What Makes an Effective Clinical Query and Querier?

Bevan Koopman
Australian e-Health Research Centre, CSIRO, Brisbane, Australia

Guido Zuccon
School of Electrical Engineering & Computer Science, Queensland University of Technology, Brisbane, Australia

Peter Bruza
School of Information Systems, Queensland University of Technology, Brisbane, Australia

In this paper, we perform an in-depth study into how clinicians represent their information needs and the influence this has on information retrieval (IR) effectiveness. While much research in IR has considered the effectiveness of IR systems, there is still a significant gap in the understanding of how users influence the effectiveness of these systems. The paper aims to contribute to this by studying how clinicians search for information.
Multiple representations of a information need — from verbose patient case descriptions to ad-hoc queries — were considered in order to understand their effect on retrieval. Four clinicians provided queries and performed relevance assessment to form a test collection used in this study. The different query formulation strategies of each clinician, and their effectiveness, were investigated.
The results show that query formulation had more impact on retrieval effectiveness than the particular retrieval systems used. The most effective queries were short, ad-hoc keyword queries. Different clinicians were observed to consistently adopt specific query formulation strategies. The most effective queriers were those who, given their information need, inferred novel keywords most likely to appear in relevant documents.
This study reveals aspects of how people search within the clinical domain. This can help inform the development of new models and methods that specifically focus on the query formulation process to improve retrieval effectiveness.

Introduction

Information retrieval (IR) systems have the fundamental purpose of promoting interactions with information that support people to achieve their goals and agendas in a wide variety of situations. Although it has long been held that IR research and practice must be based on an understanding of the people for whom the systems are intended, there is still a large gap between the study of the users of the systems as opposed to the algorithms underpinning the systems and how these are evaluated (@warning Citation 'James-Allan2012Frontiers-Chall' on page 1 undefined). This paper aims to be a stepping stone in bridging this gap by understanding how clinicians engage in medical information retrieval. In particular, our concern is both to understand what makes a good clinical *query* as well as to understand what makes a good clinical *querier*.

In this paper, we perform an in-depth study into how clinicians represent their information needs and the influence this had on retrieval effectiveness. Unlike the standard approach in IR evaluation of a single query per information need, we considered three different representations of an information need: i) verbose patient case descriptions (78 words per topic); ii) shorter patient case summaries (22 words per topic) of the patient case description; and iii) short ad-hoc queries (4.2 words per topic) expressed by clinicians. All three representations were realistic queries taken from a real-word clinical search scenario. These multiple representations of a single information need were used to retrieve clinical documents via a number of retrieval systems. Medical professionals were employed to provide relevance assessments of the retrieved documents, which allows the effectiveness of different query representations to be studied. In addition, we studied the different query formulation strategies of different clinicians to understand what constituted an effective clinical querier. More specifically, this paper aims to answer the following overarching questions:

What makes a good clinical query?

- How did different representations of the same information need — from ad-hoc queries through to verbose patient case descriptions — influence retrieval effectiveness? Were human-derived ad-hoc queries more effective than verbose case descriptions?
- What was the variation in effectiveness for different ad-hoc queries and for different clinicians? When was ad-hoc querying best or worst?

What makes a good clinical querier?

- To what extend did clinicians select keywords from the patient case description to form their ad-hoc queries; or did they derive unique keywords? Which method was more effective?
- Are there specific query strategies between clinicians that proved more effective?

Our findings confirm that different representations of information needs did indeed have a large impact on retrieval effectiveness. We find that humans were capable of formulating very effective queries but that there were large differences in effectiveness between clinicians. Specific clinician query strategies were found to

be significantly more effective. These strategies included: i) using query keywords from specific clinical tasks such as treatments or tests; ii) deriving new query keywords not mentioned in the verbose patient case descriptions; and iii) formulating short or long queries. In addition, we find that specific clinicians chose specific querying strategies, with some being more effective than others.

These findings help to uncover an understanding of how different representations of an information need and different query strategies impact retrieval. It also shows that certain people consistently adopt certain query strategies. Our purpose in revealing this is to inform the development of new models and methods that specifically focus on the query formulation process to improve retrieval effectiveness.

The Information Need — Searching for Clinical Trials

In this study, we focus on the specific information seeking task of searching clinical trials. Clinical trials are experiments conducted in the development of new medical treatments, drugs or devices. Recruiting patients for a trial is often a time-consuming and resource intensive effort, and imposes delays or even the cancellation of trials (@warning Citation 'Penberthy2012Effort-required' on page 2 undefined). Matching patients to clinical trials is essentially an information retrieval task: the query is a description of the patient (from verbose patient case descriptions to terse ad-hoc queries) and the documents are the clinical trials currently recruiting patients.

The task of searching clinical trials was chosen for a number of reasons. Firstly, it is an important real-world information seeking task in the medical domain that unpins the success of clinical trials, which are critical for the advancement of science and medicine. Secondly, and more specific to this study, it is common to have multiple, different representations of an information need (i.e., the patient), from verbose patient case descriptions through to terse ad-hoc queries. This allowed us to study how these different representations affected retrieval. Finally, there is higher task complexity in clinical search (@warning Citation 'Koopman2014Why-Assessing-R' on page 2 undefined) and research has shown that query variation is more significant with higher task complexity (@warning Citation 'Bailey2015User-variabilit' on page 2 undefined).

Related Work

Query variability is as big as system variability

In information retrieval (IR), a persons's information need is often represented as a keyword query. It is generally accepted that such a query is a significant simplification of an often complex information need (@warning Citation 'belkin1982ask' on page 2 undefined; @warning Citation 'vanRijsbergen79irbook' on page 2 undefined; @warning Citation 'ingwersen2005turn' on page 2 undefined). The choice different people make in how they formulate their information need may have a significant bearing on the effectiveness of their query — some queries and some people will be more effective than others (@warning Citation 'Bailey2015User-variabilit' on page 2 undefined).

There is evidence from initial studies showing that variability in queries had as much impact on retrieval effectiveness as variability in systems (@warning Citation 'Bailey2015User-variabilit' on page 2 undefined; @warning Citation 'Moffat2015Pooled-Evaluati' on page 2 undefined). Azzopardi (@warning Citation 'azzopardi2009query' on page 2 undefined) noted that the effectiveness of an IR system was strongly influenced by the query submitted; they further went on to quantify the likely effort involved in submitting effective queries. Bailey et al. (@warning Citation 'Bailey2015User-variabilit' on page 2 undefined) presented different people with the same information need (TREC topics) and solicited ad-hoc queries (on average 44 queries per topic). They compared the variation of query effectiveness for a topic with the variation in system effectiveness for all retrieval systems that participated in the TREC track. They found that query-derived variations were just as broad as system variations. Results obtained from the TREC 8 Query Track and CLEF 2015 eHealth Lab Task 2 confirmed these findings (@warning Citation 'Buckley1999The-TREC-8-Quer' on page 2 undefined; @warning Citation 'Palotti2015CLEF-eHealth-ev' on page 2 undefined). Therefore, improved performance was just as "likely to be derived from query reformulation as it is from system improvement" (@warning Citation 'Bailey2015User-variabilit' on page 2 undefined). On analysing the effectiveness of individual queries, Bailey et al. concluded that query formulation was critical to query effectiveness.

More complex search tasks were found to display even greater variability in query effectiveness (@warning Citation 'Bailey2015User-variabilit' on page 2 undefined). For complex tasks, query variability trumped system variability. Search within the clinical domain can be considered a complex task (@warning Citation 'Koopman2014Why-Assessing-R' on page 2 undefined), especially when dealing with clinical trials (@warning Citation 'Penberthy2012Effort-required' on page 2 undefined; @warning Citation 'Pressler2012Computational-c' on page 2 undefined). Therefore, a study of query variability is well situated within the clinical domain.

There are few resources that facilitate studies of multiple query variations. The TREC Query Track (@warning Citation 'Buckley1999The-TREC-8-Quer' on page 2 undefined) had some investigation of query variability for the same information need. For each of the 50 topics, teams provided one or more ad-hoc and sentence based query representations. The findings were: i) topics were extremely variable; ii) queries dealing with the same topic were extremely variable; iii) even short queries were rarely duplicated (16%); iv) systems were only somewhat variable. While the track did investigate multiple query variations, it was severely limited by a lack of data as only 5 teams participated and only a subset of them provided a small number of queries. The study of query variations was mainly limited to determining whether a query or topic was hard or easy in terms of effectiveness rather than determining the underlying reasons for this (@warning Citation 'Buckley1999The-TREC-8-Quer' on page 2 undefined). The organisers concluded that not enough data was available to draw strong conclusions and that the "experiment needs to be repeated." (@warning Citation 'Buckley1999The-TREC-8-Quer' on page 2 undefined). Our study aims to both repeat and address some of the shortcomings mentioned.

Other recent resources aimed at investigating query variations include the UQV100 collection (@warning Citation 'Bailey2016UQV100:-A-Test-' on page 2 undefined), which contained 5,764 query variations obtained for 100 topic backstories, and the CLEF 2015 eHealth Lab Task 2 collection (@warning Ci-

tation 'Palotti2015CLEF-eHealth-ev' on page 2 undefined), which contained 66 consumer health queries related to 23 medical conditions.

These previous studies on query variations were valuable in understanding the impact variations had on retrieval. However, they do not provide an insight into the characteristics of different queries according to their effectiveness. The question remains — what makes an effective query?

Different people use different query strategies

Studies in information seeking behaviour have shown that different people use different search strategies (@warning Citation 'Gwizdka2006What-Can-Search' on page 3 undefined; @warning Citation 'bates1979information' on page 3 undefined). Factors such as experience, verbal ability and other cognitive abilities, all influenced query effectiveness.

In addition, people have differing expectations with respect to the amount of information (e.g., the number of documents) they believe they need to complete their task (@warning Citation 'Bailey2015User-variabilit' on page 3 undefined). So much so that new evaluation measures (e.g., the INST measure (@warning Citation 'Moffat2015INST:-An-Adapti' on page 3 undefined)) were proposed to explicitly capture peoples' different expectations.

A significant amount of research has been dedicated to understanding how people seek information, their search tactics, and in modelling people's search behaviour. For example, Bates reported that one of the tactics people used when searching was to weigh up the costs and benefits of their interactions with the search system: formulate a query, examine a result, etc. (@warning Citation 'bates1979information' on page 3 undefined). While Bates did not elaborate on this tactic, subsequent research has expanded on this notion of accounting for the costs and benefits of interaction using different, but related, frameworks (@warning Citation 'russell1993sensemaking' on page 3 undefined; @warning Citation 'pirolli2007information' on page 3 undefined; @warning Citation 'fuhr2008iprp' on page 3 undefined; @warning Citation 'azzopardi2011economics' on page 3 undefined). For example, Russell et al. examined the cost structures associated with sensemaking (@warning Citation 'russell1993sensemaking' on page 3 undefined), while Pirolli adapted Foraging Theory to explore how searchers strive to maximise the gain of their search interactions over time (@warning Citation 'pirolli2007information' on page 3 undefined). Azzopardi's Economic Model for IR (@warning Citation 'azzopardi2011economics' on page 3 undefined; @warning Citation 'azzopardi2015analysis' on page 3 undefined; @warning Citation 'azzopardi2016a' on page 3 undefined) which synthesises the user behaviour as a function of benefit and cost of different interactions (querying, assessing, etc.). Much of this previous research focuses on modelling user interactions, often within a cost, benefit scenario. In this study, we are more concerned with different interaction strategies of individual users and how these differ.

We know that query formulation strongly influences effectiveness and we aim to answer what makes an effective clinical query. We also know that queriers use different query strategies. In this paper, we further investigate these different query strategies with an eye toward determining what makes an effective clinical querier.

Methods

Topic and query variation generation

Our definition of a topic was a single information need (in line with that used in TREC campaigns). However, for each topic we had multiple representations, which we will refer to as individual queries. The topics, in our case, represented a description of a patient for which relevant clinical trails were sought. We adopted the topics previously used by the TREC CDS track (@warning Citation 'Simpson2014Overview-of-the' on page 3 undefined), which comprised 60 patient case descriptions (30 from 2014 and 30 from 2015). Each topic described a patient with certain conditions and observations. Each patient case topic already had two types of query representations: a *description* (on average 78 words) and a shorter *summary* (on average 22 words). A sample topic, with description and summary is shown in Figure 1(a) and Figure 1(b), respectively. Note, that these patient case descriptions and summaries were similar to those that may be found in actual electronic medical records or at least as patient surrogates in clinician teaching material. As such, they were a realistic representation of the information need and, in the case of the task of matching an electronic patient record to a clinical trial, would constitute the actual query.

In addition to the description and summary we collected a number of ad-hoc queries; these were provided by clinicians. Our clinicians were four final year medical students who were employed to provide ad-hoc queries and to perform relevance assessments on documents. All four clinicians came from the same cohort of their medical degree and thus had received very similar training. They had completed the theory portion of their medical degree and were completely their hospital placements, gaining extensive exposure to clinical practise. Thus, they had similar (topical) expertise. Each clinician was shown the 60 topic descriptions and provided one or more ad-hoc queries after being asked "Please provide us with one or more queries you would enter when searching for eligible clinical trials for this patient". A screenshot showing an actual interaction by Clinician B with the interface used to collect queries for topic# 2014-3 is shown in Figure 2. Thus, for each of the 60 topics there was a description, a summary and a number of ad-hoc queries. (Figure 3 shows the ad-hoc queries for the sample topic of 1(a).)

In addition to providing queries, each clinician was also asked "How many clinical trials do you expect this patient would be eligible for?". This value was used to understand some of the differences between the clinicians' understanding of their information need. The value was also collected to be used within the INST evaluation measure (@warning Citation 'Moffat2015INST:-An-Adapti' on page 3 undefined) (described in detail later), which explicitly accounts for this expectation regarding the number of relevant results.

Documents and relevance assessments

A collection of 204,855 publicly available clinical trails were crawled from ClinicalTrials.gov. These form part of an IR test collection for searching clinical trails (@warning Citation 'Koopman2016A-Test-Collecti' on page 3 undefined).

Six baseline retrieval models were run to form the pool of document to be judged by clinicians. These models included: BM25, Language Model (Direchlet and Jelinek-Mercer), Divergence From Randomness (BB2 and DLH) and TF-IDF. (The Terrier IR system was used for all models and parameters left to Terrier de-

Topic# 2014-29
1. A 51-year-old woman is seen in clinic for advice on osteoporosis. She has a past medical history of significant hypertension and diet-controlled diabetes mellitus. She currently smokes 1 pack of cigarettes per day. She was documented by previous LH and FSH levels to be in menopause within the last year. She is concerned about breaking her hip as she gets older and is seeking advice on osteoporosis prevention.
(a) Description
2. 51-year-old smoker with hypertension and diabetes, in menopause, needs recommendations for preventing osteoporosis.
(b) Summary

Figure 1. Different description (a) and summary (b) representation of patient case for the same topic (information need).

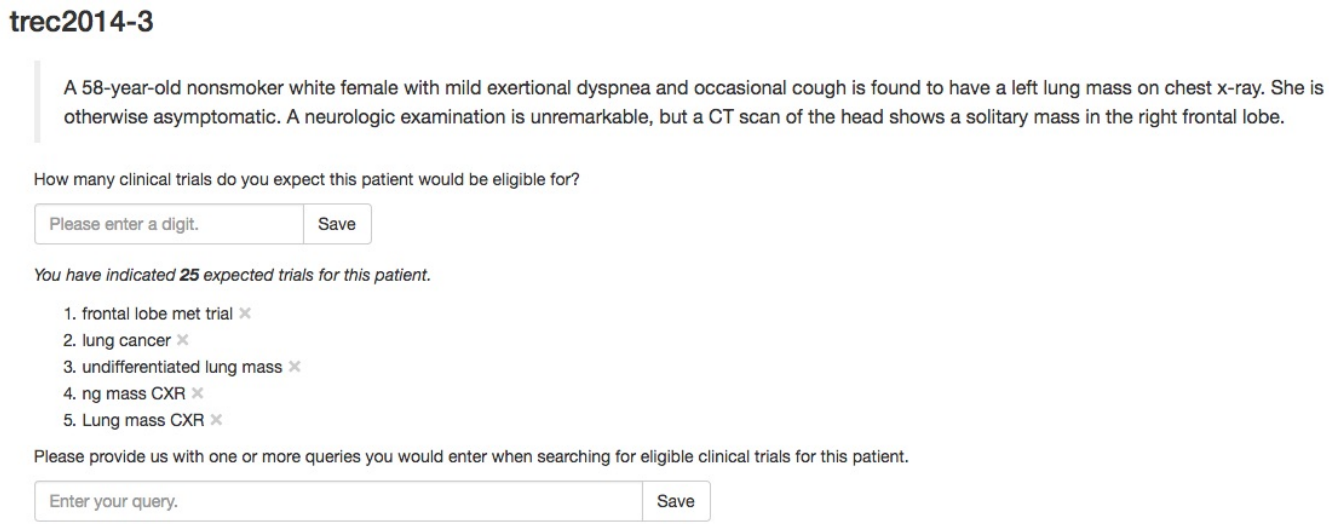


Figure 2. Screenshot of query collection interface taken from actual interaction by Clinician B with the interface used to collected queries for topic# 2014-3. The clinician has provided 5 different ad-hoc queries and indicated that they expect 25 clinical trials for which patient would be eligible.

Topic# 2014-29
3. peripheral arterial disease
4. cardiovascular disease
5. peripheral vascular disease and possible therapies to prevent ischaemic limb
6. calf Pain Excercise History of Myocardial infarct Hypertension polypharmacy
7. peripheral vascular disease trial
8. lower limb claudication trial
9. peripheral arterial disease trial

Figure 3. Multiple ad-hoc queries based on clinician's review of Fig 1(a).

faults (@warning Citation 'Macdonald2012From-puppy-to-m' on page 3 undefined).) While this was only a small number of systems, we note that Moffat et al. found that query variations were as strong as system variations in producing a diverse document pool (@warning Citation 'Moffat2015Pooled-Evaluati' on page 3 undefined); thus, we overcame the limit of having a small number of systems by including a large number of query variations. Including the description, the summary and the ad-hoc queries meant that there were, on average, 10.2 queries per topic. This equated to an average of 61 runs per topic (10.2 queries per topic * 6 baseline methods). This provided a diverse set of retrieved documents to form the pool.

To maximise the time and minimise costs associated with employing clinicians it was important to maximise the chance of sampling important documents for assessment. A standard approach to form the pool was to include all documents that were highly ranked by participating systems. However, Moffat et al. (@warning Citation 'Moffat2007Strategic-syste' on page 5 undefined) noted that not all documents provided the same benefit and they instead proposed an alternative method based on the Ranked Biased Precision (RBP) evaluation measure. Documents were ranked according to RBP across all queries; documents that were retrieved by multiple, different systems in top-ranked positions would appear higher in the RBP ranking. The pool was then formed based on the available assessment budget by setting a cut-off point of 4,000 documents in the RBP ranking — documents above the cut-off were included in the pool.[1] This pooling approach has been shown to minimise bias for fixed budget pooling (@warning Citation 'Lipani2016The-Impact-of-F' on page 5 undefined).

Documents and queries were uploaded to the Relevation! relevance assessment system (@warning Citation 'Koopman2014Relevation:-An-' on page 5 undefined) and the four clinicians who provided queries also provided the relevance assessments, according to a three-point scale:

0: *Would not refer this patient for this clinical trial*;
1: *Would consider referring this patient to this clinical trial upon further investigation*; and
2: *Highly likely to refer this patient for this clinical trial*.

Queries were divided amongst the four clinicians; a control query (topic #20158) was used to familiarise clinicians with the task and to record inter-coder reliability, which was found to be 70%. This highlights the difficulty intrinsic in judging relevance in the medical domain, as identified by other studies (@warning Citation 'Koopman2014Why-Assessing-R' on page 5 undefined; @warning Citation 'palotti2016a' on page 5 undefined). Reasons for clinicians disagreement will be investigated in future work.

The task and evaluation measures

The task of matching patients to clinical trials has three specific use cases; we use these to set the evaluation measures for the task.

The first use case is in a General Practitioner (GP) setting where the GP opens a patient's record as part of a consultation and a search is automatically initiated to find relevant clinical trials that the GP may refer the patient to. In this scenario the GP is time-pressured and would likely only review a small number of results, stopping when a single relevant trial is found. Thus for this scenario we adopted Reciprocal Rank as the evaluation measure.

The second use case is also set within a general medical professional (GP or other) but where the clinician is specifically searching for clinical trials and may dedicate more time and effort to the task. In this case they may issue an ad-hoc query themselves and be willing to evaluate a few more results. For this scenario we adopted Precision at 5 (P@5) as the evaluation measure.

The final use case is for medical specialists or patients themselves searching for trials. Here both clinician and patients may conduct longer search sessions and review far more results. They may use both short ad-hoc queries and more verbose patient case description. In addition, both clinician and patients would have an expectation about how many clinical trials the patient would be eligible for. This would influence their search behaviour: for rare diseases, they may expect to find a very small number of trials and would therefore not persist in examining results at greater rank depths. In contrast, for common diseases, they would expect to find many relevant trials and would therefore persist to greater rank depths. This notion of expected number of (relevant) results is directly modelled by T in the INST evaluation measure (@warning Citation 'Moffat2015INST:-An-Adapti' on page 5 undefined); thus we adopted INST for this scenario.

INST is a weighted precision metric where the likelihood of the user assessing a document at a specific rank depends on i) the rank position; ii) the expected number of relevant documents; and iii) the actual number of relevant documents encountered up to that rank. According to INST, the expected depth at which the user would stop viewing documents falls between approximately $T + 0.25$ (all encountered documents are relevant) and $2T + 0.5$ (no encountered documents are relevant) (@warning Citation 'Moffat2015INST:-An-Adapti' on page 5 undefined). We use INST because it explicitly accounts for the differences in peoples' information seeking behaviour.

For the binary relevance measures of P@5 and Reciprocal Rank, a relevance label of 0 was considered not relevant and a relevance label of 1 or 2 as relevant. INST is a graded relevance measure that accounts for the difference in gain between a label of 1 and 2.

Statically significance was determined by a unpaired two-tailed t-test with p values reported.

Results and Analysis

What makes a good clinical *query*?

How different query representations affected retrieval?

The retrieval results, according to the different query representations (description, summary and ad-hoc), and divided by each baseline retrieval model, are shown in Figure 4. (Note that there were on average 8.2 ad-hoc query runs per topic-system pair; therefore, we averaged the effectiveness of a system over all ad-hoc queries.) We firstly observe that there was high variability of performance across the different query representations. The clinician-provided ad-hoc queries proved most effective overall, followed by summary and, finally, the patient case description was the least effective form of query representation, although only ad-hoc queries were found to actually be statistically significantly different.

There was also variability across different baseline retrieval models. The best method for ad-hoc queries was the Jelinek-Mercer language model and DRF-DLH. However, these were only found to be statistically significant (t-test, $p < 0.05$) for INST and P@5 and not against every other model. For the longer summaries and descriptions the best method varied with no clear stand out method. This observation suggests that different baseline models were better suited to different query representations.

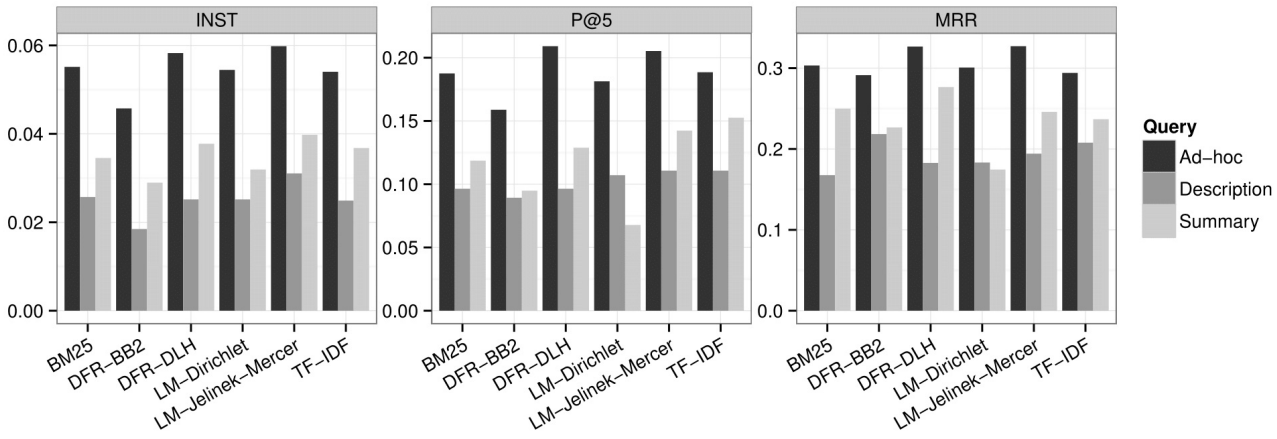


Figure 4. Retrieval results for different baselines and query representations.

Overall, we note that there was more variability across different query representations than across (baseline models) systems. Statistical testing also showed that the differences in query representations were significant (t-test, $p < 0.05$) but the differences in retrieval model was only significant under certain conditions. Although high query variation was found in other studies (@warning Citation 'Moffat2015Pooled-Evaluati' on page 5 undefined), these were considering different ad-hoc queries for the same information need; we, instead considered different representations (description, summary and ad-hoc) for the same information need. Nevertheless, the conclusion from both variations in representations and variations in ad-hoc queries (@warning Citation 'Moffat2015Pooled-Evaluati' on page 5 undefined) was the same — that how a query was expressed can have a larger impact on retrieval than what system was used to serve it.

The variation in effectiveness for different ad-hoc queries.

In the previous section, the effectiveness of ad-hoc queries for a topic was calculated as the averaged across all the individual ad-hoc queries. Here we consider the effectiveness of individual queries and how these compared with the effectiveness of the summary and description query representations. Figure 5 shows the effectiveness of the various individual queries associated with each topic on the LM-Jelinek-Mercer baseline. (Other baseline models displayed similar results.)

For a given topic there was considerable variation in the effectiveness of individual query representations. For the majority of topics, the ad-hoc queries were the most effective (58/60 topics for INST, 48/60 topics for P@5 and 53/60 topics for reciprocal rank). However, even within these there was still considerable variation in effectiveness, showing that clinicians formulated both highly effective and ineffective queries for the same information need. Although there was significant variation, ad-hoc queries were still found to be statistically significantly better than description and summary queries (t-test, $p < 0.05$ for all three evaluation measures); no statistically significant differences were found between summary and description queries.

The next section considers in more detail how queries were formulated and the effect that this had on retrieval effectiveness. In summary, a good clinical query tends to be a short, ad-hoc query.

What makes a good clinical *querier*?

Number and length of queries entered.

The number of ad-hoc queries provided by each clinician is shown in Figure 6. The number of queries varied both per topic and per clinician. The average number of queries per topic was 2.08 (sd=1.43) and the maximum of queries for single topic was 11 (Clinician B). On certain topics all clinicians tended to enter more queries, while on other topics clinicians only entered a single query. Figure 7(a) shows how the amount of queries entered per topic differed between clinicians. Clinician A and C both entered fewer queries per topic (A 1.59 and B 1.37), while clinician B and D entered a more queries per topic (B 2.54 and D 2.81).

The length of the queries (in terms of number of keywords) also varied. Figure 7(b) shows a histogram of the number of queries according to different query lengths. Some clinicians entered multiple short queries, while others preferred single, longer queries. In fact, the plot shows two distinct approaches to query formulation: an approach based on many short queries (clinicians B and D with 2.8 and 3.5 terms, respectively) and an approach which adopted more verbose queries (clinicians A and C with 5.1 and 6.6 terms, respectively).

There is often a trade off between the length of the query and the number of queries a person formulates (@warning Citation 'azzopardi2011economics' on page 6 undefined; @warning Citation 'azzopardi2013query' on page 6 undefined). Again, two distinct approaches to query formulation are apparent: an approach based on a smaller amount of longer queries (clinicians A and C) and another based on a larger amount of shorter queries (clinician B and D). Next, we consider which of these query formulation approaches was more effective and why.

How did query length affect retrieval effectiveness?

Figure 8 shows the effect of query length on retrieval effectiveness. (The LM-Jelinek-Mercer is shown but other retrieval methods exhibited similar results.) First, considering just query length, and leaving aside the individual differences between clinicians, we observe that longer queries were less effective (as shown by the far right, "(all)" column of the plot). Second, when considering individual clinicians, we observe differences in both query length and effectiveness. Clinician B and D both formulated shorter queries

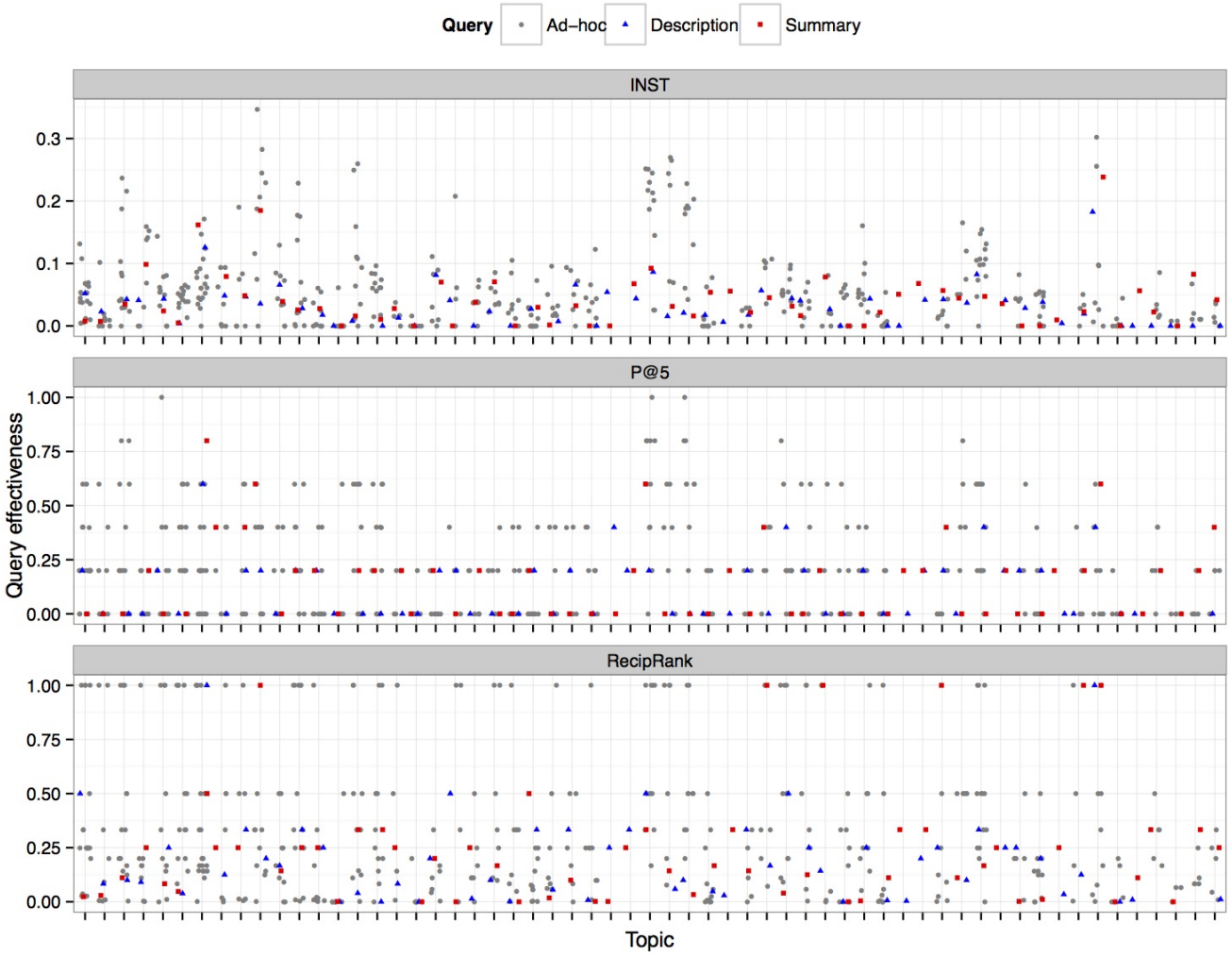


Figure 5. Retrieval effectiveness results for individual queries per-topic on the LM-Jelinek-Mercer baseline.

that proved more effective. Clinician C, in particular, formulated much longer queries that proved statistically significantly less effective (t-test, $p = 2.2^{-16}$) than the other clinicians.

Expected number of results, T.

Clinicians were also asked how many clinical trials they expected a patient would be eligible for. This was represented as T in the INST evaluation measure. The values of T for each topic, across the four clinicians, is shown in Figure 9. Values of T varied across topics, thus indicating the different information needs clinicians derived from different patients. Although T varied across topics, individual clinicians displayed similar trends across topics; e.g., clinician D typically chose lower values of T (mean=3.2, SD=1.4) and clinician C displayed higher values of T (mean 35.0, SD=19.3). This resulted in values of T that varied across clinicians for a single topic. Qualitative feedback from clinicians indicated estimating T was challenging and subjective. The clinicians were asked about their rationale for determining values of T. We found that regardless of the value of T, clinicians indicated that the main rationale was how rare or common the patient's medical condition was; secondary to that was the likelihood that clinical trails were currently being conducted on the patient's condition.

Did clinicians choose keywords from the description?

When entering ad-hoc queries, clinicians were shown the patient case descriptions; this is akin to the task clinicians regularly perform of reviewing a patient's chart. After reading the description clinicians were asked to provide keyword queries. Some queries contained keywords from the description, while others contain novel keywords. Here we consider the overlap of keywords in the clinician's ad-hoc query and corresponding description in order to understand better how clinicians formulated their queries and the differing strategies between clinicians.

The overlap of a query Q is defined as the portion of keywords in an ad-hoc query that were contained in its description, D:

$$overlap(D, Q) = \frac{|D \cap Q|}{|Q|}.$$

Figure 10 shows a histogram of the number of queries with different overlaps. Considering all clinicians (righthand plot), there was a large number of queries for which clinicians chose their own

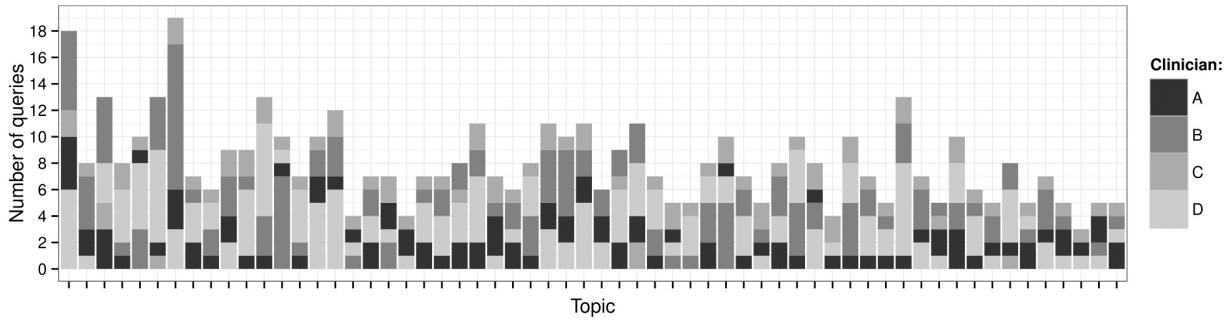


Figure 6. Number of queries supplied for each topic.

keywords and did not choose any keywords from the description. However, this overlap varied considerably between clinicians. Clinician D consistently selected their own query keywords that did not appear in the description, with 74% (125 out of 169) queries containing no keywords in common with the description. Clinician A also often selected their own keywords not found in the description. In contrast, clinician C always chose keywords from the description. Clinician B formulated a mix of different queries, some with high overlap but others with no overlap.

The results show a spectrum of clinical querier: from those who consistently re-use query keywords to those who consistently inferred novel query keywords.

How did overlap affect retrieval effectiveness?

The effect of query overlap on retrieval effectiveness is shown in Figure 11. Each point represents an ad-hoc query. Clinicians A and D both tended to express ad-hoc queries with lower overlap with the description (no statistically significant difference in overlap $p = 0.62$). Interestingly, these low overlap queries of A and D proved to be more effective (average Pearson's correlation coefficient between overlap and retrieval effectiveness was -0.31). Indeed, the most effective queries for these two assessors were those with no overlap (queries with no overlap were 13% better in reciprocal rank, 40% better in INST and 27% better in P@5). Clinician C, with the highest overlap was statistically significantly less effective than the other clinicians. Clinicians A and B had statistically significant different ($p < 0.001$) strategies in terms of overlap (clinician A with low overlap and clinician B with high overlap), however, they had similar overall effectiveness (no statistically significant difference in effectiveness: $p > 0.05$ for all three evaluation measures). In general, lower overlap was weakly correlated with higher query effectiveness (Pearson's correlation coefficient -0.21).

To provide a concrete example of the type of queries that clinicians generated and the overlap with the description, we provide two sample topics in Appendix A. These show examples of where clinicians inferred novel query keywords that resulted in very effective queries (when compared to the description and summary alone).

What clinical tasks types did clinicians choose?

Previous studies (@warning Citation 'Ely2000A-taxonomy-of-g' on page 8 undefined) have shown that clinicians pose queries within three common clinical tasks: i) searching for *diagnoses* given a list

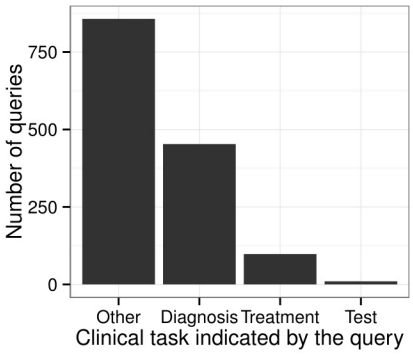
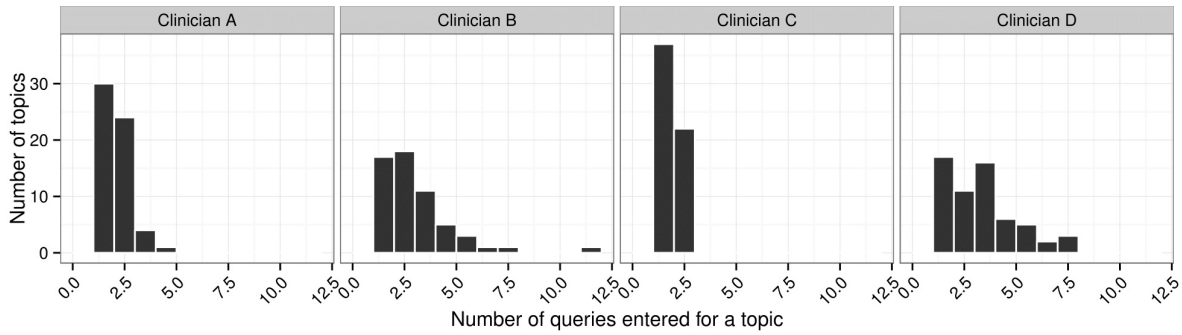


Figure 12. Number of query posed according to the clinical task of query keywords.

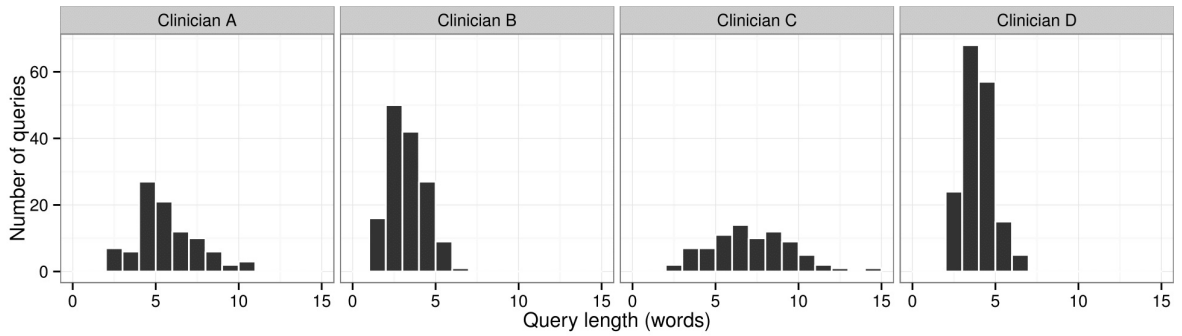
of symptoms; ii) searching for relevant *tests* given a patient's situation; and iii) searching for the most effective *treatments* given a particular condition. These three diagnoses, test, treatment tasks were used in the TREC Clinical Decision Track to describe individual topics (@warning Citation 'Simpson2014Overview-of-the' on page 8 undefined). In our study, it was also possible to identify the particular clinical task. After reading the patient case descriptions, clinicians chose ad-hoc keywords — these keywords can be mapped to one of the three clinical tasks. In this section, we investigate which clinical tasks clinicians used most and the influence that this had on retrieval effectiveness.

Each ad-hoc query was analysed to identify medical concepts belonging to the UMLS Metathesaurus (this was done using the Metamap concept extraction system). Within the UMLS Metathesaurus, each medical concept has an overarching semantic type (e.g., the concept "Headache" belongs to the semantic type "Sign or Symptom"). These semantic types could then be mapped to the clinical tasks diagnosis, treatment or test by consulting the i2b2 challenge guidelines which defined a mapping between UMLS semantic types and clinical tasks (@warning Citation 'Uzuner20112010-i2b2/VA-ch' on page 8 undefined). Using the aforementioned method, each ad-hoc query may have contained a number of concepts belonging to the three clinical tasks, with an "other" task added to capture those concepts that belonged to none.

The frequency of use for the different clinical tasks is shown in Figure 12. Clinicians often entered keywords pertaining to diagnoses, i.e., the known conditions about a patient. This was more or less to be expected as many clinical trials documents stated certain conditions in the inclusion eligibility criteria. Clinicians searched



(a) Distribution of topic according to the number of queries entered per assessor. The average number of queries for each clinician was A 1.59, B 2.54, C 1.37, D 2.81.



(b) Number of queries according to query length per clinician. The average query length was 4.5 words, sd=2.5 words. The average query length for each clinician was A 5.1, B 2.8, C 6.6, D 3.5.

Figure 7. Number of queries per topic and number of queries according to query length

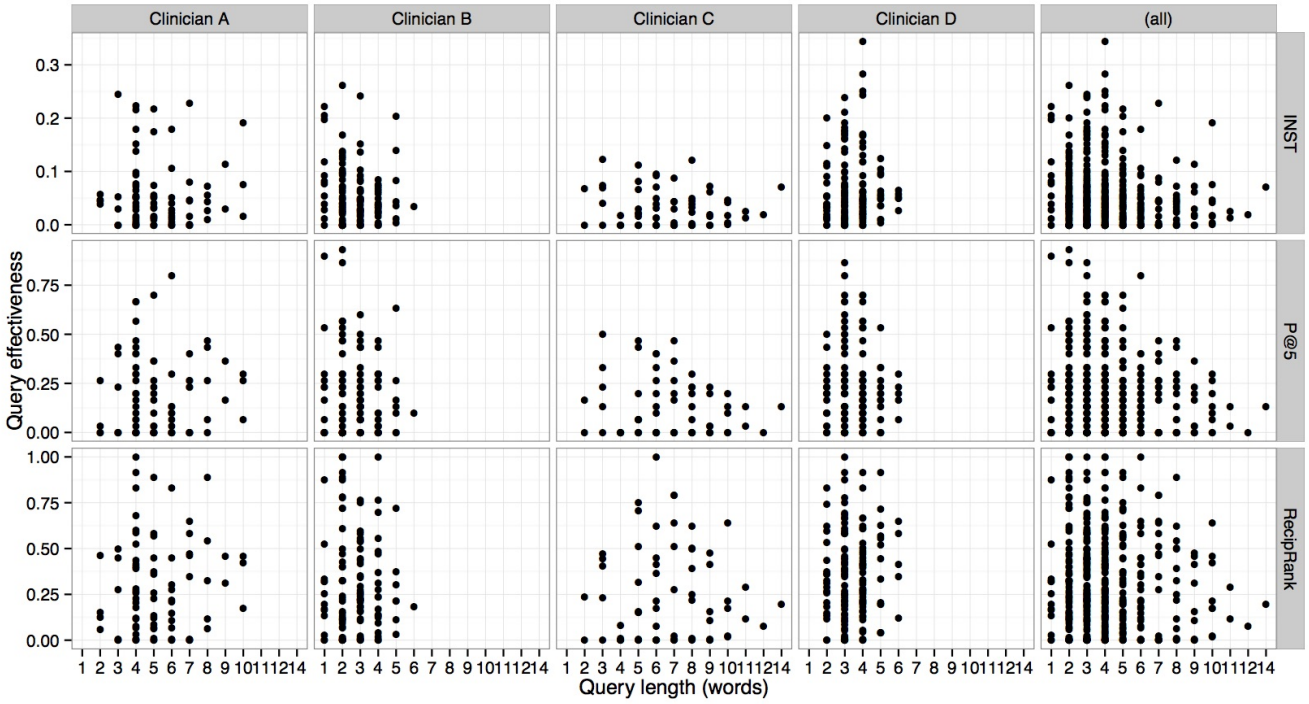
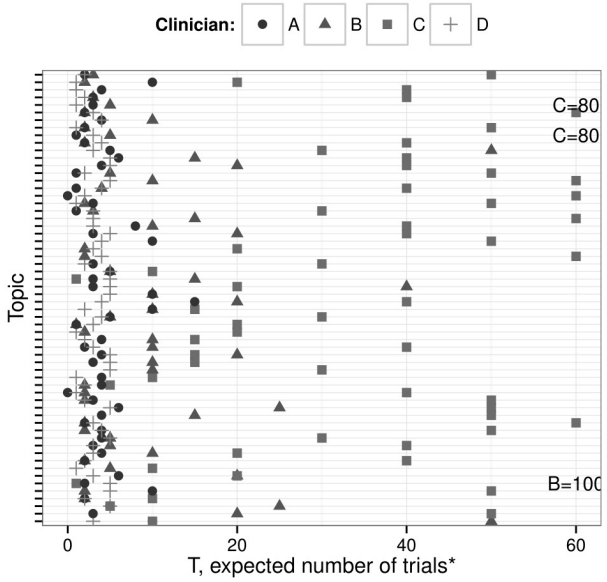


Figure 8. Query length vs. query effectiveness.



*x-axis truncated at $T = 60$ excluding outliners $T = 80, 80, 100$.

Clinician	mean T	SD T
A	4.2	3.1
B	11.9	16.5
C	35.0	19.3
D	3.2	1.4

Figure 9. T, the clinicians' expected number of clinical trials for a patient topic.

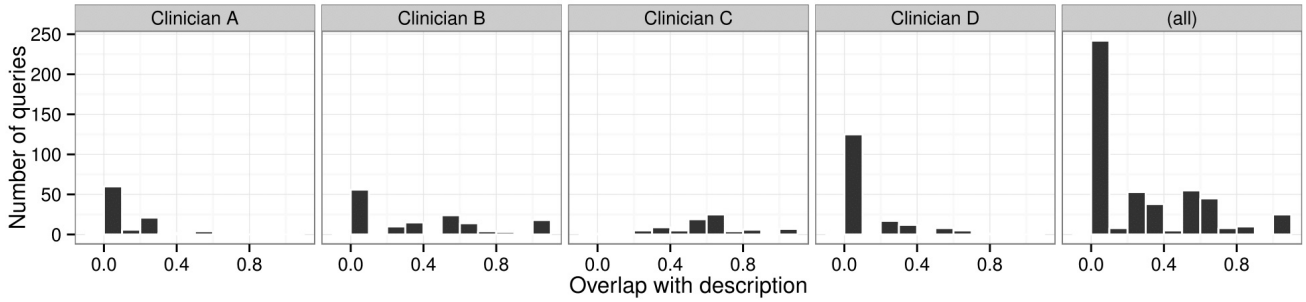


Figure 10. Distribution of overlaps of ad-hoc query with description. A large number of queries contained terms that did not appear in the description. This indicates that clinicians chose to formulate their own query terms rather than select those from the patient description.

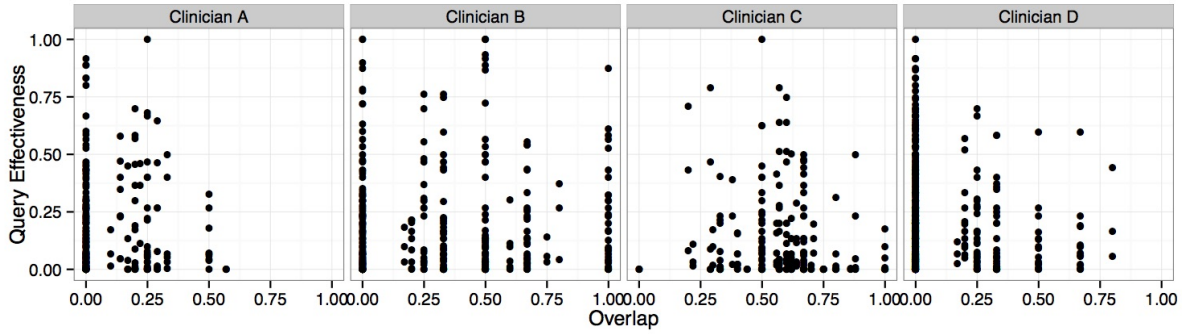


Figure 11. The effect of query lap on retrieval effectiveness (LM-Jelinek-Mercer model). Each point represents an ad-hoc query.

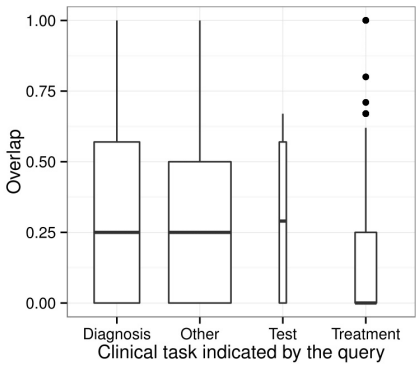


Figure 14. Overlap of ad-hoc query keywords with patient case description according to different task types. Treatment tasks has very little overlap, indicating that clinicians inferred relevant treatments not mentioned in the patient case description.

according to treatment tasks with less frequency than diagnosis tasks. The most common treatments related to specific medications (e.g., drug names) followed by specific medical procedures (e.g., surgical procedures). Clinicians only issued a small number of queries pertaining to test tasks. The most common tests were medical imaging based tests; e.g., MRI, X-ray or CT scans. Finally, a large number of query keywords fell within 'other', e.g., common examples of such queries involve medical concepts relating to body parts or clinical findings. Individual clinicians all displayed a similar trend relating to the frequency of use of the different task types when issuing queries.

How did the clinical task affect overlap and retrieval effectiveness?

Figure 13 shows the variation in query effectiveness for each clinical task. Queries in the 'other' type proved the least effective. Queries pertaining to diagnosis tasks turned out to be only marginally more effective. In turn, queries pertaining to treatment tasks proved more effective than queries pertaining to diagnosis tasks. Finally, queries containing keywords relating to tests exhibited the highest level of effectiveness. Note, however, due to the small number of such queries, the statistical significance of this superiority could not be established.

How task type influenced overlap — that is whether the clinician chose keywords from the description or inferred their own — is shown in Figure 14. Clinicians entering keywords pertaining to diagnoses and 'other' tasks often chose keywords from the description (median overlap = 0.25). In contrast, when clinicians entered keywords pertaining to treatment tasks, these were nearly always keywords not present in description (median overlap = 0.00). This result suggests that clinicians read the patient descriptions and then inferred the treatments they thought might be mentioned in clinical trials. This strategy of inferring novel keywords pertaining to treatments turned out to be effective (Figure 13) when compared to using keywords pertaining to diagnosis or 'other' tasks.

In summary, a good clinical querier is one who: issues short queries; considers specific keywords that do not appear in the patient description (i.e., low overlap); and infers related treatment (not diagnosis) terms that were not mentioned in the patient description, but may have appeared in the clinical trial.

Discussion

Multiple representations of an information need

In IR, a person's complex information need is often represented as a single keyword query, even though it is generally accepted that such a query is a significant simplification (@warning Citation 'belkin1982ask' on page 11 undefined; @warning Citation 'vanRijsbergen79irbook' on page 11 undefined; @warning Citation 'ingwersen2005turn' on page 11 undefined). Nonetheless, the common paradigm in IR evaluation is still to use test collections comprising single keyword queries to represent the information need and relevance assessments indicating the relevance of single query, document pairs. The focus is, therefore, on comparing how different systems perform on these fixed queries, that is, the focus tends to rest on system variation without considering query variation.

A consequence of focusing on system variation is that substantial research effort is spent on the system side (e.g., understanding and developing retrieval models) rather than the query side (e.g., understanding query formulation and models to improve this).

This study aims to focus squarely on how users express their information need and how that manifests as different query representations. The results confirm that different query representations of the same information need can have a large impact on retrieval effectiveness. Other studies on query variations have also found greater variability in query effectiveness than in variability across the effectiveness of systems (@warning Citation 'Buckley1999The-TREC-8-Quer' on page 11 undefined; @warning Citation 'Bailey2015User-variabilit' on page 11 undefined; @warning Citation 'Moffat2015Pooled-Evaluati' on page 11 undefined). In addition, the choices the querier makes formulating their information need will have a large impact on the effectiveness of the given query, especially in complex information seeking tasks (@warning Citation 'Bailey2015User-variabilit' on page 11 undefined) such as clinical search (@warning Citation 'Koopman2014Why-Assessing-R' on page 11 undefined). In this regard, we found via their ad-hoc queries, that it was the clinicians who largely impacted retrieval effectiveness, rather than the underlying retrieval model. However, within these ad-hoc queries there was a spectrum of more and less effective queries, and that this spectrum varied across the different clinicians.

What makes a good clinical querier?

Some key query formulation patterns emerged that indicated an effective clinical querier:

- A trade-off was observed between the query length and the number of queries posed per topic. Clinicians either posed a small number of long queries or a large number of short queries. The most effective clinicians were those who entered short queries. One caveat to this finding should be stated, namely, most retrieval models (including those used in this study) are optimised for shorter queries.
- Effective queriers tended to infer their own keywords rather than simply re-using those from the patient case descriptions. These novel keywords were more likely to occur in relevant documents and thus improved retrieval effectiveness. There was large variation in the extent to which clinicians inferred novel keywords, with those doing so being far more effective than those who did not.
- There are three common clinical task types — searching for diagnoses, searching for treatments and searching for tests. While diagnoses were used most frequently and tests least frequently, the

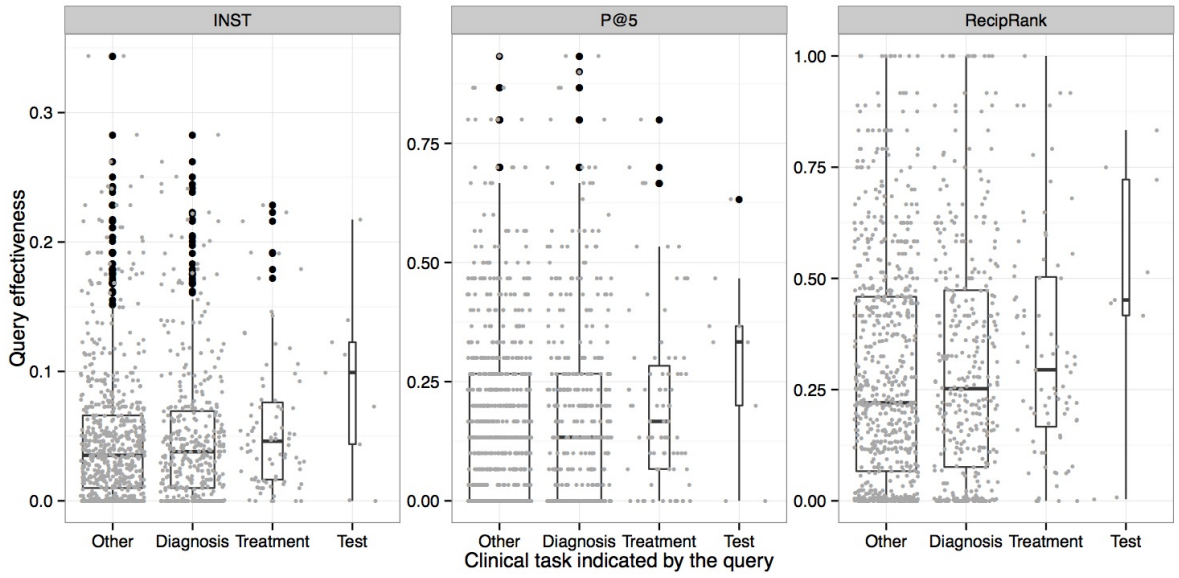


Figure 13. Effectiveness (LM-Jelinek-Mercer model) of different queries according to the clinical task of the keywords used in the query. Small, points represent the performance of an individual query and boxplots represent the distribution of effectiveness across queries. Queries containing tests proved most effective; followed by queries containing treatments and then queries containing diagnoses; queries containing keywords in the other category were least effective.

most effective queriers were those who posed queries around treatments more than diagnoses. Effective queriers were found to infer treatments not mentioned in, but based on, the patient case description; these treatments were, however, found in relevant clinical trials.

Limitations

Un-judged documents.

The dataset used to analyse the query behaviour of the clinicians comprised 4,000 relevance judgements. Documents to be judged were selected according to a pooling strategy based on RBP (@warning Citation 'Moffat2007Strategic-syste' on page 12 undefined). Because of the pooling strategy, there is no guarantee that measures like P@5 and reciprocal rank were computed based on complete assessments. For example, about 40% of the documents that contributed to the computation of P@5 values reported in Figure 5 were un-judged. For evaluation, un-judged documents were considered not relevant, as commonly done in information retrieval. This implies that we may under-estimate the effectiveness of a run due to the presence of un-judged documents (assumed to be not relevant, although they may actually have been relevant). There is no commonly accepted method to address this possibility. Specific evaluation measures have been developed to evaluate systems in presence of incomplete assessments, e.g., (@warning Citation 'yilmaz2008simple' on page 12 undefined). However, these methods are not applicable to our work because they make assumptions regarding sampling and distributions of assessments that are not valid for our collection. Others have reported residuals of an evaluation measure, e.g., RBP and INST (@warning Citation 'moffat2008rank' on page 12 undefined; @warning Citation 'Moffat2015INST:-An-Adapti' on page 12 undefined). Residuals attempt to estimate uncertainty intervals when faced with un-judged documents, providing an indication of the upper (all un-judged documents are considered relevant) and lower bound (all un-judged documents are considered not relevant) of the system effectiveness. However, the fact that one may measure large residuals may be misleading, as the actual chances that *all* un-judged documents are relevant may be in practice rather low. Recent work has investigated reducing the bounds on the residuals (@warning Citation 'park2016uncertainty' on page 12 undefined), but this work is still at early stages and the reporting of residuals is not a widely accepted practice in information retrieval. If we assume that, in our collection, un-judged documents had the same probability of being relevant as judged documents (a conservative assumption given the pooling strategy used here), then residuals may be as big as 11% for P@5.[2] Nevertheless, note that the pooling strategy we used to build our test collection explicitly attempts to maximise the chance that an un-judged document appearing in the top ranking is in fact not relevant, thus minimising the residuals. Recent work by Lipani et al. (@warning Citation 'Lipani2016The-Impact-of-F' on page 12 undefined; @warning Citation 'lipani2017fixed' on page 12 undefined) has further found that the pooling strategy used here minimised pool bias, thus attempting to form pools that are fair for the evaluation of all systems.

Limited number of subjects.

Our user study considered the search behaviour of four users. Although this is a small user sample to be able to generalise our results, we note that the users were sampled from a tight cohort of experts. Furthermore, we highlight the difficulty in getting hold of a large amount of time-poor clinicians, as well as the high costs involved. We also note that initial TREC guidelines for the Interactive Track called for a setting consistent with ours, with a minimum of four users per system (see (@warning Citation 'swan1998aspect' on page 12 undefined)), although later this has been revised to 16 users per conditions in a need for additional rigour and consistency in results and findings (@warning Citation 'dumais2005trec' on

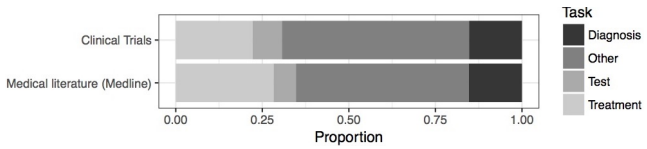


Figure 15. A comparison of clinical task types between the clinical trials collection used in this study and general medical literature documents.

page 12 undefined; @warning Citation 'julien2008controlled' on page 12 undefined).

Artefacts of the Clinical Trials Collection

Clinical querying, in this study, involved the task of searching for clinical trials given a patient case descriptions. Clinical trials documents are a specific type of document which may have some unique characteristics. In particular, the finding that queries with novel treatment terms were most effective could be an artefact of collection; i.e., clinical trial documents generally contain a large number of treatment terms, thus biasing treatment related queries. To verify (or refute) this hypothesis, we compared the clinical trials collection with a collection of medical journal articles (taken from the TREC Clinical Decision Support collection). Specifically, we annotated all documents in both collections according the four task types — diagnosis, test, treatment and other. The results are shown below in Figure 15. Clinical trials actually have a lower proportion of treatment related terms than medical literature. Clinical trials also have a slightly higher proportion of test related terms. However, both these differences were not found to be statistically significant. The purpose of clinical trials is indeed to verify treatments; however, the actual content of a clinical trial document is not primarily describing a treatment — while the treatment(s) is certainly mentioned, much of the focus is on describing the methodology of the trial and the eligibility criteria for participating patients. Thus the clinical trials collection is not biased toward treatments with respect to general medical documents.

Implications and further studies

From this study a number of different query strategies were observed — some more effective than others. If clear query-strategy patterns exist, then it may be possible to identify these automatically. It would then be possible to identify when a person may be using a sub-optimal strategy. In such cases, certain system interventions, such as query suggestion to decrease overlap, or query summarisation to shorten queries, may be employed in order to steer the person toward a more optimal strategy. An investigation on query strategy prediction is left to future work. However, we have completed an initial study that applied a machine learning approach to identify the features of an effective query and thus how such a query could be automatically generated (@warning Citation 'Koopman2017Generating-Clin' on page 13 undefined). Preliminary results using standard methods from the literature showed that more work is needed to develop automatic methods that approach the effectiveness of the human ad-hoc queries.

Although the pool of clinicians was small, we did observe that clinicians consistently used a single strategy (e.g., clinician D chose short, low overlap queries, whereas clinician C chose long, high overlap queries). It is, however, intriguing to speculate why it is that a clinician may consistently employ a single strategy and not others, and why do they adopt this strategy in the first place? Investigation into the reasons behind a person's query strategy could be the starting point for further, more psychologically inspired, studies in this area.

Although set within the clinical domain, this study has implications more generally applicable to the field of information retrieval. It confirms previous studies indicating that for the same information need, people formulate a large variety of queries — both in terms of length of query, keywords chosen and importantly, the effectiveness of the query. However, within this large variety of queries, people tend to consistently adopt specific query strategy patterns. We posit that these patterns could be leveraged to create more realistic user models and simulations for the evaluation of IR systems. Finally, our results highlighted that users often elicit implicit expert knowledge by reasoning about the information need and inferring novel keywords. (This was seen when clinicians inferred relevant treatments found in clinical trials but not mentioned in the patient description.) This human process of inferring relevant query terms warrants retrieval techniques that try to do the same by exploiting semantic inference processes to discover query expansions (@warning Citation 'Zhou2007Knowledge-inten' on page 13 undefined; @warning Citation 'Limsopatham2013A-Task-Specific' on page 13 undefined; @warning Citation 'Koopman2015Information-Ret' on page 13 undefined).

Conclusion

This paper provides an in-depth study into how clinicians formulate queries to represent an information need. Multiple representations of an information need — from verbose patient case description to short, ad-hoc queries — were considered in order to understand their effect on retrieval. The way a query was formulated impacted retrieval effectiveness far more than the particular retrieval systems used. The most effective queries were short and ad-hoc.

Observations of four different clinicians highlight that people consistently adopt specific query formulation strategies. Future work may be directed toward automatically identify these different query strategies and their use in modelling and in simulation for IR evaluation. The underlying reasons for why people choose specific strategies is also an interesting line of investigation.

The most effective queriers were those who, given their information need, inferred novel keywords most likely to appear in relevant documents. This suggests that inference-based retrieval methods are required to bridge the gap in how systems and humans formulate queries.

In general, this study aims to provide a deeper understanding of how clinicians search, with an eye for the development of new models and methods that specifically focus on the query formulation process to improve retrieval effectiveness.

The data for the clinical trials test collection, including all the query variations is available at:
`http://doi.org/10.4225/08/58e2e83d92c2b`.

Notes

[1]The persistence parameter, p, for RBP was set to 0.8 following the findings of (@warning Citation 'Zhang2010Click-based-evi' on page 13 undefined).

[2]29% of judged documents were found to be relevant, with 40% of

Topic# 2014-23 Description:
A 63-year-old man presents with cough and shortness of breath. His past medical history is notable for heavy smoking, spinal stenosis, diabetes, hypothyroidism and mild psoriasis. He also has a family history of early onset dementia. His symptoms began about a week prior to his admission, with productive cough, purulent sputum and difficulty breathing, requiring him to use his home oxygen for the past 24 hours. He denies fever. On examination he is cyanotic, tachypneic, with a barrel shaped chest and diffuse rales over his lungs. A chest x-ray is notable for hyperinflation with no consolidation.

Clinician	Query keywords	Overlap
A	Viral infective exacerbation of COPD	0.00
A	Treating multiple diseases at once in the setting of an infective exacerbation	0.00
B	COPD smoking	0.50
B	acute exacerbation of COPD	0.00
C	Clinical Trial Cough Smoker Diabetes X-ray hyperinflation	0.57
D	COPD exacerbation trial	0.00
D	COPD infective exacerbation trial	0.00
D	COPD antibiotics trial	0.00
D	**COPD corticosteroids trial**	0.00

Table A1. Sample topic showing different queries generated by four different clinicians. Overlap indicated the proportion of query keywords shared with the above description. The most effective query shown in bold typeface.

pooled document found to be un-judged; therefore, $29\% * 40\% = 11\%$. Similar values were found for the other evaluation measures.

A Sample topics with associated clinician queries

In this section, we provide two sample topics with topic descriptions and all the ad-hoc queries provided by the four different clinicians. These topic demonstrate cases where clinicians inferred novel query keywords that resulted in very effective queries (when compared to the description and summary alone).

Table A1 provides a sample topic# 2014-23 with queries generated by different clinicians according to the provided description. Many of the queries exhibit low overlap: they do not use keyword taken from the shown description. All four clinicians inferred, from the patient description, that the patient suffered from COPD: a group of lung diseases that block airflow and make it difficult to breathe. Many relevant clinical trials contained COPD and thus these queries were effective. The most effective query ("COPD corticosteroids trial") inferred the steroid hormone corticosteroids. The inclusion of this keyword helped retrieve those clinical trials focused on pulmonary diseases, rather than the large number trials that mentioned COPD but were not the focus of the trial.

Table A2 provides a further example where clinicians inferred novel keywords. Clinician C generated the most effective query by inferring the correct diagnosis (endometriosis) from the patient description, which contained a large number of symptoms and some past diagnoses (ectopic pregnancy) that may be misleading for the current query.

Topic# 2015-10 Description:
A 38 year old woman complains of severe premenstrual and menstrual pelvic pain, heavy, irregular periods and occasional spotting between periods. Past medical history remarkable for two years of infertility treatment and an ectopic pregnancy at age 26.

Clinician	Query keywords	Overlap
A	Early onset menopause	0.00
B	**Endometriosis middle aged female**	0.00
C	Premenstrual menstrual pelvic pain	1.00
C	Menstruation severe pain irregular spotting	0.80
D	fibroids clinical trial	0.00

Table A2. Sample topic showing different queries generated by four different clinicians. Overlap indicated the proportion of query keywords shared with the above description. The most effective query shown in bold typeface.