms of agreement and difficulty?
- **RQ3:** Are the levels of performance of traditional IR systems based on query-document term matching significantly different with respect to domain expertise of users?

To provide answers to those questions, we carried out a study using crowdsourcing platforms and using medical cases issued from major IR evaluation campaigns namely TREC and CLEF. Those answers would help guide the design of systems that provide users with the appropriate assistance considering both the level of expertise and search interests.

Table 1
Summary statistics of the datasets used in the study.

Feature	TREC Filtering	CLEF e-Health
#Medical cases	63	55
#Documents	293,856	1,000,000
Average length of documents (words)	100	312
Average relevant documents per topic	50	10

Table 2
Example of TREC-Ohsumed topic.

< *top* >
< *num* > Number: OHSU1
< *title* > 60 year old menopausal woman without hormone replacement therapy
< *desc* > Description: Are there adverse effects on lipds when progesterone is given with estrogen replacement therapy?
< */top* >

Table 3
Example of a CLEF eHealth query.

< *query* >
< *id* > *qtest*2014 < */id* >
< *title* > Cornoray artery disease < */title* >
< *desc* > What does coronary artery disease mean < */desc* >
< *narr* > The documents should contain basic information about coronary artery disease and its care< */narr* >
< *profile* > This positive 83 year old woman has had problems with her heart with increased shortness of breath for a while. She has now received a diagnosis for these problems having visited a doctor. She and her daughter are seeking information from the internet related to the condition she has been diagnosed with. They have no knowledge about the disease< */profile* >
< */query* >

3.2. Medical case descriptions

All of the participants, experts or novices, aimed to solve the same health-related search tasks for which information needs were extracted from two clinical information datasets issued from major IR evaluation campaigns, namely TREC Filtering (Robertson & Hull, 2000) and CLEF E-Health (Suominen et al., 2013). The datasets include a total of 113 medical cases, called also topics, that have been employed for query generation within search tasks submitted to the crowd workers. Various general properties and statistics of the data used in this study are described in Table 1. Below, we describe the datasets.

- *TREC Filtering*. This track (Robertson & Hull, 2000) attempts to measure the ability of an IR system to select relevant documents that fit the needs of a persistent user represented by profiles. Note that the medical dataset provided within this track for an ad hoc retrieval task was used in this study rather than a filtering task. Specifically, the OHSUMED test collection was used, consisting of a set of 348,566 references from MEDLINE, the online medical database of a five-year journal (1987–1991) provided by Hersh, Buckley, Leone, and Hickam (1994). This collection is known as a large-scale, standard collection for ad hoc medical IR (Stokes, Cavedon, & Zobel, 2009). We used one of the subsets of the TREC-9 filtering track medical cases developed by Hersh and Hickam (1994) for their medical information retrieval experiments. The ad hoc task simulated the use case by performing an assessment of the use of MEDLINE by physicians in a clinical setting. Eleven medical reference librarians and eleven physicians experienced with MEDLINE were recruited. Moreover, each physician had to have an active clinical practice in an ambulatory setting. The topics included the patient information provided in a full-text form (TI) and the request description (AB), excluding Human-assigned MeSH terms (MH). An example of an OHSUMED filtering query is presented in Table 2.
- *SHARE eHealth CLEF track*. The overall goal of the ShARe eHealth track is to evaluate systems that assist novices in understanding their health-related information Suominen et al. (2013). Here, the dataset provided within *Task 3* is utilized, which is an ad hoc health-related IR task. The dataset provided to participants includes the following: 1) either a document collection from medical certified websites[3] or from commonly used websites such as Diagnosia[4] or Drugbank[5], which address a wide range of search topics, and 2) a set of general public queries that users may realistically issue based on the content of their discharge summaries. Each topic description contains additional information about the patient discharge summary that was assumed to have triggered the corresponding query. An example of a query is presented in Table 3.

[3] Certification according to the HONcode principle http://www.hon.ch/HONcode/Patients-Conduct.html.
[4] http://www.diagnosia.com/.
[5] http://www.drugbank.ca/.

3.3. Participants

Our study relies on the CrowdFlower[6] and the Upwork[7] crowdsourcing platforms. The former is used for recruiting novices while the latter is used for recruiting medical experts. Participants were mostly from United States and were required to be fluent in English. All the participants, either novices or experts, were asked to provide demographic data such age, gender and level of education. Novice workers were asked to indicate themselves (self-rating) on a 3-point scale (*basic, low* and *high*), their health literacy level.

- *Novice participants*. To ensure reliable task outcomes, we submitted the tasks for experienced crowd workers with a high level of performance (Level=3). The latter is assessed by the platform using an average measure of the correctness of their answers regarding the test questions over all the tasks they achieved. Furthermore, for additional quality control, we include for each task predefined pairs of question and answer as the gold standard. Only crowd workers who correctly answered the question tests are finally recruited. A total of 119 novices participated in the study, 41 female (35%) and 78 male (65%) and the average age was 35 years old ($SD = 11.3$), ranging from 18 years old and 70 years old. The most frequent study level is bachelor's for 65 users (55%), Master for 23 users (19%), Doctorate for only 2 users (2%) and other for 29 users (24%) . We only retained judges for whom the health literacy was basic otherwise medium or high but with a level of education is the bachelor at the most. Novice participants indicated that their health literacy was basic for 83 users (70%), low for 20 users (17%) and high for 16 users (13%). Participants were compensated financially for each task: 20 cent for *Task 1* and 25 cent for *Task 2*.
- *Expert participants*. A total of 5 experts participated in the study: 2 male (40%) and 3 female (60%), and their average age was 35 years old ($SD = 7.3$), ranging from 28 years old and 42 years old. The level of education was high, all of them are medical doctors and among them 2 were medical researchers with a long experience in medical writing. We assume that their health literacy is high. They were compensated financially for each task: 35$ for *Task 1* and 44$ for *Task 2*.

Below, we list the tasks performed by the crowd workers and guided by the research questions RQ1-RQ3.

3.4. Tasks

We created two (2) Human Intelligence Tasks (HIT) on each of the crowdsourcing platforms in which crowd workers formulated queries and assessed the relevance of documents returned by an IR system to those queries. Since the tasks are based on subjective criteria, it is likely that different workers (either experts or novices) have different levels of agreements even in the same category while performing them. Therefore, we assigned each task to at least two novices and two medical experts. The tasks are detailed below.

3.4.1. Task 1: query formulation

To study the impact of domain expertise on query formulation (RQ1), we designed a query formulation task which was expression oriented and presented crowd workers with a simulated search task. They were asked to build the appropriate query that allows achieving the search task. According to the study objectives, the search tasks were mapped from the formulation of the TREC and CLEF medical cases described using a pair of facets provided to the crowd workers: 1) *search context* which corresponds to the medical case that triggers the information need and 2) *the information need* which gives clues about the desired content of relevant documents. For the TREC Filtering topics, we used the *Title* as the context and a reformulation of the *Description* as the information need. For the CLEF e-Health, we used the *Profile* as the context and the *Narrative* as the information need. A total of 113 topics were submitted for both novices and experts. The data obtained from each task achieved by novices has been checked for coherence and accuracy by two human judges and the tasks were resubmitted until achieving reliable outputs. Each topic was submitted to 3 experts and 3 novices for self-query generation; therefore, 6 queries were formulated for each topic. A total of 678 queries were analyzed including 339 queries formulated by novices and 339 queries formulated by experts. For each novice worker, an elementary task consists in formulating one query test and one another corresponding to a real TREC or CLEF topic. Each novice formulated 2.8 queries on average (from 1 to 5 queries per user). For the experts, an elementary task consists in formulating $35 - 38$ queries. Each expert performed 1 or 2 tasks leading to the formulation of a range between 35 and 113 queries (only one task included 38 topics).

3.4.2. Task 2: relevance assessment

With respect to research question RQ2, our first goal behind this task is to test the differences between novices' and experts' relevance assessments. To achieve this goal, we employ the formulated queries obtained from *Task 1* (a total of 6 queries per topic including 3 queries from each category of users) from which we build a single system ranking of candidate relevant documents. To ensure a right balance between system rankings issued from the different formulated queries, we apply an interleaving algorithm (Radlinski, Kurup, & Joachims, 2008) with respect to the following three steps:

[6] http://www.crowdflower.com/.
[7] http://www.upwork.com/.

1. For each topic, we built 6 rankings of documents related to each formulated query obtained through *Task 1*: 3 queries from novices and 3 queries from medical experts. Each ranking results from the query evaluation process using 1) the Terrier search engine[8], 2) the appropriate collection (CLEF E-Health, TREC Filtering document collections) and 3) the Language Model (LM) IR model (Song & Croft, 1999) with the Dirichlet smoothing method with $\mu = 1000$.
2. For each category of participants (experts vs. novices) and each topic, we built one interleaved ranking by processing pairs of rankings using the Team Draft Interleaving (TDI) (Radlinski et al., 2008). Thus we obtain a pair of rankings for each topic, namely one expert-based ranking and one novice-based ranking.
3. We further interleave the expert-based ranking and the novice-based ranking. Thus we obtain a single ranking for each topic. Accordingly, each participant in *Task 2* gets about the same number of documents from participants belonging to his own category or other participants from another category (experts vs. novices).

Crowd workers were then provided with a medical case (among the 113 TREC and CLEF medical cases) and a list of top 10 candidate relevant documents from the interleaved ranking built as detailed above. For each query-list of top 10 results, we obtained relevance labels from 2 experts and 2 novices. As previously done in TREC evaluation campaigns, the assessors (here the crowd workers) were instructed to rate the topical relevance of documents in a 3-point scale (*Relevant, Partially Relevant, Not Relevant*). Task instructions stated that (Roberts et al., 2015): "*a document is relevant or partially relevant to a given topic within its context if they find it useful in addressing the generic information need posed by the given topic. The document has to: 1) provide information of importance to the information need, 2) provide information that is topically relevant to the information need. A document that suggests a particular diagnosis, test, or treatment that sound reasonable given the information available in the topic/context task should be judged Relevant. A Partially Relevant document should contain meaningful information to find suggested diagnosis, test, treatment. In the cases where the document suggests information that are not appropriate to the given information need and task context or does not even describe the medical condition at all, should be assessed as Not Relevant*".

The relevance ratings obtained through this task allow us to evaluate the relevance agreement between assessors (RQ2) but also to achieve the second goal behind this task which consists in measuring the effect of domain expertise on the stability of retrieval system performance considering the domain expertise of users (RQ3). To do so, we made use of various sets of gold standard and averaged the performance measures over different retrieval scenarios according to different assumptions of document relevance.

Moreover, we asked all the crowd workers to estimate the time they spent to achieve the task and rate the difficulty of the task on a 3-point scale (*Easy, Moderate* and *Difficult*). They also were asked to provide free-text qualitative comments about the reasons for the difficulty vs. easiness of the task.

For the novices, an elementary task financially compensated with 25 cents consists in assessing the relevance of 10 documents related to one query test and 10 documents related to one real TREC or CLEF query in addition to answering the questions related to time and task difficulty. For the experts, an elementary task consists in assessing the relevance of 10 documents related to 5 topics in addition to answering the questions related to time and task difficulty, with a financial compensation of 44$.

3.5. Query features

We detail below the query features used in our study and motivate their choice.

- ***Query length.*** The query length is considered to be a relevant attribute to characterize medical and health-related information needs, as shown in previous work (Lykke et al., 2012). Furthermore, because both experts and non-experts might use medical terminologies, two facets of query components are retained: 1) $LgW(Q)$, which refers to the query length based on the number of stems or significant words (not including empty words), and 2) $LgC(Q)$, which refers to the query length based on the number of concepts in the query identified through terms that reference related preferred entries issued from a reference medical terminology.
  In this study, the following medical resources and methods of concept-based representations were used:
  - *The MeSH terminology*: This terminological resource is chosen because previous work clearly shows that it is the most-used general resource in the biomedical domain (Stokes et al., 2009). This choice allows the results issued from this study to be comparable to other results issued from prior studies in the literature review.
  - *The concept extraction method* (Dinh & Tamine, 2011; Dinh, Tamine, & Boubekeur, 2013): This method relies on an IR-based approach for concept recognition built upon Metamap[9]. The key component of this method consists of representing the text (here a query) semantic kernel as the top relevant concepts, which are extracted by measuring the concept relevance for the text.
- ***Query specificity.*** Specificity is usually considered as a criterion for identifying index terms or descriptors (Jones, 1972). In the medical domain, specificity is used for identifying hierarchical semantic levels of queries (Ely et al., 2000). Considering the problem addressed in this article, we expect, as shown in previous studies (White et al., 2008), that experts are

[8] http://www.terrier.org.
[9] http://metamap.nlm.nih.gov.

more willing to express focused queries with a more specific language than novices would do. Two types of specificity, which have previously been shown to be uncorrelated regardless of the medical task under consideration (Tamine et al., 2015) are used:

1. *Posting specificity PSpe*(*Q*): The posting specificity represents the uniqueness of the query words in the index collection; the basic assumption behind the posting specificity is that the fewer documents involved with the query words, the more specific the query topics are (Kim, 2006).

$$PSpe(Q) = \frac{1}{LgW(Q)} \times \sum_{w \in words(Q)} -log\frac{N_w}{N} \tag{1}$$

   where *words*(*Q*) is the set of words belonging to the query *Q*, N_w is the number of documents that contain the word *w*, and *N* is the document collection size.

2. *Hierarchical specificity HSpe*(*Q*): The hierarchical specificity is based on the deepness of meaning of the query words as defined in a reference terminology through the "is-a" taxonomic relation (Kim, 2006). The basic underlying assumption is that the more specific concepts are involved with the query words, the more specific the query topics are. The hierarchical specificity of a query is computed as follows:

$$HSpe(Q) = \frac{1}{LgC(Q)} \sum_{c \in Concepts(Q)} -\log \frac{level(c)}{2 * Maxlevel(MeSH)} \tag{2}$$

   where *Concepts*(*Q*) is the set of concepts belonging to the query *Q*, *level*(*c*) is the MeSH level of concept *c*, and *Maxlevel*(*MeSH*) is the maximum level of MeSH.

- ***Query difficulty.*** Intuitively speaking, a difficult query (an easy query) leads to a low (high) retrieval performance. Our motivation behind the study of this feature is to explore in what extent the vocabulary of the query matches the vocabulary of the document. We use difficulty pre-retrieval predictors based on similarity score that has been shown to be effective in both general web document (Zhao, F. Scholer, & Tsegay, 2008) and medical document collections (Limsopatham, Macdonald, & Ounis, 2013). The main underlying idea is that a query is more likely to be easy when it is similar to more documents in the collection. The Normalized similarity score *NSC*(*Q*) was used:

$$NSC(Q) = \frac{SCQ}{LgW(Q)}, SCQ = \sum_{w \in words(Q)} \left(1 + ln(N_c(w)) \times ln\left(1 + \frac{N}{N_w}\right)\right) \tag{3}$$

  where $N_c(w)$ is the frequency of word *w* in the collection.

4. Results

The central goal of this study was to investigate the similarities and commonalities between the formulations, relevance assessments and performance of expert searches and non-expert ones. The statistical analysis were performed using the SAS (http://www.sas.com/) software, version 9.3. Document indexing and retrieval were performed using Terrier framework (http://www.terrier.org) version 4.0. In this section, the results are grouped by research question and the main findings that arise from the results are highlighted.

4.1. Query formulation (RQ1)

This study began by analyzing how users characterized by different levels of expertise formulated their information needs. The primary goal of this analysis was to investigate the impact of domain knowledge on query formulation. Several comparative statistical analysis were completed between queries expressed by experts and queries submitted by novices. As stated in the description of the query formulation task (See Section 3.4.1), a total of 678 queries were generated from 113 topics including 339 queries formulated by novices and 339 other queries formulated by experts.

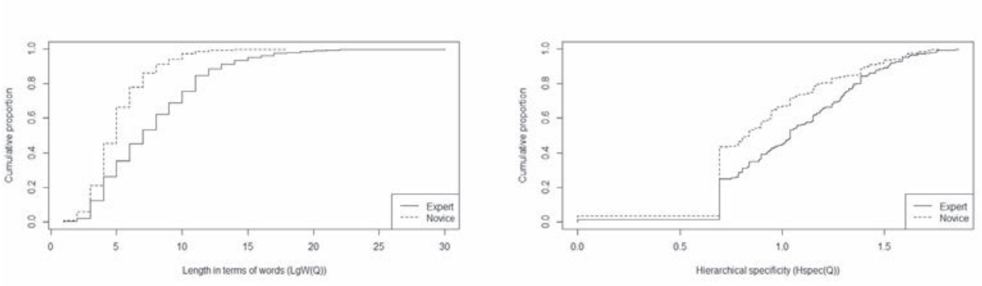
Table 4 provides a summary of query feature counts of the groups based on the standard statistical indicators of the mean value of the feature ($\underline{M}$), the standard deviation ($\underline{SD}$) and the median value within the different query groups described above. Linear mixed-effects models for repeated measures (Davis, 2002) were conducted to test the differences between expert and novice groups. These models take into account repeated measures: each user formulated several queries (between 1 and 113) and 6 queries are submitted for each topic. The significance of the differences was estimated using the *p-value*. Table 4 shows the obtained *p-value* ranges: not significant *ns*, moderately significant * ($0.01 < p < 0.05$), significant ** ($0.01 < p < 0.001$) and highly significant *** ($p < 0.001$). Moreover, averages of each feature were calculated by topic in each user group.

Because the primary objective of this study was to analyze the expertise effect on query formulation, the features were compared between queries issued by experts and those issued by novices.

Table 4 indicates that queries issued by experts are significantly different than those issued by novices in terms of almost all features. The query length in terms of words (*LgW*(*Q*)) indicates that experts generally issued longer queries than did novices ($M = 7.8$, vs. $M = 5.2$, *p-value* < 0.001), as previously shown both in web search either in the medical domain and out of the medical domain (Hembrooke, Granka, Gay, & Liddy, 2005; White et al., 2008). For 89% of topics, experts have

Table 4
Description of features and significance of feature-based differences between query groups.

Groups		*Experts*			*Novices*			
# Queries		339			339			
		M	*SD*	*Median*	*M*	*SD*	*Median*	*p-value*
Length	Words (*LgW*(*Q*))	7.8	4.1	7.0	5.2	2.3	5.0	***
	Concepts (*LgC*(*Q*))	2.8	1.4	3.0	1.9	1.0	2.0	***
% of concepts among words		**38.8**	**17.3**	**36.4**	**39.2**	**19.5**	**33.3**	*ns*
Specificity	Posting (*PSpe*(*Q*))	3.7	1.3	3.7	3.4	1.4	3.3	*
	Hierarchical (*HSpe*(*Q*))	1.05	0.34	1.04	0.91	0.34	0.84	***
Difficulty	Similarity score (*NSC*(*Q*))	39.2	5.7	40.9	38.4	6.3	40.5	*



(a) Query length in terms of words ($LgW(Q)$)

(b) Hierarchical specificity ($HSpe(Q)$)

Fig. 1. Empirical distributions of length and hierarchical specificity features according to participant's level of expertise.

formulated longer queries than novices, and in half cases, it was observed the occurrence of 2.7 terms more in queries issued by experts than those issued by novices. In Fig. 1.a, the empirical distribution of query length is plotted across the two groups of crowd workers (experts and novices); this figure clearly confirms our previous observation.

Examining the usage of a technical lexicon, here the MeSH concepts, it was also observed that experts use more concepts than novices ($M = 2.8$ vs. $M = 1.9$, $p\text{-}value < 0.0001$) which is consistent with the results reported in previous findings (White et al., 2009, 2008) on web search. Experts used more concepts than novices in 81% of submitted topics, and in half of cases, experts used more than one concept compared to novices. However it appears that on average, the MeSH terminology can only cover less than 40% of the query words whether users were experts or not ($M = 38.8$ vs. $M = 39.2$, $p\text{-}value = 0.87$). These results may demonstrate that novices searching in specialized medical repositories are able to use technical words but are significantly less able than experts. Another interesting result arising from this analysis is that the expertise significantly impacts the posting and hierarchical specificities; this result offers insight into the uniqueness of query words. As shown in Table 4, queries issued by experts exhibit higher hierarchical specificity ($M = 1.05$ on average) than queries submitted by novices ($M = 0.91$, $p\text{-}value < 0.001$). Fig. 1.b clearly confirms this result. On average, hierarchical specificity is higher for experts in 74% of all the submitted topics. In contrast, differences between the two groups of users in terms of posting specificity are less pronounced ($M = 3.7$ for queries issued by experts and $M = 3.4$ for queries issued by novices, $p\text{-}value < 0.05$). From these findings, we hypothesize that to express focused clinical information needs, novices are more likely to use unique words, whereas experts are more adept to the use of fine-grained medical concepts. Looking at the normalized similarity score which measures query difficulty, based on the corpus-query term overlapping, we can observe that expert queries were slightly more similar to documents of the collection than non-expert queries ($M = 39.2$ vs $M = 38.4$, $p\text{-}value < 0.05$), suggesting that the gap between the query vocabulary and the documents vocabulary is less important in the case of experts than in the case of novices.

In summary, we found that even in clinical searches, experts formulate longer queries and make use of more technical concepts than novices but the latter are slightly as able as experts to use unique and specific words. We also showed that in comparison to novice's language, expert's language used for query formulation better matches the language used to express the content of clinical documents.

4.2. Relevance assessment (RQ2)

The main objectives of this analysis were 1) to evaluate the relationship between the expertise level of assessors, the level of their relevance assessment, and the level of the relevance assessment task difficulty (Section 4.2.1); 2) to examine the level of agreement vs. disagreement on relevance assessment between the two categories of users namely, experts and novices but also among the participants of the same category Section 4.2.2), and 3) to gauge the level of difficulty of the

Table 5
Description of relevance according to groups of participants.

Groups	*Experts*		*Novices*	
# Topics	113		113	
	M (*SD*)	*Median*	*M* (*SD*)	*Median*
Relevance Score	0.65 (0.58)	0.50	0.87 (0.53)	0.60
Task difficulty				
Easy for 2 judges	0.60 (0.62)	0.40	0.83 (0.58)	0.50
Easy for 1 judge	0.74 (0.50)	0.70	0.95 (0.47)	1.00
Not easy for 2 judges	0.68 (0.38)	0.75	0.92 (0.38)	0.90

Table 6
Description of relevance judgments concordance within and between groups of participants.

Groups	*Experts*		*Novices*	
# Topics	113		113	
	M (*SD*)	*Median*	*M* (*SD*)	*Median*
Weighted Kappa value	0.20 (0.31)	0.08	0.09 (0.36)	0.00
Task difficulty				
Easy for 2 judges	0.21 (0.32)	0.02	0.13 (0.36)	0.00
Easy for 1 judge	0.25 (0.30)	0.19	0.08 (0.38)	0.00
Not easy for 2 judges	0.08 (0.29)	0.04	0.08 (0.27)	0.00
Agreement levels	*N*	%	*N*	%
Less than chance (<0)	18	16%	35	31%
Slight (0.01–0.20)	51	45%	44	39%
Fair (0.21–0.40)	18	16%	17	15%
Moderate (0.41–0.60)	9	8%	4	3%
Substantial (0.61–0.80)	10	9%	2	2%
Almost perfect (0.81–1)	7	6%	11	10%

relevance assessment task for both user's categories and explore the reasons for the difficulty if any and their relationship with the time spent to achieve the task (Section 4.2.3).

4.2.1. Analysis of assessor's relevance ratings

According to *Task 2* guidelines (Section 3.4.2), for each topic, 10 documents were presented to crowd workers and their relevance was assessed by 2 experts and 2 novices according to three levels of perceived relevance: *Relevant, Partially Relevant* and *Not Relevant*. For each assessor and each topic, a numerical document relevance score was first calculated based on qualitative relevance assessment (2: *Relevant*, 1: *Partially Relevant* and 0: *Not Relevant*) and then averaged across documents with respect to each pair of assessors from the same category (experts vs. novices). Thus, the computed numerical relevance scores ranged from 0 (if the 10 documents were assessed *not relevant* by both assessors) and 2 (if the 10 documents were all assessed *relevant* by both assessors).

Table 5 provides a summary of relevance scores for the two groups of crowd workers (*M* mean value, *SD* Standard Deviation and median value) and according to the difficulty of the task assessed by each assessor. We can observe from Table 5 that mean values for relevance scores were significantly higher for novices (0.87 on average) than for experts (0.65 on average, $p\text{-}value < 0.0001$). This observation may be explained by two complementary reasons: 1) expert' relevance assessments are more targeted and context-specific than novice' relevance assessments leading thereby to lower relevance scores from the experts' side, and that 2) novices are more likely to rely on content matching between query content and document content to assess relevance which clearly fits the principle used by the IR system to return candidate top relevant documents. The latter are more likely to be assessed as *relevant* or *partially relevant* by novices.

Turning our attention to both the relevance scores and the relevance assessment task difficulty, we can surprisingly notice that the relevance scores were lower for tasks assessed as *Easy* than those assessed as *Moderate or Difficult* for the two groups of users (experts vs. novices) ($p\text{-}value < 0.05$). Moreover, relevance scores were still higher for novices than for experts whatever the assessed difficulty of the task. One explanation is that the more comfortable the user is with both the search topic and the related documents, the more able he is to assess the right level of relevance with a bias toward low scores. Instead, when the task is perceived as difficult, the level of relevance is more likely to be unreliable with a bias toward higher scores. However, more investigations are needed to understand the reasons of the perceived task difficulty and its relationship with relevance assessment. This will be the focus of our subsequent analyses.

4.2.2. Analysis of assessor's agreement

Assessor's agreement was estimated using the weighted Cohen's Kappa cœfficient between assessors belonging to different groups and those belonging to the same group. Table 6 provides a summary of the Kappa values of concordance for the

Table 7
Categories of reasons behind the difficulty of document relevance assessment.

Code	Description	Example issued from an expert	Example issued from a novice
DU	Document Understanding: all statements that involve difficulty of interpretation of document content	"Pancreatitis is a cause for pancreatic pseudocycts, and there were several documents that described pancreatitis. Deciding whether those documents were A to the topic was a bit challenging"	"topic and the articles were slightly difficult to understand"
MI	Missing Information: all statements that involve that additional information was require to better assess the relevance of the document	"would have been easier if age of patient was provided, also ACE inhibitor medications have many types so need to read through to make sure there is not one medication mentioned in the study"	"when describing the disease is not clearly stated whether or not they are related to the T - cells"
SKR	Specific Knowledge Required: all statements that involve the lack of sufficient knowledge to assess the relevance of the documents	-	"a lot were about heart attack and treatment but not a lot to do with basic explanation for families"
QU	Query Understanding: all statement about the cognitive ability to interpret the topic	"needed to make sure it was about lactase deficiency THERAPY and not just lactase deficiency"	"Understanding the topic was a bit difficult"
GR	Graded Relevance: all statements related to a fine-grained level of relevance assessment	"Most of the documents described endarterectomy and related morbidity, but they didn't address the question when to perform"	-
DRR	Deep Reading of documents Required: all statements involving the need of in-depth reading of detailed or long documents before assessing their relevance	"abstracts were detailed and needed to be read fully to find both neuropathy and edema in them, however, most only B"	-

two groups of crowd workers (*M* Mean value, *SD* Standard Deviation and median value), and the repartition of assessments in the usual agreement levels (from *less than chance* for negative values to *almost perfect* for Kappa values greater than 0.8).

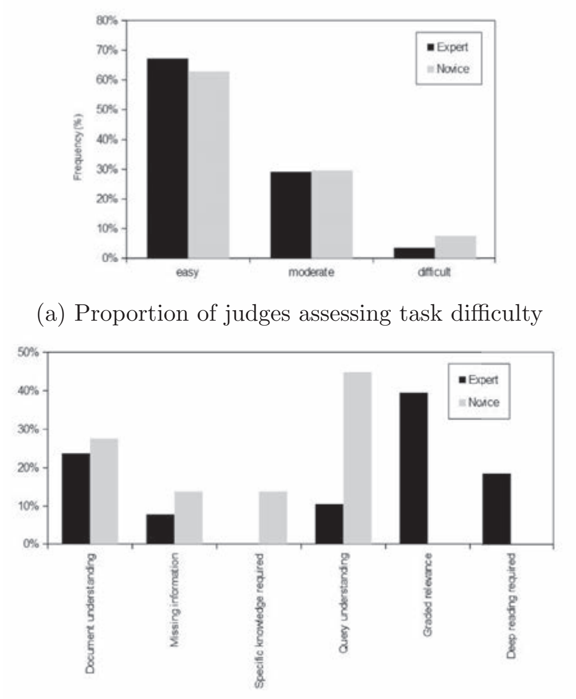
We can observe from Table 6, that the mean values of Kappa coefficient are very low for experts as well as for novices (0.20 for experts and 0.09 for novices). This indicates that relevance assessment agreement is low whether users are experts or not. More precisely, we can see that for half of the studied topics, Kappa values are less than 0.08 for expert assessors and 0.09 for novice judges ($p\text{-}value < 0.05$). We can also observe poor agreement for 16% of experts and for 31% of novices. When comparing the agreement averages using the paired *t*-test (since two agreements are related to the same topic), we can notice a slight significative difference between experts and novices agreements in favor of experts ($p\text{-}value < 0.05$). We can also observe that within the two groups of assessors, 15% and 12% of experts and novices agreements were substantial or more. Previous research has also shown that relevance agreement is low for medical experts primarily in clinical settings (Hersh et al., 2004; Hersh et al., 2005; Koopman & Zuccon, 2014; Roberts et al., 2015). Through this analysis, we extend the observation to novices as well and show that experts are however more concordant than novices in their perceived relevance. This observation could be explained by the fact that experts have stronger beliefs than novices about the relevance (vs. irrelevance) of documents but their domain knowledge better constrains the disparity of their beliefs in comparison with novices for whom domain knowledge is rather limited. We also computed the Kappa values depending on whether the task is considered as *Easy* by the two assessors, by only one assessor or by any assessor. From Table 6, we see that on average, the lower the agreement between the assessors belonging to the same category, the more difficult the relevance assessment task but this relationship was found statistically not significant neither for experts nor for novices ($p\text{-}value = 0.40$).

In summary, the agreement level between experts is slightly higher than the agreement level between novices, but whether expert or not, the agreement level is low and was not impacted by the assessed difficulty of the task.

4.2.3. Qualitative analysis of the relevance assessment task difficulty

In addition to considering the agreement between crowd workers regarding the relevance assessment task, we specifically examined the difficulty of this task. We collected from the crowd workers both difficulty ratings and qualitative comments about the reasons of the difficulty if any. To exploit these comments, 4 annotators who are students (2 graduate and 2 post-graduate students) analyzed and manually annotated the entire pool of comments. Each comment was qualitatively examined for content by 2 students and then categorized. For categorization, strong evidence regarding the participants feelings had to be expressed in the comment, according to the category description inferred by the human annotator. The agreement level between the two assessors was estimated using Cohens Kappa coefficient. We obtained a value equal to 0.70 for students who annotated expert's comments and a value equal to 0.80 for students who annotated novices' comments, which indicated substantial agreement. To check the annotation consistency between the annotators of the two types of comments (issued from experts vs. novices), the students met to make consensual decisions about the final categorization. Table 7 presents the proposed categorization.

First we assessed the number of participants involved in each group of participants (experts vs. novices) that assessed the task as *Easy, Moderate* or *Difficult*; the results are shown in Fig. 2a. Surprisingly, as can be seen from Fig. 2a, most of the



(a) Proportion of judges assessing task difficulty

(b) Proportion of reasons why the task was judged difficult or moderate

Fig. 2. Qualitative analysis of task difficulty.

assessors found that the task was *Easy*: 67% for experts and 63% for novices; χ^2 test between the assessor's group and the difficulty level factors ($\chi^2(2) = 3.58,\ p\text{-}value = 0.17 > 0.05$) found that the reported differences are not significant.

Second, we aimed to deepen our understanding of the reasons of the relevance assessment task difficulty by performing a qualitative analysis on the pool of comments issued from participants who found the task *Moderate* or *Difficult* (Cf. 3.4.2). Among the 452 judgements, only 158 judgers assessed that the task was *Moderate* or *Difficult*. However we found only 80 comments (50, 6%), 51 from experts and 29 from novices. Fig. 2b shows the qualitative differences in terms of what made the task difficult or moderate, according to categories of reasons presented in Table 7. We can see that for experts, the most frequent reason (39%) is about the graded relevance (GR) related to the fine-grained analysis of each document and of the results as a whole before assessing their relevance. This reason is followed by the one related to document understanding (DU) which involves a cognitive load related to an in-depth reading and interpretation of document content. For novices, query understanding (QU) was the most frequent reason of the relevance assessment task (45% of given reasons). However, we consider that this reason is not realistic in the daily-life search activity since the crowd workers performed simulated tasks with provided laboratory-controlled topics even they self-generated the queries. So the most effective frequent reason of relevance assessment difficulty to consider here is more likely to be document understanding (DU). Looking at the differences between the reasons mentioned in the experts' and novices' comments, we can see that two categories of difficulty reasons, namely graded relevance (GR) and deep reading of documents (DRR) are only mentioned by the experts while one specific category of difficulty reasons, namely specific knowledge required (SKR) is mentioned by the novices. This observation can obviously be explained by the differences in domain expertise of the crowd workers. It is worth to mention that even for common general reasons of relevance assessment difficulty as mentioned by both experts and novices (DU, MI, QU), the comments suggest that the practical difficulties faced by experts during the relevance assessment task are different than those faced by novices. For instance, regarding document understanding (DU), experts generally mentioned difficulties related to the lack of specific differentiation in document content leading to ambiguity considering the clinical case at hand. Unlikely, novices argue the difficulty to understand the general content of the document and recognize the need of higher-level skills to assess reliable relevance scores. This qualitative analysis reinforces the significant differences observed in the previous quantitative comparative analysis between experts and novices according to relevance assessment agreement. Furthermore, it partially explains the observed bias toward lower relevance scores assigned by experts than novices. Experts are more demanding of specific technical content and context-specific relevance indicators before providing high relevance scores.

Looking at the time spent assessing the relevance of documents, reported in Table 8, we can see that the novices spent less time assessing the relevance of documents than experts: only 33% of novices spent more than 2 min on the task against

Table 8
Time statistics regarding relevance assessment task.

Groups	*Experts*		*Novices*	
# Topics	113		113	
Spent time (in seconds	M (SD)	*Median*	M (SD)	*Median*
All	181 (83)	180	227 (416)	50
Easy	152 (74)	120	88 (93)	32
Moderately difficult	235 (62)	240	467 (624)	90
Difficult	308 (73)	300	453 (541)	240
More than 2 min	n / N	%	n / N	%
All	130/221	58%	75/226	33%
Easy	61/149	41%	36/142	25%
Moderately difficult	62/65	95%	29/67	43%
Difficult	7/7	100%	10/17	58%

58% of experts ($p\text{-}value < 0.0001$). This percentage was greater when the task is assessed as *Difficult*: from 41% for an *Easy* task to 100% for a *Difficult* task in the expert group, and from 25% for an *Easy* task to 58% for a *Difficult* task in the novice group ($p\text{-}value < 0.0001$). This observation 1) confirms the time-consuming issue that experts specifically may encounter in seeking for relevant clinical resources during their professional activities as outlined by previous work (Ely et al., 1999, 2002; Koopman & Zuccon, 2014); the findings about the qualitative reasons for relevance assessment difficulty gives insights on the main reasons that make the relevance assessment longer: in-depth reading of documents and context-specific targeted relevance assessment; 2) suggest a quick assessment of the returned results for novices. Qualitative reasons of the search difficulty as perceived by novices suggest that the lack of appropriate knowledge (SKR) and query understanding (QU) make the user not fully engaged in the search. Altogether the findings about relevance assessment agreement, task difficulty and time spent to assess the relevance of documents imply that: 1) relevance agreement depends on both domain expertise and perceived relevance considering document content interpretation, specifically for experts and 2) that time spent is undoubtly related to the difficulty level of the relevance assessment task and more precisely to the qualitative reasons for the perceived difficulty that significantly differs between experts and novices.

4.3. Impact of expertise on system performance (RQ3)

Our third research question concerns the evaluation of a traditional IR system towards the particular profile of the assessor (expert vs. novice) who provided the relevance assessments. To evaluate how levels of retrieval performance change according to the variability of the ground truth, we also considered different scenarios of building the ground truth within the assessors of the same category as detailed below. For the purpose of evaluating and comparing query performance across and within the groups of participants, both score-based and level-based performances were analyzed, as detailed below. The following evaluation resources were used under version 4.0 of the Terrier search engine[10]:

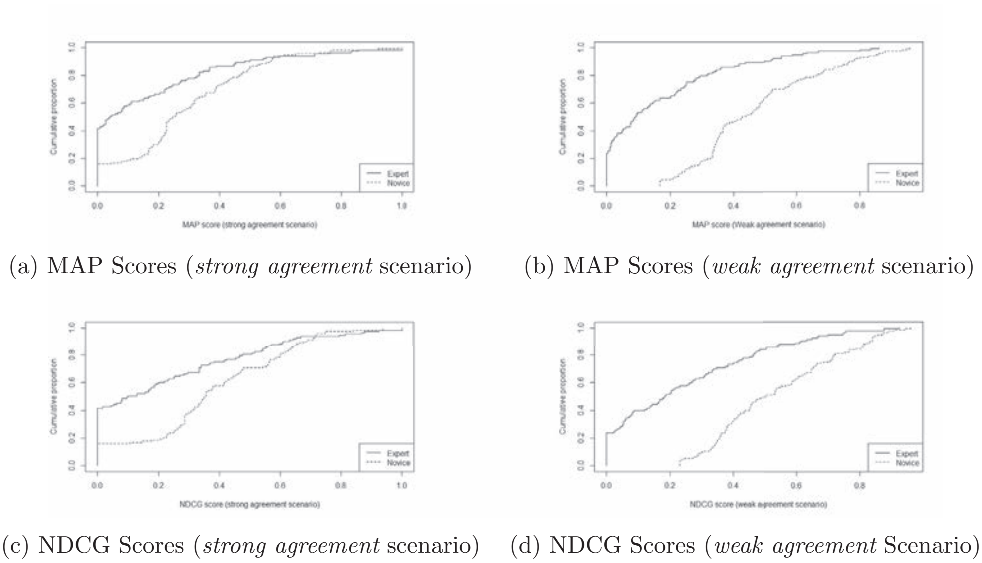
- *Performance measure:* The Mean Average Precision (MAP) and the Discount Cumulated Gain (DCG) are used to provide a single, overall measure of search performance. For evaluating the MAP measure, we considered both relevant and partially relevant documents under the same category of relevant documents. The performance measures have been computed using the standard TREC-eval tool[11].
  More precisely, for each topic (among the 113 topics), we considered the 6 formulated queries during *Task 1*. With respect to each category of crowd workers, we averaged the MAP and NDCG performance scores obtained using each of the 3 different queries related to the same topic.
- *Relevance assessments:* We used 2 scenarios for building the ground truth according to the assumed relevance assessment at the document level: 1) *weak agreement* where we assume that a document is relevant to a topic if at least one of the two assessors assessed the document as *partially relevant* or *relevant*. The relevance score of a document is computed as the maximal score obtained from the two assessors; this corresponds to a score aggregation built using the OR operator; 2) *strong agreement* where we assume that a document is relevant to a topic if at least the two assessors assessed the document as *Partially relevant* or *Relevant*. The relevance score of a document is computed as the minimal score obtained from the two assessors; this corresponds to a score aggregation built using the AND operator. This allows building for each topic 4 ground truth sets, 2 for each participant's group (experts vs. novices) and for each group, 2 ground truth sets related to both scenarios of relevant assessment assumptions (weak agreement vs. strong agreement).

Table 9 lists the MAP and NDCG performance scores of queries formulated by experts vs. novices according to the two scenarios described above. The search performance scores issued from expert and novice groups were compared by the

[10] http://www.terrier.org.
[11] http://trec.nist.gov/trec_eval.

Table 9
MAP and NDCG performance scores for experts and novices according to relevance agreement levels.

Groups # Topics (# Queries)	*Expert group* 113 (339)			*Novice group* 113 (339)			
	M	*SD*	*Median*	*M*	*SD*	*Median*	*p-value*
MAP performance scores							
Strong agreement	0.17	0.23	0.09	0.29	0.21	0.25	***
Weak agreement	0.17	0.24	0.18	0.47	0.20	0.43	***
NDCG performance scores							
Strong agreement	0.22	0.27	0.10	0.37	0.23	0.35	***
Weak agreement	0.25	0.24	0.18	0.53	0.19	0.50	***



(a) MAP Scores (*strong agreement* scenario)

(b) MAP Scores (*weak agreement* scenario)

(c) NDCG Scores (*strong agreement* scenario)

(d) NDCG Scores (*weak agreement* Scenario)

Fig. 3. Empirical distributions of two performance scores by expertise level for two scenarios.

paired *t*-test and the corresponding *p*-values are given in the last column. We can clearly see that the overall result tendencies are the same for both MAP and NDCG performance measures. Interestingly, as shown by the mean score values (*M*), we can see from Table 9 that the performance scores are higher for novices than for experts: for example, for MAP (scenario *weak agreement*), $M = 0.17$ in the expert group vs. $M = 0.47$ in the novice group ($p\text{-}value < 0.0001$). This suggests that-basically speaking- novice queries are more successful than expert queries. This may be due to several concomitant reasons that we could infer from our previous analysis. Since traditional IR systems (like the one we used in our experiments) are based on simple term matching between documents and queries, top results are more likely to match expert queries than novice queries. Indeed as outlined by the query formulation analysis (Section 4.1) the gap between the query language and the document language is less pronounced in the case of experts than in the case of novices. However, given a topic, a document is more likely to be assessed as relevant by novices (who assess the relevance of documents answering a provided topic) who mostly rely on the same evidence to assess document relevance than the IR system to rank the documents at the top. Unlikely experts look at more specific relevance indicators-beyond term sharing between queries and documents)- as outlined by our qualitative analysis about relevance assessment difficulty (Section 4.2.3) to assess document relevance; experts rather look deeply at the documents with regard to the different interpretations of the topic. Moreover, as found in our previous analysis of relevance rating distribution (Section 4.2.1), expert relevance ratings are lower for experts than for novices leading to lower the score performance either in the case of *Weak agreement* or *Strong agreement* scenarios. From the empirical distributions of the scores plotted in Fig. 3, it can be confirmed that domain expertise leads to significant differences in the performance scores considering whether the *Weak agreement* or the *strong agreement* scenario, with higher values for the novice group. However, it was observed that differences in performance between expert queries and novice queries were more pronounced in the case of the *weak agreement* scenario, than in the case of the *strong agreement* scenario. This observation could be simply explained by the higher level of agreement between experts than between novices in relevance assessment as found in our previous analysis about assessor's agreement (Section 4.2.2). It is also interesting to reveal that the impact of the level of assessor's agreement (*weak agreement* vs. *strong agreement*) on performance scores was more pronounced for novices than for experts.

Table 10
Topic repartition into the four performance level categories (*failure, low, middle and high*) considering user's expertise level : #topics (% among topics in the expertise level).

Performance	User's expertise level			Comparison
level	Values range	Expert	Novice	(χ^2 *p*-value)
	based on MAP with *strong agreement* scenario			
Failure	0	47 (42%)	118 (16%)	
Low	[0–0.04[	9 (8%)	0 (0%)	
Middle	[0.04–0.29[	30 (26%)	44 (39%)	
High	$\geq$0.29	27 (27%)	51 (45%)	***
	based on MAP with *weak agreement* scenario			
Failure	0	27 (24%)	0 (0%)	
Low	[0–0.17[	45 (40%)	4 (4%)	
Middle	[0.17–0.40[	25 (22%)	48 (42%)	
High	$\geq$ 0.40	16 (14%)	61 (54%)	***
	based on NDCG with *strong agreement* scenario			
Failure	0	47 (42%)	18 (16%)	
Low	[0–0.10[	10 (9%)	0 (0%)	
Middle	[0.10–0.38[	27 (24%)	46 (41%)	
High	$\geq$ 0.38	29 (25%)	49 (43%)	***
	based on NDCG with *weak agreement* scenario			
Failure	0	27 (24%)	0 (0%)	
Low	[0–0.27[	41 (36%)	6 (5%)	
Middle	[0.27–0.48[	26 (23%)	49 (43%)	
High	$\geq$ 0.48	19 (17%)	58 (52%)	***

To complete the comparative study on system performance, levels of performance are analyzed using qualitative intervals rather than ordinal scales. To this end, the topics were categorized into *Failure, Low, Middle* and *High* performance based on the MAP and NDCG scores. As performed in previous work (Cronen-Townsend, Zhou, & Croft, 2002), the topic categorization was established by performing kernel density estimation, whose 33% and 66% percentiles were computed (and denoted by $P_{33\%}$ and $P_{66\%}$, respectively). More specifically, the *Failure* category included topics with a performance score equal to 0; the subsequent categories, namely, *Low, Middle* and *High*, included topics with a performance score ranging within the intervals of $]0 \ldots P_{33\%}]$, $]P_{33\%} \ldots P_{66\%}]$ and $[P_{66\%} \ldots 1]$, respectively. Table 10 presents the percentiles obtained using each scenario and each performance level. To assess the relationship between the domain expertise level and the performance level, the two groups of users were categorized into *Failure, Low, Middle* and *High*-performance categories. The statistics related to each group and each performance category are presented in Table 10. The comparisons were tested using a χ^2 test, and the corresponding *p-values* are given in the last column. As can be seen from Table 10, regardless of performance scores (MAP or NDCG) or scenario (*weak agreement* or *strong agreement*), significant differences were found in the performance levels between expert-based searches and novice-based searches (p-*value* < 0.0001 for all scores and all scenario), as outlined in the previous comparative study based on the ordinal performance scores. Queries formulated by experts were characterized by a stronger percentage of failure and low performance (ranging from 50% to 64%) in comparison with the queries formulated by novices (ranging from 4% to 16%). In the same context, the percentage of high performance is largely higher among novices (ranging from 43% to 54%) than among experts (ranging from 14% to 25%).

5. Discussion and design implications

The study investigated the differences and commonalities between expert-oriented and novice-oriented clinical searches using library resources. The results show that queries issued by experts are longer than those issued by non-experts, as previously shown in web searches (White et al., 2009, 2008); Moreover, consistent with previous findings in the medical web searches (White et al., 2008), the results show that experts searching medical repositories are more adept at utilizing the technical lexicon than those searching on the web. In addition to analyzing the length, the specificity of query formulations was investigated. It appears that novices are more likely to use unique words to express specific notions; however, experts appeared to be more adept at using fine-grained technical concepts without any correlation between term specificity and semantical hierarchical specificity as previously shown (Tamine et al., 2015). These results are also consistent with previous preliminary findings (Palotti et al., 2014) who indicated that the number of words and number of concepts in the queries was found as good indicators for inferring user expertise in the medical domain.

The findings regarding relevance assessment of search results may be partially consistent with previous work which mainly focused on the study of relevance assessment agreement among experts. These findings identified that there is a low agreement between medical experts (Hersh et al., 2004; Hersh et al., 2005; Koopman & Zuccon, 2014; Roberts et al., 2015). For instance in the TREC Genomics track, the overlap of relevance judgments was only able to achieve 0.51 in 2004 (Hersh et al., 2004) and 0.59 in 2005 (Hersh et al., 2005), while the agreement does not exceed 0.44 in the TREC Medical track

(Roberts et al., 2015). It is worth to mention that similar findings are reported by previous studies even out-of the medical domain (Vakkari & Sormunen, 2004; Voorhees, 2000). Our study reveals the same trend among experts but also among novices. We believe that the higher level of agreement among experts than among novices is due to domain knowledge that constraints the interpretation of document content for relevance appraisal. Our study is consistent with previous work in the medical domain (Hersh et al., 2002; Koopman & Zuccon, 2014; Zhu & Carterette, 2012) in the sense that it reinforces the idea that judging is highly subjective and multidimensional, thereby leading to diverse interpretations among experts as the main reason of disagreement between them. In addition, our study suggests that the lack of appropriate domain knowledge increases the risk of disparity and erroneous relevance judgments among novices that could lead to misleading interpretations as shown in previous work (White & Horvitz, 2009). Results also indicated that the reasons for relevance assessment difficulty significantly differ with regard to domain expertise of users. Regarding experts, our findings highlight that the cognitive load is mostly due to the need of in-depth reading and interpretation of the document contents and the need of assessing graded relevance with respect to the different possible interpretations of the results. This implies a significant amount of time to provide accurate and reliable relevance assessments. From a wide point of view, these findings are consistent with the preliminary results provided in Koopman and Zuccon (2014). In contrast, the lack of appropriate knowledge is the most reason mentioned by novices who exhibited furthermore relative quick relevance assessments. For both novices and experts, the difficulty of the relevance assessment task is not without relationship with the time spent to achieve the task.

To further probe the differences between experts and non-experts, query performance was computed using the MAP and NDCG scores based on various pools of gold standard. Our aim was to determine the impact of domain expertise but also the variations in relevance assessment agreement on system performance comparisons for effectiveness measures considering both numerical and qualitative scores. In our study, the difference in relevance assessments for a particular result originates from 1) the difference in the expertise level and 2) the difference in personal opinions of assessors. It is worth to mention that the agreement vs. disagreement made for a document occurs with respect to different query formulations issued from Task 1 collected from the same information need (topic or clinical case description). Since laboratory-based evaluation significantly contributed to the validation of IR models, a long-standing previous research focused on the issue of evaluating the impact of variations in relevance assessments-according to diverse attributes such as expertise and document content- on system rankings (Bailey et al., 2008; Carterette & Soboroff, 2010; Demeester, Aly, Hiemstra, Nguyen, & Develder, 2015). For instance, Bailey et al. (2008) showed that task and domain expertise have significant effects on document rankings. Through our study results, we confirm those previous findings. It clearly appears that expertise, by nature, significantly impacts clinical search performance. Interestingly, we found that whatever the level of agreement between assessors of the same group, performance scores of queries issued by novices are higher than performance scores of those issued by experts suggesting that novices assess relevance using the same evidence used by traditional IR models to rank documents at the top. In contrast, experts leverage from their knowledge and their personal understanding of the medical case to build a self-perception of multi-evidence based relevance that goes far beyond term overlapping between the query and the document. The results have the same trend when considering qualitative ranges of performance.

The findings provide a useful step forward in a number of research directions. We discuss here two theoretical and one practical implications for designing future medical IR systems that are revealed by our results.

- Our findings indicate that queries issued from experts are significantly different from those issued from novices according to several pre-retrieval facets including length, specificity of the vocabulary used for their formulation and their difficulty in terms of the degree of matching between the query and document vocabularies. The study results particularly points out the gap between the vocabulary used by novices for formulating their information needs and the vocabulary used by experts for reporting their clinical findings in library clinical documents. Based on these findings, the implications for further theoretical investigation is to develop models to predict expertise based on those query-related features. The expertise prediction would be a prior step to an evolving automatic query suggestion that would leverage from user (novice)-system interactions with the aim of reducing the effect of the language barrier. Methods already exist for automatic clinical query suggestion (Lu, Wilbur, McEntyre, Iskhakov, & Szilagyi, 2009) but should be revised toward a better personalization of the suggestion process through the use of evidence issued from the user's search intent with respect to his level of expertise instead of using popular queries.
- The retrieval performance evaluation results reported in the study demonstrated that traditional IR models which mainly consider the presence or absence of query terms within documents are particularly unsuccessful for experts who rather leverage from their knowledge and past experience to assess a multi-dimensional relevance. Ranking documents according to multiple dimensions of relevance is not new in the IR field (Taylor, Cool, Belkin, & Amadio, 2007) but a further research is needed to explore the relevant dimensions to particularly consider in clinical search settings as well as their interactions with domain expertise. Studying expert search behavior within multi-session searches and across a number of taxonomic searches including searching for potential diagnosis given a set of symptoms, searching for effective treatments given a medical case, may lead to the identification of a set of dimensions of relevance experts may rely on. Such dimensions could help system designers to formalize new relevance-based expert models. Additionally, even though traditional recall-precision measures give a general view of system performance, they may have a limited value in the assessment of how well the IR models work in realistic clinical search settings within the constraints imposed by such dimensions; Thus, a relevant theoritical investigation is needed to formalize evaluation measures that put empha-

sis on the cognitive load, speed of task completion and topic coverage to cite but a few. Measures such as the cube-test (Luo, Wing, Yang, & Hearst, 2013) and the $\alpha nDCG$ (Clarke et al., 2008) could be used as the basis but need to be improved for better addressing the evaluation of a clinical search task.

- The qualitative comparative analysis of the relevance assessment task in terms of difficulty revealed significant differences in the underlying reasons perceived by experts vs. novices. Experts spent a considerable amount of time for assessing a graded relevance through in-depth reading and interpretation of the documents with respect to their understanding of the topic. In contrast, novices mainly acknowledged the lack of adequate knowledge to assess the relevance vs. irrelevance of documents. One relevant practical implication we envision is the design of system-mediated collaboration between novices and experts and also among experts through social document annotation. Experts could tag, while reading the clinical documents, for ease retrieval and understanding by themselves or others (experts or novices) or provide their point of view to share context and experience with others (experts or novices). While social tagging techniques (Gupta, Li, Yin, & Han, 2010) have already shown their merits in providing search assistance, our study findings highlight that there still much that can be done to achieve targeted solutions in the specific case of clinical search. Examples of remaining challenges are: 1) making expert tags understandable by novices with regard to the language barrier which has been revealed by the query formulation analysis; 2) evaluating the objectivity vs. subjectivity of the tags while providing assistance to experts or novices since the findings highlighted a low level of relevance assessment agreement that may be due to personal interpretations of document content.

Beyond medical search, we believe that the trend of our results remain in the case of health search; however further research is needed to deal with the specifities of health-related queries (eg. queries about diverse impairments of human beings formulated by diverse health professionals) before assessing the reliability of those implications on health search in general.

This study is not without limitations. First, since we used crowdsourced users and given that the topics were pre-defined, such users may have been not self-motivated to accurately complete the relevance assessment task, particularly in the case of novices. Hence, bias may have been introduced within document relevance ratings and time spent to achieve the task under time pressure. While it is difficult to assess user's engagement in the relevance assessment task, the time used for the task achievement was used with care by comparing to levels of time intervals rather than absolute values. We believe that the trend of our results remain however reliable. Second, the features used in this study are insufficient in terms of revealing other aspects of the possible differences that could arise between experts and novices and impact their perceived relevance of the results, as well as their feeling about task difficulty. Further work is needed to develop additional features that capture factors beyond the query formulation and search performance such as user behavioral facets (eg. clickthrough data, session length, formulated queries, etc.) that can expand the study findings.

6. Conclusion

Medical information search is a common pursuit in the daily life of an increasing number of users either experts or not. Even medical search services grown in popularity, there is a lack of studies that investigated the differences between medical-related searches involved by experts and novices using clinical resources. We employed two crowdsourcing platforms to gain access to experts and novices. In this study, it was found that expert-based searches are significantly different than novice-based searches with respect to all considered facets. The analyses revealed that there is a more pronounced gap between novice's query language and document language and that novices formulate shorter and less technical queries even they have been found to be able to employ specific medical terms. The findings also highlight that the levels of relevance agreement are low for both experts and novices with a greater concordance between experts. The analyses revealed that even the quantitative differences about the perceived difficulty of the relevance assessment task between experts and novices were not different, the qualitative reasons were significantly different. Experts are faced to document and relevance interpretation difficulties while novices are faced to the lack of appropriate knowledge for relevance appraisal. These reasons directly impact the time spent to achieve the task. We also showed that IR systems based on traditional query-document matching models favor the success of queries issued by novices within various sets of gold standard built using different scenarios related to both domain expertise and levels of agreement between assessors.

Because the study focused on understanding the peculiarities of expert-based vs. novice-based clinical information searches, it is hoped that the findings may help the design of future medical and health-related IR systems that consider the level of expertise of seekers and the use of such evidence to provide more targeted answers that lead to user's satisfaction.

7. Acknowledgment

We acknowledge the support of Paul Sabatier University under the grant project ERC.

References

Alonso, O., Rose, D. E., & Stewart, B. (2008). Crowdsourcing for relevance evaluation. *SIGIR Forum, 42*(2), 9–15.

Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vries, A., & Yilmaz, E. (2008). Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st annual international acm sigir conference on research and development in information retrieval. SIGIR '* (pp. 667–674). ACM.

Bhavnani, K. (2001). Important cognitive components of domain-specific search knowledge. In *Proceedings of the tenth text retrivel conference. TREC '01* (pp. 571–578).

Bin, L., & Lun, K. (2001). The retrieval effectiveness of medical information on the web. *International Journal of Medical Informatics, 62*(3), 155–163.

Carterette, B., & Soboroff, I. (2010). The effect of assessor error on ir system evaluation. In *Proceedings of the 33rd international acm sigir conference on research and development in information retrieval. SIGIR '10* (pp. 539–546). New York, NY, USA: ACM.

Choudhury, M., Morris, M., & White, R. W. (2014). Seeking and sharing health information online: comparing search engines and social media. In *Proceedings of the acm chi conference* (pp. 536–545).

Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., & MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international acm sigir conference on research and development in information retrieval. SIGIR '08* (pp. 659–666). ACM.

Cohen, A., Stavri, P., & Hersh, W. (2004). A categorization and analysis of the criticisms of evidence-based medicine. *International Journal of Medical Informatics, 73*(1), 35–43.

Cronen-Townsend, S., Zhou, Y., & Croft, W. B. (2002). Predicting query performance. In *Proceedings of the 25th annual international acm sigir conference on research and development in information retrieval. SIGIR '02* (pp. 299–306). New York, NY, USA: ACM. doi:10.1145/564376.564429.

Davis, C. S. (2002). *Statistical methods for the analysis of repeated measurements*. Springer.

Demeester, T., Aly, R., Hiemstra, D., Nguyen, D., & Develder, C. (2015). Predicting relevance based on assessor disagreement: analysis and practical applications for search evaluation. *Information Retrieval Journal, 19*(3), 284–312.

Dinh, D., & Tamine, L. (2011). Combining global and local semantic contexts for improving biomedical information retrieval. In *European conference on information retrieval (ecir)* (pp. 375–386).

Dinh, D., Tamine, L., & Boubekeur, F. (2013). A study on factors affecting the effectiveness of biomedical document indexing and retrieval based on terminologies. *Artificial Intelligence in Medicine Journal, 57*(2), 155–167.

Ely, J., Jerome, M., Osheroff, A., Chamblis, M. L., Mark, M., Ebbell, H., … Rosenbaum, E. (2005). Answering physicians's clinical questions: obstacles and potential solutions. *Journal of the Ameican Medical Informatics Association, 12*(2), 217–224.

Ely, J., Jerome, M., Osheroff, A., Saverio, M. D., Maveglia, M., Marcy, M. D., & Rosenbaum, E. (2007). Patient-care questions that physicians are unable to answer. *Journal of the American Medical Informatics Association, 14*(4), 407–414.

Ely, J., Osheroff, A., Ebbell, H., Bergus, G., Levy, B., Chamblis, M. L., & Evans, E. (1999). Analysis of questions asked by family doctors regarding patient care. *British Medical Journal, 319*(7206), 358–361.

Ely, J., Osheroff, A., Ebbell, H., Chamblis, M. L., Vinston, M. L., & Stevermer, J. J. (2002). Obstacles to answering doctor's questions about patient care with evidence. *British Medical Journal, 324*, 1–7.

Ely, J., Osheroff, J., Gorman, P., Ebell, M. H., Chambliss, M. L., Pifer, E., & Stavri, P. (2000). A taxonomy of generic clinical questions: classification study. *Biomedical Journal, 32*(1), 429–432.

Eysenbach, G., Powell, J., & Englesakis, M. (2004). Health virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *Biomedical journal, 328*(7449), 1166.

Fox, K., & Duggan, M. (2013). Health online 2013. *Technical Report*. Pew Internet & American Life Project.

Fox, S. (2011). 80% of internet users look for health information online. *Technical Report*. Pew Research Center.

Francke, A., Smit, M., & de Veer, A. (2008). Factors influencing the implementation of clinical guidelines for health care professionals: a systematic meta-review. *BMC Medical Information Decision Making, 8*(38), 1–11.

Gupta, M., Li, R., Yin, Z., & Han, J. (2010). Survey on social tagging techniques. *SIGKDD Explorations Newsletter, 12*(1), 58–72.

Hembrooke, H. A., Granka, L. A., Gay, G. K., & Liddy, E. D. (2005). The effects of expertise and feedback on search term selection and subsequent learning: research articles. *Journal of American Society in Information Science and Technology, 56*(8), 861–871.

Hersh, H., Buckley, C., Leone, T., & Hickam, D. (1994). Ohsumed: an interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th annual international acm sigir conference on research and development in information retrieval. SIGIR '94*. ACM.

Hersh, W., Bhupatiraju, R. T., Ross, L., Jonhson, P., Cohen, A. M., & Kraemer, D. F. (2004). Trec 2004 genomics track overview. In *Proceedings of the text retrieval conference trec*. NIST.

Hersh, W., Cohen, A. M., Yang, J., Bhupatiraju, R. T., & Roberts, P. (2005). Trec 2005 genomics track overview. In *Proceedings of the text retrieval conference trec*. NIST.

Hersh, W., Crabtree, M., Hickam, D., Lynetta, M., Charles, M., Friedman, P., … Kraemer, D. (2002). Factors associated with success in searching medline and applying evidence to answer clinical questions. *Journal of the American Medical Informatics Association, 9*(3), 283–293.

Hersh, W., Crabtree, M., Hickam, D., Sacherek, L., Rose, L., & Friedman, C. (2000). Factors associated with successful answering of clinical questions using an information retrieval system. *Bull Medical Library Association, 88*, 323–331.

Hersh, W., & Hickam, D. (1994). Use of a multi-application computer workstation in a clinical setting. *Bulletin of the Medical Library Association, 82*, 382–389.

Ingwersen, P., & Belkin, N. (2004). Information retrieval in context - irix: workshop at sigir 2004 - sheffield. *SIGIR Forum, 38*(2), 50–52. doi:10.1145/1041394.1041405.

Jeong, J.-W., Morris, M. R., Teevan, J., & Liebling, D. J. (2013). A crowd-powered socially embedded search engine.. *ICWSM*.

Jones, S. (1972). A statistical interpretation of term specificity and its application to retrieval. *Journal of documentation, 28*(1), 11–20.

Kim, G. (2006). Relation between index term specificity and relevance judgment. *Information Processing and management (IPM), 42*, 1218–1229.

Kittur, A., Chi, E., & Suh, B. (2008). Crowdsourcing user studies with mechanical truck. In *Conference on human interaction chi* (pp. 453–456). ACM.

Koopman, B., & Zuccon, G. (2014). Why assessing relevance in medical ir is demanding. In *Proceedings of the medical ir workshop, in conjunction with the 2014 annual international acm sigir conference on research and development in information retrieval*. Gold Coast, Australia: ACM.

Lacroix, E., & Mehnert, R. (2002). The us national library of medicine in the 21st century: expanding collections, nontraditional formats, new audiences. *Health Information Librarian Journal, 19*(3), 126–132.

Lease, M., & Yelmaz, E. (2011). Crowdsourcing for information retrieval. *ACM SIGIR Forum, 45*(2), 66–75.

Lesk, M. E., & Salton, G. (1968). Relevance assessments and retrieval system evaluation. *Information storage and retrieval, 4*(4), 343–359.

Limsopatham, N., Macdonald, C., & Ounis, I. (2013). Learning to selectively rank patients' medical history. In *Proceedings of the 22nd acm international conference on conference on information knowledge management. CIKM '13* (pp. 1833–1836). New York, NY, USA: ACM. doi:10.1145/2505515.2507874.

Liu, C., Liu, J., Cole, M., Belkin, N., & Zhang, X. (2012). Task difficulty and domain knowledge effects on information search behaviors. In *Proceedings of the american society in information science and technology. ASSIST '12* (pp. 1–10).

Liu, J., Kim, C., & Creel, C. (2014). Exploring search task difficulty reasons in different task types and user knowledge groups. *Information Processing and Management, 51*(3), 273–285.

Lu, Z., Wilbur, W. J., McEntyre, J. R., Iskhakov, A., & Szilagyi, L. (2009). Finding query suggestions for pubmed. In *AMIA Annual Symposium Proceedings* (pp. 396–400).

Luo, J., Wing, C., Yang, H., & Hearst, M. (2013). The water filling model and the cube test: multi-dimensional evaluation for professional search. In *Proceedings of the 22nd acm international conference on information knowledge and management. CIKM '13* (pp. 709–714). New York, NY, USA: ACM.

Lykke, M., Price, S., & Delcambre, L. (2012). How doctors search: a study of query behaviour and the impact on search results. *Information Processing Management, 48*(6), 1151–1170.

Natarajan, K., Stein, D., Jain, S., & Elhadad, N. (2010). An analysis of clinical queries in an electronic health record search utility. *International journal of medical information, 79*(7), 515–522.

NLM (2012). Key MEDLINE indicators. *Technical Report*. US National Library of Medicine, National Institutes of health.

Palotti, J., Hanbury, A., & Henning, M. (2014). Exploiting health related features to infer user expertise in the medical domain. *Workshop on log-based personalization*.

Pratt, W., & Wasserman, H. (2000). Querycat: automatic categorisation of medline queries. In *Proceedings of the amia conference. AMIA'00* (pp. 655–659).

Radlinski, F., Kurup, M., & Joachims, T. (2008). How does clickthrough data reflect retrieval quality. In *Conference on information and knowldge management CIKM*. ACM.

Roberts, K., Simpson, M., Demner-Fushman, D., Voorhees, E., & Hersh, W. (2015). State-of-the art in biomedical literature retrieval for clinical cases: a survey of the trec 2014 CDS track. *Information retrieval*.

Robertson, S., & Hull, D. (2000). The tre-9 filtering track final report. In *Proceedings of trec 2000. TREC'00*, NIST Special Publication 500-249, 25–40.

Sackett, D. L. (1997). Evidence-based medicine. *Seminars in perinatology, 21*(1), 3–5.

Soldaini, L., Yates, A., Yom-Tov, E., Frieder, O., & Goharian, N. (2016). Enhancing web search in the medical domain via query clarification. *Information Retrieval, 19*(1), 149–173.

Song, F., & Croft, W. B. (1999). A general language model for information retrieval. In *Proceedings of the 22nd international acm sigir conference on research and development in information retrieval (sigir)* (pp. 279–280).

Spink, A., Yang, Y., Jansen, J., Nykanen, P., Lorence, D. P., & Ozmutlu, S. (2004). A study of medical and health queries to web search engines. *Health information and libraries journal, 21*, 44–51.

Stokes, N., Cavedon, Y., & Zobel, J. (2009). Exploring criteria for succesful query expansion in the genomic domain. *Information retrieval, 12*, 17–50.

Suominen, H., Salanter, S., Velupillai, S., Chapman, W., Savova, G., Elhadad, N., … Zuccon, G. (2013). Overview of the share/clef ehealth evaluation lab 2013. In P. Forner, H. Muller, R. Paredes, P. Rosso, & B. Stein (Eds.), *Information access evaluation. multilinguality, multimodality, and visualization. Lecture Notes in Computer Science: 8138* (pp. 212–231). Springer Berlin Heidelberg.

Tamine, L., Chouquet, C., & Palmer, T. (2015). Analysis of biomedical and health queries: lessons learned from trec and clef evaluation benchmarks. *Journal of the Association for Information Science and Technology, 66*(12), 2626–2642.

Taylor, A. R., Cool, C., Belkin, N. J., & Amadio, W. J. (2007). Relationships between categories of relevance criteria and stage in task completion. *Information Processing and Management, 43*(4), 1071–1084.

Vakkari, P., & Sormunen, E. (2004). The influence of relevance levels on the effectiveness of interactive information retrieval. *J. Am. Soc. Inf. Sci. Technol., 55*(11), 963–969.

Voorhees, E. M. (2000). Variations in the relevance of judgments and the measurement of retrieval effectiveness. *Information Processing and Management, 36*(5), 697–716.

Voorhees, E. M., & Hersh, W. (2012). Overview of the trec 2012 medical records. In *Proceedings of the text retreival conference trec*. NIST.

White, R., Dumais, S., & Teevan, J. (2009). Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the second acm international conference on web search and data mining. WSDM'09*.

White, R., & Hassan, A. (2014). Belief dynamics in web search. *Journal of the Association for Information Science and Technology, 65*(1), 2165–2178.

White, R., & Horvitz, E. (2009). Cyberchondria: studies of the escalation of medical concerns in web search. *ACM Transactions on Information Systems, 27*(4), 1–37.

White, R., & Horvitz, E. (2015). Belief dynamics and biases in web search. *ACM Transactions on Information Systems, 33*(4), 18.

White, R. W., Dumais, S., & Teevan, J. (2008). How medical expertise influences web search interaction. In *Proceedings of the 31st annual international acm sigir conference on research and development in information retrieval. SIGIR '08* (pp. 791–792). New York, NY, USA: ACM. doi:10.1145/1390334.1390506.

Wildemuth, B. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology, 55*(3), 246–258.

Yang, L., Mei, Q., Zheng, K., & Hanauer, D. A. (2011). Query log analysis of an electronic health record search engine. In *Proceedings of the annual symposium amia*. In *AMIA '11* (pp. 915–924).

Zhang, Y., & Fu, W. T. (2011). Designing consumer health information systems: what do user-generated question tell us? In *Proceedings of the 6th international conference on foundations of augmented cognition: directing the future of adaptive systems* (pp. 536–545).

Zhao, Y., F. Scholer, F., & Tsegay, Y. (2008). Effective pre-retrieval query performance prediction using similarity and variability evidence. In *Advances in information retrieval. Lecture Notes in Computer Science: 4956* (pp. 52–64). Springer Berlin Heidelberg.

Zhu, D., & Carterette, B. (2012). Combining multi-level evidence for medical record retrieval. In *International workshop on smart health and well being* (pp. 49–56).