SPi	Journal Code	Article ID	Dispatch: 22-JUL-20	CE:
	ASI	24398	No. of Pages: 15	ME:

Received: 10 February 2020 | Revised: 4 June 2020 | Accepted: 7 July 2020

DOI: 10.1002/asi.24398

RESEARCH ARTICLE

JASIST WILEY

Do better search engines really equate to better clinical decisions? If not, why not?

Anton van der Vegt[1] | Guido Zuccon[1] | Bevan Koopman[2]

[1]School of Information Technology and Electrical Engineering, The University of Queensland, St Lucia, Queensland, Australia

[2]Australian eHealth Research Centre, The Commonwealth Scientific and Industrial Research Organisation, Brisbane, Queensland, Australia

Correspondence
Anton van der Vegt, School of Information Technology and Electrical Engineering, The University of Queensland, St Lucia, QLD 4072, Australia.
Email: a.vandervegt@uq.net.au

Abstract

Previous research has found that improved search engine effectiveness—evaluated using a batch-style approach—does not always translate to significant improvements in user task performance; however, these prior studies focused on simple recall and precision-based search tasks. We investigated the same relationship, but for realistic, complex search tasks required in clinical decision making. One hundred and nine clinicians and final year medical students answered 16 clinical questions. Although the search engine did improve answer accuracy by 20 percentage points, there was no significant difference when participants used a more effective, state-of-the-art search engine. We also found that the search engine effectiveness difference, identified in the lab, was diminished by around 70% when the search engines were used with real users. Despite the aid of the search engine, half of the clinical questions were answered incorrectly. We further identified the relative contribution of search engine effectiveness to the overall end task success. We found that the ability to interpret documents correctly was a much more important factor impacting task success. If these findings are representative, information retrieval research may need to reorient its emphasis towards helping users to better understand information, rather than just finding it for them.

1 | INTRODUCTION

Is there a disconnect between the effectiveness of an information retrieval (IR) system and the success of the searcher's end task? This is a fundamental question that goes to the heart of the IR discipline. Since Cleverdon (1960) first began evaluating IR systems, the systems approach to IR system evaluation dominated research in the field. The systems approach, such as that employed within Text Retrieval Conference (TREC) programs (Voorhees & Harman, 2005), typically evaluated IR systems, without users, on the basis of the relevance of a ranked list of documents, selected by the system in response to a query. The underlying assumption was that improvements measured in the lab translated to *real* improvements for searchers, and there is no doubt that searchers today reap the benefits of the many gains that have been made via systems oriented research.

However, there is also mounting evidence (Allan, Carterette, & Lewis, 2005; Al-Maskari, Sanderson, & Clough, 2007; Hersh, Turpin, et al., 2000; Turpin & Hersh, 2001) that gains in the lab do not always translate to gains for searchers, potentially undermining the value of some of the systems research. Furthermore, the retrieval algorithm is just one component of the search process; how important is its role in task success when compared to the role of the searcher or the corpus? Perhaps the disconnect, mentioned in the first sentence, has little to do with search engine effectiveness and much more to do with the user's abilities, or corpus content. Answering these questions has broader implications for the future direction of IR research.

To explore these themes, a user study was conducted to assess the benefits of medical literature search systems. The medical domain is an obvious candidate for such studies because of the importance that finding clinical evidence can have on patient outcomes and overall healthcare efficiency (Marshall, 1992; Marshall et al., 2013). There is also a long tradition of assessing the benefits of IR systems within the medical domain, starting with Hersh, Pentecost, and Hickam (1996). Hersh (1994) identified numerous problems associated with the systems approach to clinical search system evaluation and asserted that a topical and situational view of relevance was insufficient. Drawing upon, among others, Saracevic's (1975) broader consideration of relevance and Schamber, Eisenberg, and Nilan's (1990) user-oriented thinking, Hersh proposed an outcomes-oriented approach for medical IR system evaluation. In Hersh et al. (1996), this outcomes-oriented approach was realized by comparing the effectiveness of two MEDLINE IR systems by their ability to support medical students to answer clinical questions. Since then, this and similar approaches have been widely utilized by researchers (Hersh, Turpin, et al., 2000; Hersh et al., 2002; McKibbon & Fridsma, 2006; Westbrook, Coiera, & Gosling, 2005).

The study presented in this work was built off this long tradition; however, unlike prior clinical studies of this nature, it investigated retrieval system effectiveness as a variable. To the best of our knowledge, this was the first such study to do this. In our study, clinicians and final year medical students had to answer a set of realistic clinical questions, first with just their prior knowledge, and then with the aid of a medical literature search system. Two search systems, with widely varied effectiveness, as evaluated with a batch-style approach on the same corpus with similar types of medical questions, were provided in alternating fashion to the participant. In this way, task success could be measured by clinical decision accuracy, and task efficiency could be measured by the time to complete the task. The specific research questions investigated were:

RQ-1 What is the impact of varying retrieval effectiveness, as evaluated using a batch approach, on clinical decision making, in the context of medical literature search? This impact to be measured on task effectiveness (decision accuracy) and task efficiency (time to find answer).

RQ-2 Are search engine differences reported in batch evaluations also found when evaluating on multiple real user queries?

RQ-3 What is the relative contribution to end task success of search engine retrieval effectiveness when compared to that of the corpus and the searcher?

1.1 | Related work: search engine effectiveness and user task success

Turpin and Hersh (2001) provide two explanations for why IR systems evaluated as more effective in the lab (i.e., using a batch approach) do not always translate to better task success: (1) the system effectiveness of the batch-evaluated system does not translate to similar effectiveness in the interactive user environment because of the varied effectiveness of the user's multiple queries for the same topic; and (2) the lab system does translate to similar effectiveness in the user environment, but this does not convert to improved outcomes for the searcher. In Figure 1, we extend and refine these potential change factors to demonstrate how they may impact the extent to which an IR system's batch evaluation results translate to final task success.

Figure 1 shows that the batch evaluated IR system is tested over a number of topics, with usually a single query per topic. A topic defines the information need, whereas the query is the search phrase input to the IR system to find relevant documents for the topic. For each query, the system produces a document ranking, which is evaluated against a set of, usually expert derived, document relevance assessments (QRELs) and calculated using standard IR metrics (e.g., MAP, nDCG). The set of information needs represented by the choice of topics in the batch environment, may not be representative of the actual information needs specified in the interactive environment, and therefore, represent a potential change factor, referred to as Δ_{topic}. The impact of this factor will depend on how well the IR system can generalize its

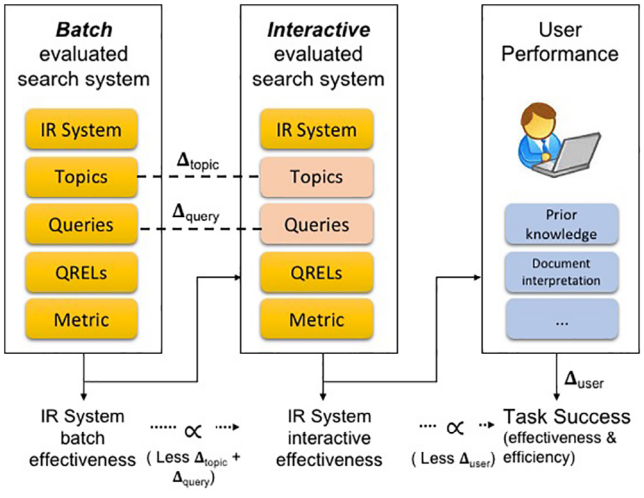


FIGURE 1 The potential change factors that may impact the extent to which an information retrieval (IR) system's batch evaluation results translate to final task success [Color figure can be viewed at wileyonlinelibrary.com]

effectiveness to topics outside of those in the batch test collection.

Turpin and Hersh's (2001) first potential change factor (i.e., (1) above) is related to the user's query variations, rather than topic variation, and is denoted by Δ_{query} in Figure 1. Their second potential change factor (i.e., (2) above), is denoted by Δ_{user} and may include the user's prior knowledge, search skills, and ability to interpret the documents found; each of these may impact the final task success, irrespective of the IR system interactive effectiveness.

We can apply this model and terminology to relevant prior research. Hersh, Turpin, et al. (2000) first identified a disconnect between IR system batch evaluation results and task success. Twenty-four participants were required to search for as many instances (i.e., aspects of a topic, for example, countries growing wheat) as possible, for six topics, in a 20 min/topic time-frame. Half the searches were performed with a baseline IR system and the other half on an IR system with a significantly higher batch effectiveness, that is, an 81% higher average precision (AP). Despite the higher AP, instance recall only improved by 18%, and this gain was not significant. Different topics were used between the batch and interactive user environments, which Hersh, Turpin, et al. (2000) suggested may have been the cause of the diminished IR system interactive effectiveness; that is, the impact of Δ_{topic} reduced the 81% batch effectiveness difference to an 18% interactive effectiveness difference. This final difference in interactive effectiveness was inline with the final difference in task success of 18%, perhaps indicating that interactive effectiveness is a better indicator of end task success than batch evaluated retrieval effectiveness.

In Allan et al. (2005), task success was directly compared to variations in IR system batch effectiveness, so that the impact of the Δ_{topic} and Δ_{query} potential change factors was excluded from the study. This was achieved by developing ranked lists of documents, at specified retrieval effectiveness levels, for each topic, to present to users, irrespective of their query. The bPref measure was used as the effectiveness metric; it is calculated as the number of relevant documents that are ranked before nonrelevant documents. The search task was also an instance recall task. Allan et al. (2005) reported that average time on task (task efficiency) reduced as bPref increased; however, the differences were only significant between document rankings with bPref levels of 50 and 93, 60 and 90 and then from 80 and above with higher levels. This means, for example, that although a user may utilize a retrieval system, evaluated using a batch approach with 80% greater effectiveness (i.e., bPref = 50 to 90) than another system, their task efficiency would remain the same. Similarly, for recall (task effectiveness), IR system batch effectiveness gains above bPref = 50 and below 80 had no significant impact. In addition, the maximum reported recall was 60% across document rankings at all bPref levels. In these scenarios, given a floor bPref value of 50, task success was disconnected from improvements in IR system batch effectiveness. To further improve task success would require better understanding of the impact of Δ_{user} to see if the user was limiting improvement; this is the subject of RQ-3 in our study.

So far, the experiments described above have employed recall-based search tasks to identify a disconnect between batch-style search system effectiveness and user task success. Turpin and Scholer (2006) sought to understand if the same disconnect existed for simple precision search tasks. Like Allan et al. (2005), the search system itself was removed from consideration by generating fixed rankings of specified mean average precision (MAP) levels; so Δ_{topic} and Δ_{query} change factors were excluded from this study. Thirty students were tasked with finding the first relevant document for 50 topics. No dependency was found between MAP and the time to find the first relevant document, signifying a similar disconnect between IR system batch evaluation and task success, to recall search tasks.

In summary, prior work has demonstrated that significant changes to IR system batch effectiveness often have little or no impact on search task success (effectiveness and efficiency). A surprising and controversial outcome. Yet, all of these studies concentrated on simple precision and recall tasks for general search. Also, although realistic in terms of the type of task a user may engage in (e.g., find as many instances within a time frame), they were not realistic specific tasks that were likely to be pertinent to the personal or working life of the participants. We were, therefore, interested to see whether the same findings hold, without these limitations, by posing complex and realistic search tasks, of pertinence to the searcher group, with task-oriented decision-making outcomes. The clinical domain provides a good use-case because of the complex nature of clinical evidence and because of the robust methodology that already exists for evaluating medical search systems (see next section). Our study employed clinicians to answer realistic clinical questions where the quality of the clinical decision, based on their search, could be assessed.

1.2 | Related work: medical search

Numerous studies have measured and demonstrated the benefits of using clinical evidence search systems; however, none of these studies considered IR system effectiveness as a variable. As mentioned above, Hersh

et al. (1996) began this work by comparing two MEDLINE search systems: one employed Boolean retrieval and the other natural language retrieval. The evaluation found that prior to using the search systems, the 12 medical students had a lower-than-random correct answer rate for the 12 clinical questions they answered. However, after search, this improved to 10 correct answers, with no significant difference between both systems. In addition, there was no significant difference in the search time. This method of evaluating search system effectiveness on the basis of the change in the correct answer rate (effectiveness) and the time to search (efficiency) became the standard for future clinical search system evaluations.

Two further studies were conducted with MEDLINE only search systems (Hersh, Crabtree, et al., 2000; Hersh et al., 2002), which is similar to our study in which a subset of MEDLINE was used for search. In Hersh et al. (2002), 45 medical and 21 nurse practitioner students answered a total of 324 questions. The correct answer rate improved from 32.1% (104/324) pre-search to 46.3% (150/324) post-search. After this, a number of studies were conducted with multiple search corpora.

McKibbon and Fridsma (2006) assessed how well 23 clinicians could answer 2 clinical questions of their choice from 23 available questions. The clinicians could reference multiple data sources of their choice, including PubMed and MEDLINE. Their correct answer rate only improved from 39% (18/46) pre-search to 41% (19/46) post-search. Westbrook et al. (2005), on the other hand, found more extensive improvement. They studied the answer accuracy of 8 clinical questions presented to 75 clinicians, including nurses and doctors. The study showed that the introduction of a clinical evidence search system improved the correct answer rate from 29.0% (124/600) pre-search to 49.7% (298/600) post-search. The search system comprised six sources of evidence, MEDLINE included.

In general, these studies indicate that clinical evidence search systems do benefit clinicians and enable them to answer around a half of the clinical questions correctly. However, none of the studies considered search system effectiveness as an independent factor and nor did they account for why around half of the clinical questions remained unanswered, even after using the search system. As far as we are aware, our study is the first to do so.

2 | EXPERIMENTAL METHODS

The protocol for this study is reported by van der Vegt, Zuccon, Koopman, and Deacon (2019). A summary of the study design is provided here.

2.1 | Study design overview

One hundred and nine participants consisting of practicing clinicians and final year medical students were provided with 16 clinical scenarios, each with a single question. Figure 2 depicts the study steps. The participants had to first answer the questions without any supporting evidence. In the second stage of the study, the same set of clinicians were provided with the same 16 clinical scenarios and a medical search system. Unbeknownst to the participants, for each question the underlying search algorithm is alternated between two systems of significantly different effectiveness levels. Also, the time allowed to search for an answer is also controlled and constrained to one of 3, 6, or 9 min. This allowed us to assess the impact of the search systems under differing levels of time pressure.

2.2 | Participants

A convenience sample of 109 practicing clinicians and final year medical students, including nurses, general practitioners and hospital physicians, were asked to participate. The practicing clinical participants were Australian registered clinicians, residing in Australia. All participants required access to a computer with Internet access. Participants were offered a small honorarium ($50 gift card) to complete the assessment and were recruited via e-mail and online noticeboards directed to clinical departments in hospitals, public health area networks and medical faculties at Australian universities.

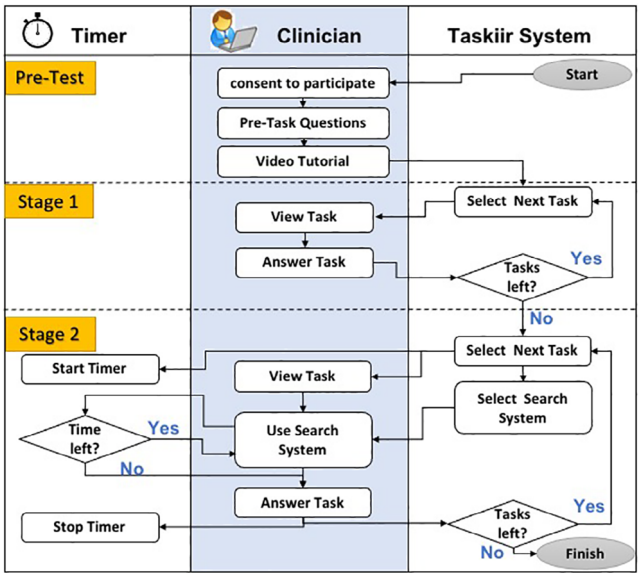


FIGURE 2 Process flow diagram of study showing both stages [Color figure can be viewed at wileyonlinelibrary.com]

2.3 | Procedures

Participants were asked to complete a 2-hr, web-based assessment of a medical search system called Taskiir. After voluntary consent was received, the participants were allocated their login details via email. In the e-mail, the participant was advised that they could perform the study in multiple sittings, within a 2-week period, at a time to suit them and that they had to use their laptop/computer (not iPad) to access the study on the web.

After initial login, the participant was asked seven questions to capture demographic data, search, and medical experience. A 5–10-min, video tutorial followed, where the study was described in more detail and the participant was shown how to use Taskiir, the medical search engine. Once complete, the participant was shown specific instructions that reinforced their obligation to perform the test alone, before they were permitted to move onto the two-stage assessment.

In stage one, 16 clinical tasks were presented to the participant, one-at-a-time. To complete each task the clinician had to answer a single question, within a few minutes, although this time limit was not enforced. Fourteen of the 16 tasks required the participant to select one of four answers (yes, no, conflicting evidence, and do not know) and the other tasks required a 1–2 word answer. At the end of the last task, the system moved the participant to stage two of the study.

In stage two, the participant had to complete the same 16 tasks, in the same order as stage one; however, the participant had to use Taskiir to help them to answer the question and to find evidence to support their answer. Evidence was collected by the participant selecting text and/or images from the source documents they read. The time allocated to search for each task was assigned to 3, 6, or 9 min, based on the timing-cohort the participant was placed into. The participant was told of the time allocation at the start of each question and a minute-by-minute countdown timer was always visible to the participant; warnings were given 30 s prior to time out. At time-out, the screen was blocked and the participant was taken to the task completion screen to enter their final details.

2.4 | Clinical tasks

Six of the 16 clinical questions were those produced and used by Westbrook et al. (2005). The tasks consist of real-life scenarios and a clinical question for each scenario. Westbrook et al. derived the tasks using clinical experts and designed them to be clinically relevant and of mixed complexity. Four questions were sourced from Hersh et al. (2002, table 2), which were also clinical questions and used for the same purposes as this study. Three questions were modified from the TREC 2015, Clinical Decision Support (CDS) topic set (Simpson, Voorhees, & Hersh, 2014b). These questions were provided with diagnoses, which our medical physician (Dr Anthony Deacon, MBBS), modified into a question of a similar format to the other questions. Finally, our medical physician also devised a further three other clinical questions for the purposes of this test. To ensure that at least one relevant document existed in the corpus for each task, our medical physician searched through the corpus, using Taskiir, to identify one or more relevant documents. A sample question is:

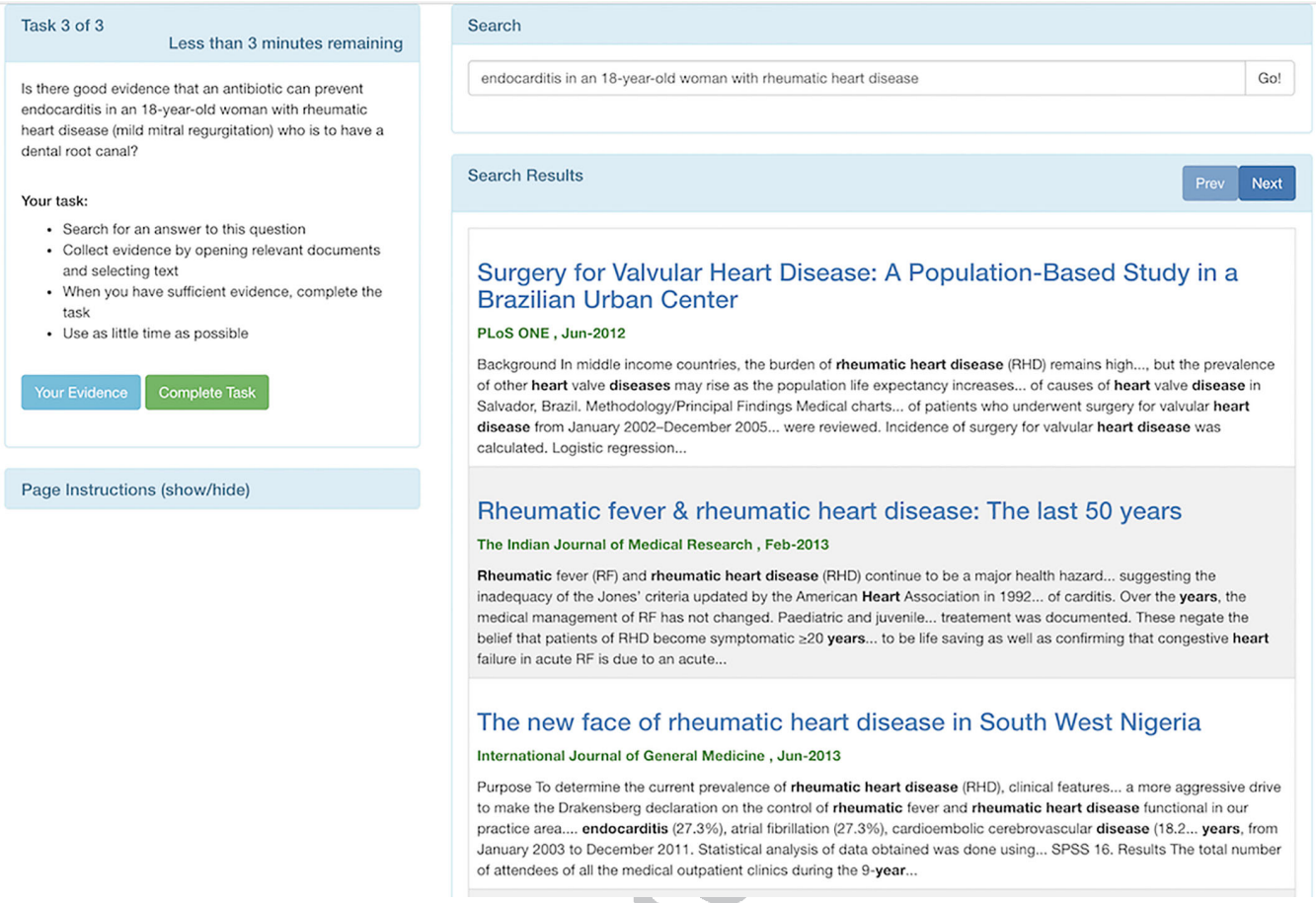
> *A 48 year old man presents with severe right sided loin pain and is diagnosed with a 4 mm distal ureteric calculus. Has Tamsulosin been shown to increase the chances of the calculus passing?* Answer = Yes; source evidence PMIDs = (3364475, 2943682)

A full listing of questions and answers can be found in van der Vegt et al. (2019). To avoid the confounding effects of fatigue and question order, a Latin square experimental design was constructed for 16 tasks and 16 participants with randomized columns.

2.5 | Corpus and medical search system (Taskiir)

The clinical information corpus used was the TREC, CDS Track 2014 and 2015 document collection (Simpson et al., 2014b; Simpson, Voorhees, & Hersh, 2014a). This consists of a snapshot of the Open Access Subset of PubMed Central taken on January 21, 2014. It contains a total of 733,138 articles.

A custom document search engine and interface, together called Taskiir, was employed for the evidence search process (see Figure 3). Similar to normal commercial search engines, Taskiir allowed the participant to write their query and perform a best match search of documents in the corpus. A snippet, highlighting matching query terms, was then provided in the Search Engine Results Page (SERP), which showed up below the query. Users could then select documents of interest to view the full text. While viewing the full text document, the participant could also select (with their mouse) any text or graphics which they wanted to use as evidence for their



[Color online, B&W in print]

FIGURE 3 Screen shot of the Taskiir custom search system interface. Shows the task in the top left, search query box, top right, and search results below [Color figure can be viewed at wileyonlinelibrary.com]

final answer. The participant could view their evidence or complete the task at any time. Instructions on using the system were provided on each page and a mandatory walk-through tutorial was provided prior to starting the study. Taskiir utilized two document retrieval algorithms:

- State-of-art (SOA): An improved version of the TREC 2015 CDS Task A best performing system by Balaneshin-kordan, Kotov, and Xisto (2015). The TREC track was targeted to identify the state-of-art IR system because the topics in Task A were of a similar clinical nature to our questions and the search corpus was the same as that used in this study. The two improvements made over the system included the removal of negated UMLS terms from the UMLS query expansion terms as well as a change to the pseudo relevance feedback term weighting (from 0.75 to 0.5). All improvements resulted from tuning parameters on the CDS 2014 test collection and testing on the 2015 collection, to avoid over-fitting.
- Baseline document retrieval system (BM25): BM25 standard retrieval system is a widely adopted best-match retrieval method. It is the default, out-of-the-box method employed by many search engines including the very popular Elasticsearch[1] and Lucene[2] systems. The parameters were set to default values ($K = 1.2$, $b = 0.75$).

Document retrieval effectiveness figures for both systems are shown in Table 1. The measures depicted were the standard set chosen for the TREC 2014 and 2015 CDS tasks. IR system effectiveness measures are usually calculated for a ranked retrieval of 1,000 documents.

3 | RESEARCH QUESTION 1: IMPACT OF VARYING RETRIEVAL EFFECTIVENESS ON CLINICAL DECISION MAKING

Participants. A total of 109 participants (16 doctors, 8 nurses, and 85 final year medical students) answered 16 questions. Of the 1,744 samples, 85 were discarded because the participant failed to search for the answer,

TABLE 1 Comparison of document retrieval effectiveness figures, across the TREC 2015 test collection, for systems used in this study and the best performing TREC CDS 2015 system Balaneshin-kordan et al. (2015)

System	infNDCG	infAP	P@10	R-prec	MAP
WSU system[a]	0.2928	0.0777	0.4633	0.2329	0.1851
SOA	0.3159	0.0849	0.4800	0.2401	0.1930
BM25	0.2168	0.0461	0.3600	0.1717	0.1114
SOA vs. BM25	+46%***	+84%*	+33%***	+40%***	+73%***

[a]As per TREC 2015, CDS, Task A, Automatic Runs listed in Simpson et al. (2014b, table 4) for summary topics.
*Significance using paired *t* test with *p*-values < .05.
***Significance using paired *t* test with *p*-values < .0005.

TABLE 2 Summary table comparing the post correct answer results and answer direction results for the two search systems

	Time-constraint cohort						All	
	3 min		6 min		9 min		Cohorts	
	BM25	SOA	BM25	SOA	BM25	SOA	BM25	SOA
Sample size	276	281	268	282	277	269	821	832
Pre-search correct								
# of samples	104	86	92	100	89	91	285	277
% of cohort samples	38	31	34	35	32	34	35	33
Post-search correct								
# of samples	141	153	140	155	155	142	436	450
% of cohort samples	51	54	52	55	56	53	53	54
Improvement								
# of samples	37	67	48	55	66	51	151	173
% of cohort samples	13	24	18	20	24	19	18	21
% improvement	36	78	52	55	74	56	53	62
Answer direction								
Right-to-wrong (RW)								
# of samples	44	28	24	32	27	34	95	94
% of cohort samples	16*	10*	9	11	10	13	12	11
Wrong-to-right (WR)								
# of samples	81	95	72	87	93	85	246	267
% of cohort samples	29	34	27	31	34	32	30	32

Note: Data are provided by search constraint as well as overall.
*BM25: SOA *p* adj = .0356 (Tukey HSD).

indicating that the search system was not used; a further 6 samples were discarded due to a system failure. This left 1,653 samples for analysis.

The gender split of the participants was slightly biased overall towards females (53%). The median self-reported rating for computer skills was 4 (*very good*) for both students and overall; however, the median for doctors and nurses was 3 (*good*). In terms of MEDLINE/PubMed usage, the median, self-reported usage across all participants was 3 (2–3 times per month) with the median for nurses being slightly lower on 2 (once per month).

3.1 | Results: impact on task effectiveness

The impact of retrieval effectiveness differences (independent variable) on task effectiveness was assessed by identifying changes in answer accuracy. The correctness of all

questions was assessed before using the search system, and then after. Table 2 provides a comparison of the two systems. It was found that across all participants that the correct answer rate improved by a significant 20 percentage points, from 34% pre-search to 54% post-search ($\chi^2 = 148.62$, $df = 1$, p-value < 2.2e−16, McNemar's chi-square test). Participants using the SOA system improved their correct answer rate by 62% compared with 53% for BM25 system users; however, this difference was not significant. Also, across all search time cohorts (3, 6, and 9 min), there were no significant differences between the two systems.

Changes to the underlying answer directions were also considered, that is, where the participant changed their pre-search answer, after using the search system. The data are plotted in Figure 4 and reveal that under the most limited search conditions (i.e., 3 min), there was a significant difference between the right-to-wrong answer direction for the two systems. In this case, participants using the SOA system incorrectly changed their pre-search answer 38% less than participants using the BM25 system (from 16 to 10%, p adj-value = .0356 using Tukey HSD). At this same time constraint (3 min), both the wrong-to-right answer rate and the post-search correct answer rate also showed improvements when using the SOA system; however, none of these differences were significant. At all other time constraint levels there were no significant differences for either answer correctness or answer direction. In summary, a statistically significantly more effective batch evaluated search system did not lead to more accurate clinical decisions.

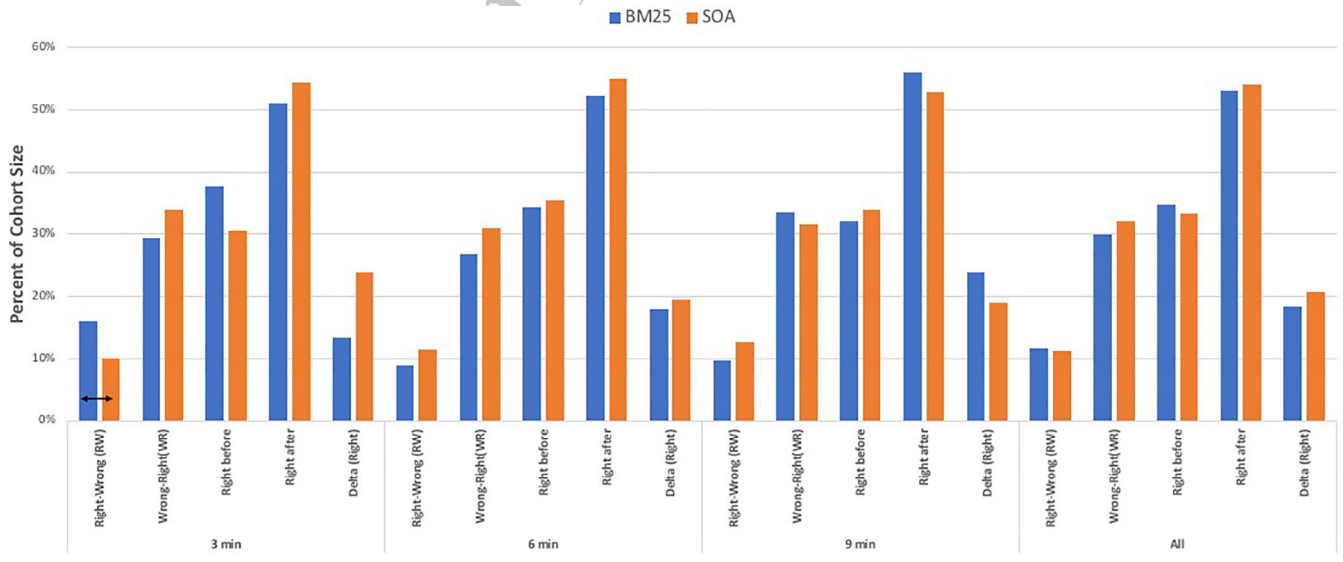
3.2 | Results: impact on task efficiency

The impact of retrieval effectiveness differences (independent variable) on task efficiency was assessed by identifying changes in the time to search for, and answer, the question. The average time to complete the task was 253 s for SOA users, down by 3% when compared with the BM25 system users; however, this difference was not significant. Table 3 provides further details relating to the average read time, time spent on the SERP as well as search behaviors, such as the task averages for the search count and number of document views. Although the SOA system provided minor task efficiencies, none of the improvements was significant.

4 | DISCUSSION

With the aid of a medical search system, participants in our study were able to answer eight questions correctly (standard deviation [*SD*] 2.2, range 0–13), improving their answers from 34% correct pre-search to 54% post-search. This improvement is in line with previous studies, including Hersh et al. (2002) and Westbrook et al. (2005), in which clinicians improved their correct answer rate by 20 and 21 percentage points, respectively.

We found that a step-change in batch-evaluated search system effectiveness (73% significant increase in MAP) had minimal impact on both task effectiveness and task efficiency for clinicians. In terms of a lack of impact, our findings match those of Hersh, Turpin, et al. (2000)



Color online, B&W in print

FIGURE 4 Answer correctness and direction results compared for the two search systems, stratified by the time constraint task cohorts (3, 6, and 9 min) and the overall result. Any significant differences are indicated by a black arrow between the columns [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Summary table of average task time and search behaviors for each system

	Search system		
	BM25	SOA	Difference[a]
Cohort size	821	832	1%
Averages per task			
Total task time (s)	260	253	-3%
SERP time (s)	90	84	−7%
Read time (s)	141	138	−2%
Search count	2.5	2.3	−7%
Document view count	2.2	2.1	−3%
Documents viewed per search	0.9	0.9	5%
Read/view time per document viewed	64	65	1%

[a]No differences are significant.

and Turpin and Scholer (2006); however, we can now extend their findings to include much more realistic and complex search tasks that require expert decision making.

Using a subset of MEDLINE (around 3%) represents a limitation of this study. It is possible that across the whole of MEDLINE, the differences between the two search systems may have been more prominent. However, because this study took place over 12 months, it would not have been possible to control for changes to the MEDLINE corpus, which may have biased the results. By limiting the corpus to the open access subset of MEDLINE, this potential confounding factor was mitigated.

The only statistically significant difference identified was found when operating under the greatest search time constraint (3 min): participants using the SOA system changed their correct pre-search answer 38% less than BM25 users. This means that a better search system is less likely to mislead a searcher who is expert in the search domain. In addition to search time pressure, we also investigated whether the two systems had a different impact across varied task difficulty. We ranked task difficulty based on the number of correct post-search answers.

Figure 5 graphs the percentage of tasks that were correctly answered post-search for each of the two system cohorts. The graph reveals that the percent correct rate for both systems was very similar on a question-by-question basis; however, when ordered from most to least difficult, it appears that the SOA system was associated with better correct answer rates for the most difficult half of the question set. Indeed, the correct rate for the SOA system for the hardest eight questions was 38% (155/413) compared with 31% (129/413) for the BM25 system, which represents a 20% improvement; however, this difference was only weakly significant (Tukey HSD p adj = .05693). The corresponding differences in correct answer rate for the easiest eight questions was 70% (294/418) for the SOA system and 75% (307/408) for the BM25 system, which was not significant. Without further data or corroborating evidence we cannot say that the better search system had a significant positive impact on harder clinical questions. Further research, potentially with a larger data set, may resolve this question.

5 | RESEARCH QUESTIONS 2 AND 3

Despite the aid of a medical search system, approximately half of the clinical questions were incorrectly answered. Were these failures a result of the search system, the content of the corpus or the user (i.e., RQ-3)? We employ a factor analysis method, detailed in this section, to address this question. Also, although we have identified that a step change in batch-evaluated system effectiveness did not translate into a similar change in user task performance, we did not ascertain the impact of the potential change factors: (a) Δ_{topic}; (b) Δ_{query}; or (c) Δ_{user}. The same factor analysis is used to also address this question (RQ-2).

5.1 | Method: introduction to the success factor model

Figure 6 provides a process flow chart of the study, stage 2, interactive search subprocess. The three key actors in the subprocess are the participant, the search system, and the corpus. Highlighted by yellow tags on the flow chart are factors that can impact the post-search correct answer rate. Each of the factors is described below, together with how they are implemented; but in order to understand

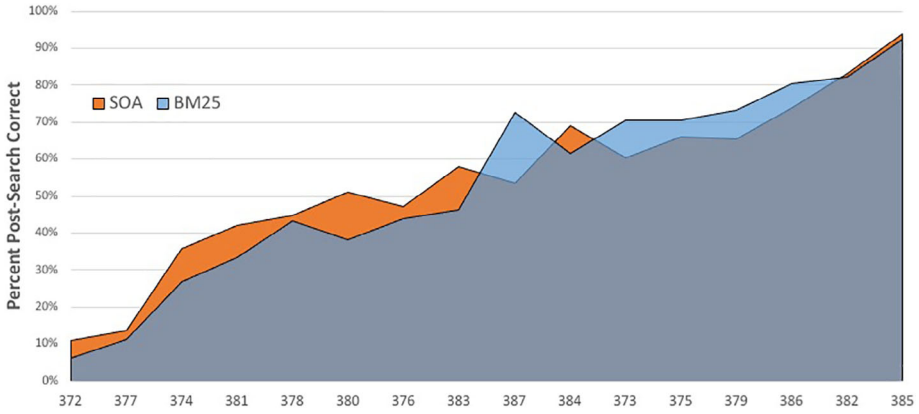


FIGURE 5 The average percent post-search correct answer rate by question for each search system cohort; BM25 (blue) and SOA (orange). Questions are ordered from hardest (lowest correct answer rate) to easiest (highest correct answer rate) [Color figure can be viewed at wileyonlinelibrary.com]

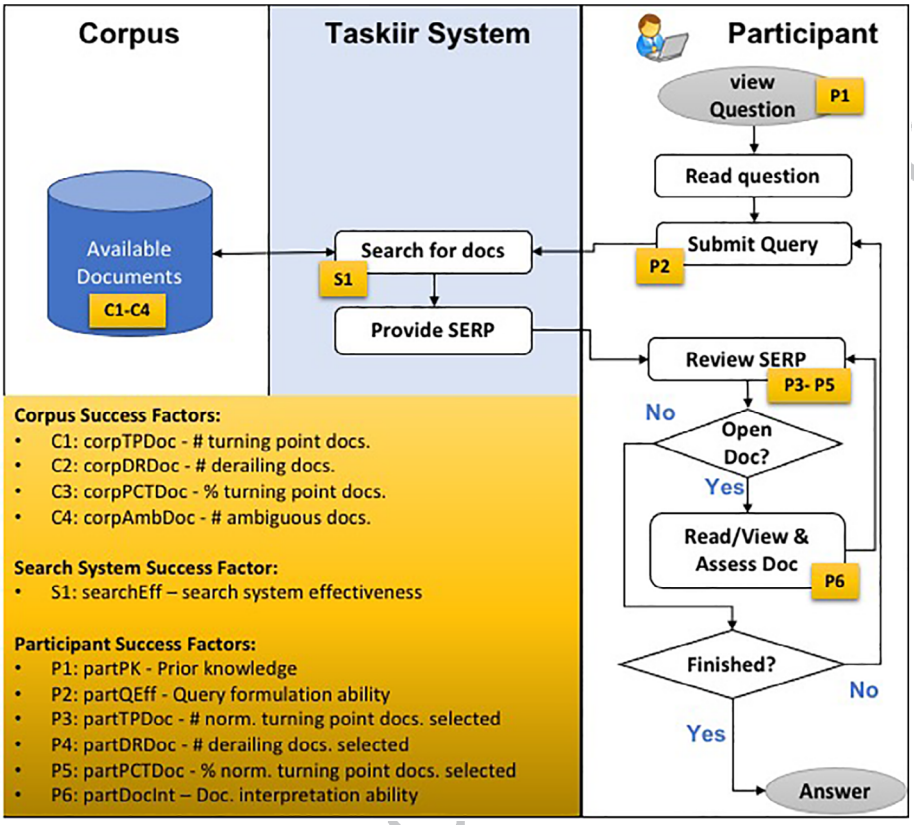
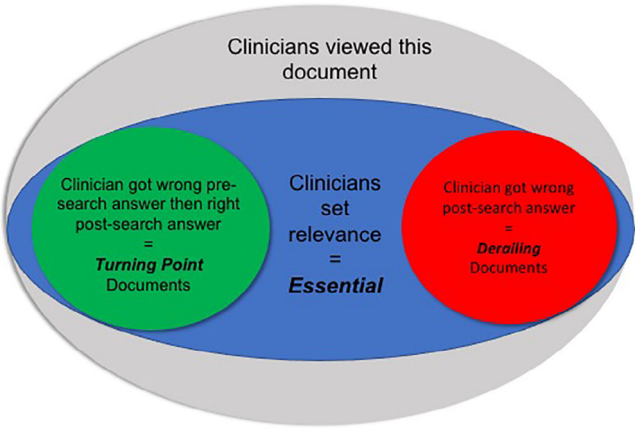


FIGURE 6 Interactive search subprocess for stage 2 of the study. The yellow tags identify process success factors which are described briefly in the yellow box [Color figure can be viewed at wileyonlinelibrary.com]

these factors we need to first explain how relevance assessments were collected.

5.2 | Document relevance assessments

In our study, relevance assessments were collected from the participants during the study. Each time a participant viewed a document in the Taskiir search engine, they had to select a relevance rating for that document, prior to closing it. The possible ratings were: (1) *essential*; (2) *helpful*; (3) *not helpful*; (4) *essential duplicate*; and (5) *helpful duplicate*. We found that using ratings (1) and (4) were more significantly related to the post-search correct rate when just wrong-to-right tasks were considered, $F(1,1651)= 90.85$, $p < 2e{-}16$. The intuition for selecting this basis of relevance assignment is that the participant must have used these documents to change their thinking to the correct answer and thus these documents, called *turning point* documents, are more likely to be relevant. Similarly, we define *derailing* documents as those that the participant viewed and marked as essential, but they then went on to answer the question incorrectly. *Ambiguous* documents are those that are identified as both turning point documents and derailing documents. The Venn diagram depicting these document sets is provided in Figure 7.



Color online, B&W in print

FIGURE 7 Venn diagram showing the selection of Turning Point and Derailing documents [Color figure can be viewed at wileyonlinelibrary.com]

For easier questions, typically more turning point documents are found. To normalize the count of turning point documents across all questions, we identify the top 3 by count, for each question, and call these, *normalized turning point documents*.

5.3 | Summary of success factors and their grading

The success factors are listed in Table 4. Each factor is graded for each participant's post-search result into one of three categories: (1) green, denoting a good result for this factor; (2) amber, denoting a normal result; and (3) red, denoting a bad result for the factor. The basis of grading is also provided in Table 4. Other details relating to the factors are provided below.

The corpus factors. These are a limited estimate of the extent of corpus content available that is relevant to each question. They are not derived through an exhaustive search and assessment of corpus documents, but instead are inferred by the findings of the participants using the two search systems. True figures are therefore likely to be higher in the corpus. Document ambiguity is based on the assumption that the documents themselves are ambiguous, rather than assuming that the participants who use them have poor interpretation skills.

Search system factors. searchEff is the only factor representing search system effectiveness. It is evaluated using normalized turning point documents as a proxy for document relevance. We selected normalized discounted cumulative gain (nDCG) as the measure for system effectiveness because it is well matched to the searcher's behavior. In this study, searchers primarily use the first SERP page. The mean depth of document selection across all users is at snippet 2.9 ($SD = 2.3$) and the average maximum depth is at snippet 4.2 ($SD = 4.2$).

To calculate nDCG using turning point documents (i.e., $nDCG_{tp}$), binary relevance values are used: Normalized turning point documents = 2, all others (i.e., nonrelevant) = 0. $nDCG_{tp}$ is calculated for each participant's question as follows: the search system identifies 10 ranked documents for the SERP for each search that the participant conducts for a single question; the ideal (where turning point documents are positioned at the top of the ranking) and actual discounted cumulative gain is evaluated for these 10 documents, depending on whether the search was the first search (i.e., ranks 1 to 10), or a latter search (i.e., ranks 11 to 20, etc.). By dividing the sums of actual and ideal cumulative gain across all their searches, a single average ($nDCG_{tp}$) value is derived by participant by question.

Participant factors. The partPK factor represents a participant's prior knowledge. partQEff is an attempt to identify the participant's ability to formulate effective queries, in terms of the query's evaluated $nDCG_{tp}$. Factors partTPDoc, partDRDoc, and partPCTDoc attempt to identify good searcher behavior. When participants perform a search they make a multitude of decisions regarding which documents to view, how far to look down the ranking for a single search or when to try a new search. The ultimate outcome of this searcher behavior is that turning point, derailing or nonrelevant documents are opened and viewed. The selection of turning point or derailing documents could result because few turning point documents are present in the SERP or because the snippets are misleading. Research to separate these factors is left for future work. The final participant factor is document interpretation, denoted partDocInt. Once documents are viewed, participants have to interpret the documents and arrive at an answer. The measurement of partDocInt considers the proportion of normalized turning point and derailing documents that are viewed by the participant as well as whether or not the participant answered correctly.

5.4 | RQ-2 results and discussion: are search engine differences reported in batch evaluations also found when evaluating on multiple real user queries?

With reference to Figure 1, to address RQ-2 we need to compare the difference in batch evaluated search system effectiveness between the SOA and BM25 system with the difference in interactive evaluated search system

TABLE 4 Details for the red and green grading of the success factors

Factor	Green grading method	Red grading method
Corpus: document mix—assessed by question		
corpTPDoc	The corpus contains above normal (average + 1 *SD*) count of turning point documents	The corpus contains below normal (average − 1 *SD*) count of turning point documents
corpDRDoc	The corpus contains below normal (average + 1 *SD*) count of derailing documents	The corpus contains above normal (average − 1 *SD*) count of derailing documents
corpPCTDoc	Corpus contains above normal (avg. + 1 *SD*) ratio of turning point: derailing docs	Corpus contains below normal (avg. − 1 *SD*) ratio of turning point: derailing docs
corpAmbDoc	Corpus contains above normal (avg. + 1 *SD*) count of ambiguous docs	Corpus contains below normal (avg. − 1 *SD*) count of ambiguous docs
Search system: effectiveness—assessed by question		
searchEff*	$nDCG_{tp}$ above normal (average + 1 *SD*) for question	$nDCG_{tp}$ below normal (average – 1 *SD*) for question
Participant: clinical expertise—assessed by participant and question		
partPK*	Correct pre-search answer	Incorrect pre-search answer
Participant: query formulation ability—assessed by participant and question		
partQEff	The participant formulates queries that are above normal (average + 1 *SD*) in effectiveness compared to their peers	The participant formulates queries that are below normal (average − 1 *SD*) in effectiveness compared to their peers
Participant: searcher behavior—assessed by participant and question		
partTPDoc	Participant selects above normal (average + 1 *SD*) normalized turning point documents	Participant selects below normal (average − 1 *SD*) normalized turning point documents
partDRDoc	The participant selects below normal (average − 1 *SD*) derailing documents	The participant selects above normal (average + 1 *SD*) derailing documents
partPCTDoc*	Participant views more normalized turning point documents than derailing documents	Participant views less normalized turning point documents than derailing documents
Participant: document interpretation—assessed by participant and question		
partDocInt	The participant viewed more derailing documents than normalized turning point documents and yet still answered the question correctly post-search	The participant viewed more normalized turning point documents than derailing documents and yet still answered the question incorrectly post-search

Note: The factors appended with a * indicate that no Amber grading level exists for that factor, otherwise, the tasks are assigned an amber grading if they do not fall into either the green or red grading categories.

effectiveness for the same two systems. Table 1 provides the batch evaluated document retrieval effectiveness of the two systems. To derive the interactive evaluation of the two systems we can generate an average $nDCG_{tp}$ across all participant tasks for each system. The results show that the SOA system was 9% more effective than the BM25 system ($nDCG_{tp}$ = 0.5260 and 0.4824 respectively; difference tested using Tukey HSD p-adj = .0016).

This difference is much less than the effectiveness difference that was identified using the TREC CDS 2015 test collection, which demonstrated at least a 33% difference across a range of evaluation measures (although not nDCG). As indicated by Figure 1, two potential change factors may account for this diminished effectiveness difference: (i) Δ_{topic}: the 2015 CDS test collection topic set is different to the topic set (questions) in our study; (ii) Δ_{query}: the 2015 CDS test collection topic set uses fixed queries (summary or description) for system evaluation; whereas in our user study, multiple queries can be used for the same question. Interestingly, the difference in improvement in correct answer rate, from pre-search to post-search, is 9 percentage points between the SOA and BM25 systems, however this difference is not significant.

It is also possible that the evaluation method ($nDCG_{tp}$) itself is confounding the results and undervaluing the differences between the systems. It is possible that with more relevance assessments, beyond those that were done by the participants alone, that system differences may increase. However, our findings are in line with the original work by Hersh, Turpin, et al. (2000), who reported four-fold decreases in the effectiveness (AP) between their systems, when compared in the lab and then with users.

5.5 | RQ-3 results and discussion: what is the relative contribution to end task success of search engine retrieval effectiveness when compared to that of the corpus and the searcher?

In order to address RQ-3, the success factor model is employed to compare the state of the factors between the success cohort (correct post-search answer) and the failure cohort (incorrect post-search answer). The task grade counts for the two cohorts are provided in Table 5. Amber graded results for factors indicate normal behavior. Therefore, in order to identify the most significant factors that change between cohorts, we need to understand how the green and red mix changes. To do this, we define a factor statistic called the positive factor mix:

$$\text{Postive factor mix (factor } (i)) = \frac{G_{ci} - R_{ci}}{G_{ci} + A_{ci} + R_{ci}}, \quad (1)$$

where R = red, G = green, and A = amber factor grades and X_{ci} is the count of participant tasks that were assessed as grade X for factor i.

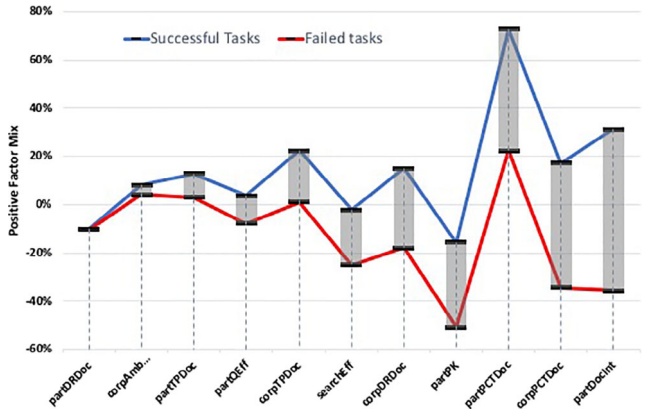
The positive factor mix is then plotted in Figure 8, for all the factors, to reveal which factors are shifting most between the success cohort and the failure cohort. Red and green count data for those factors can then be compared for significant differences. Figure 8 suggests that the top three factors accounting for failure overall are: (1) poor document interpretation (partDocInt), (2) the low proportion of turning point documents relative to the number of derailing documents in the corpus (coprPCTDoc), and (3) the low proportion of turning point documents selected by participants relative to the number of derailing documents selected (partPCTDoc). All changes to red and green factor counts are significant for these factors.

It is not surprising that document interpretation is the most critical factor impacting answer correctness. Reported in a systematic review of 2,592 articles, conducted by Sadeghi-Bazargani, Tabrizi, and Azami-Aghdash (2014), the top barriers for implementing evidence based medicine are *research barriers*, of which some of the key issues identified were, "Conflicting results, ..., lack of replication, poor generalizability, ..., literature not being compiled in one place, implications for practice not being made clear, limited relevance of research to practice"(Sadeghi-Bazargani et al., 2014, table 1). In a nutshell, medical literature is often difficult to interpret and adapt for the specific clinical case faced by the physician.

TABLE 5 Count of red, amber and green graded tasks for each of the success factors, summed for the success cohort and the failure cohort

	All-success cohort			All-failure cohort			Delta	
Factor	**Green**	**Amber**	**Red**	**Green**	**Amber**	**Red**	**Green**	**Red**
Corpus factors								
corpTPDoc	207	670	9	104	567	96	−103***	87***
coprDRDoc	260	499	127	49	531	187	−211***	60***
corpPCTDoc	248	543	95	61	382	324	−187***	229***
corpAmbDoc	254	451	181	160	480	127	−94***	−54*
Search system factor								
searchEff	205	1,031	417	31	513	223	−143***	29***
Participant factors								
partPK	562	0	1,089	189	0	578	−184***	65***
partQEff	255	1,115	283	98	511	158	−59**	33***
partTPDoc	306	1,178	169	135	520	112	−36	55***
partDRDoc	135	1,214	304	59	570	138	−17	−28
partPCTDoc	1,234	0	419	469	0	298	−296***	177***
partDocInt	277	1,104	272	0	495	272	−277***	272***

Note: The difference between the two cohorts (delta) is provided for red and green graded factors, together with their significant difference, as calculated using Tukey HSD multiple comparisons of means.
*p-Value < .05 (Tukey HSD adjusted p-values).
**p-Value < .01 (Tukey HSD adjusted p-values).
***p Value < .001 (Tukey HSD adjusted p-values).



Color online, B&W in print

FIGURE 8 Positive factor mix, by factor, for the success cohort (blue) and the failure cohort (red). The factors are ordered from those exhibiting the least change in positive factor mix, between the success and failure cohorts, and those exhibiting the greatest change [Color figure can be viewed at wileyonlinelibrary.com]

With respect to RQ-3, this factor analysis highlights that although system effectiveness is an important factor associated with effective task completion, it is less important than other factors, in particular, document interpretation, the content of the corpus in terms of the mix of derailing and turning point documents, and the ability of the searcher to select turning point over derailing documents to read. This indicates that in many cases, participants are being served up relevant, turning point documents in the SERP by the search engine, but they do not always select them and if they do select them, they are incorrectly interpreting them. This is important information for search system designers; we speculate that, for example, research on a search system which can help clinicians to better integrate the knowledge from multiple documents may yield better clinical decisions than further research on state-of-the-art retrieval models. Assessing whether these findings generalize to other or all search is warranted and has been left for future work.

6 | CONCLUSIONS

Does better retrieval effectiveness equate to better clinical decisions? In our study of 109 clinicians and final year medical students, the use of a significantly more effective search system, as evaluated using a batch-style approach in the lab, had a minimal impact on the efficiency and effectiveness of searchers performing medical search to answer clinical questions. We found that most of the system effectiveness differences reported in the lab were significantly diminished when evaluated in an interactive user environment, with different topics. We assigned the losses in effectiveness differences to losses associated with a lack of generalization to different search topics and the leveling impact of searchers using multiple queries to meet their information need. These findings agree with prior work, however they also extend the findings to much more complex and realistic search tasks, involving expert search and decision making.

Participants used two search systems to help them to answer clinical questions, but despite these systems, around half of the questions were answered incorrectly. Was this because of the search system, or were there other more important factors at play? The contribution of search system effectiveness to overall search task success was also considered in this study. Using a factor analysis approach, document interpretation was identified as the most important factor impacting end task success. The analysis demonstrated that searchers could find and view the same relevant documents, but come to different conclusions, resulting in either success or failure. These findings suggest that search system effectiveness is sufficient for medical searchers; the bottleneck now, is information interpretation.

Since Hersh (1994) first proposed an outcomes-oriented approach to evaluate clinical IR systems, many studies have incorporated clinical answer correctness as the basis for clinical search system evaluation. This study has added two further dimensions to this past work. First, it has affirmed Hersh's (1994) assertion that batch evaluated system comparison for clinical search is insufficient. Second, it has identified new factors, both user-oriented and corpus-oriented, that are independent of the search system and can have a greater impact on clinical answer accuracy. This has significant implications for medical search research and the design of medical search systems.

Probably the most important implication relates to the concept of relevance and the paradigm of document retrieval. For medical search, this study confirms that providing clinicians with a ranked list of relevant documents is insufficient. To help clinicians to correctly answer more of their questions, the IR system may need to help clinicians to interpret information, potentially from within documents and across them. We hypothesize this might mean moving beyond document retrieval, to IR. The notion of relevance, also, may need to encompass subdocument and cross-document assessment of interpretability. One such example of the provision of such information, rather than documents, has been investigated in the form of information cards, which offers new and promising directions of research towards more accurate clinical answering (Jimmy, Zuccon, Koopman, & Demartini, 2019; van der Vegt, Zuccon, Koopman, & Bruza, 2018).

In this study, the focus was the medical domain. It is difficult to generalize these findings, but perhaps other domains that require expert search might reveal similar findings. Expanding this study's approach to more general search tasks, may provide the next steps towards sizing up this opportunity.

ENDNOTES

[1] https://www.elastic.co/products/elasticsearch, accessed February 3, 2020.

[2] https://lucene.apache.org/core/, accessed February 3, 2020.

REFERENCES

Allan, J., Carterette, B., & Lewis, J. (2005). When will information retrieval be "good enough"? User effectiveness as a function of retrieval accuracy. In *SIGIR '05 Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr*. (pp. 433–440).

Al-Maskari, A., Sanderson, M., & Clough, P. (2007). The relationship between IR effectiveness measures and user satisfaction. In *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.* (pp. 773–774).

Balaneshin-kordan, S., Kotov, A., & Xisto, R. (2015). WSU-IR at TREC 2015 clinical decision support track: Joint weighting of explicit and latent medical query concepts from diverse sources. In *Proc. 2015 Text ..., 3625593*.

Cleverdon, C. W. (1960). ASLIB Cranfield research project – Report on the first stage of an investigation into the comparative efficiency of indexing systems. *Aslib Journal of Information Management*, *12*(12), 421–431.

Hersh, W. (1994). Relevance and retrieval evaluation – Perspectives from medicine. *Journal of the American Society for Information Science*, *45*(3), 201–206. https://doi.org/10.1002/(SICI)1097-4571(199404)45:3<201::AID-ASI9>3.0.CO;2-W

Hersh, W., Crabtree, M. K., Hickam, D. H., Sacherek, L., Friedman, C. P., Tidmarsh, 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. K., Hickam, D. H., Sacherek, L., Rose, L., & Friedman, C. P. (2000). Factors associated with successful answering of clinical questions using an information retrieval system. *Bulletin of the Medical Library Association*, *88*(4), 323–331.

Hersh, W., Pentecost, J., & Hickam, D. (1996). A task-oriented approach to information retrieval evaluation. *Journal of the American Society for Information Science*, *47*(1), 50–56. https://doi.org/10.1002/(SICI)1097-4571(199601)47:1<50::AID-ASI5>3.0.CO;2-1

Hersh, W., Turpin, A., Price, S., Chan, B., Kramer, D., Sacherek, L., & Olson, D. (2000). Do batch and user evaluations give the same results? In *Proc. 23rd Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr*. (pp. 17–24).

Jimmy, J., Zuccon, G., Koopman, B., & Demartini, G. (2019). Health cards for consumer health search. In *Proc. 42nd Int. ACM SIGIR Conf. Res. Dev. Inf. Retr*. (pp. 35–44).

Marshall, J. G. (1992). The impact of the hospital library on clinical decision making: The Rochester study. *Bulletin of the Medical Library Association*, *80*(2), 169–178.

Marshall, J. G., Sollenberger, J., Easterby-Gannett, S., Morgan, L. K., Klem, M. L., Cavanaugh, S. K., ... Hunter, S. (2013). The value of library and information services in patient care: results of a multisite study. *Journal of the Medical Library Association: JMLA*, *101*(1), 38.

McKibbon, K. A., & Fridsma, D. B. (2006). Effectiveness of clinician-selected electronic information resources for answering primary care physicians' information needs. *Journal of the American Medical Informatics Association*, *13*(6), 653–659.

Sadeghi-Bazargani, H., Tabrizi, J. S., & Azami-Aghdash, S. (2014). Barriers to evidence-based medicine: A systematic review. *Journal of Evaluation in Clinical Practice*, *20*(6), 793–802.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, *26*(6), 321–343.

Schamber, L., Eisenberg, M., & Nilan, M. (1990). A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing and Management: An International Journal*, *26*(6), 755–776.

Simpson, M. S., Voorhees, E., & Hersh, W. (2014a). Overview of the TREC 2014 clinical decision support track. In *Trec 2014* (Vol. 2, pp. 1–7).

Simpson, M. S., Voorhees, E. M., & Hersh, W. (2014b). Overview of the TREC 2015 clinical decision support track. In *Proc. twenty-third text Retr. Conf. TREC 2014* (Vol. 2, pp. 1–8).

Turpin, A., & Hersh, W. (2001). Why batch and user evaluations do not give the same results. In *Proc. 24th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr*. (pp. 225–231).

Turpin, A., & Scholer, F. (2006). User performance versus precision measures for simple search tasks. In *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr*. (pp. 11–18).

van der Vegt, A., Zuccon, G., Koopman, B., & Bruza, P. (2018). A task completion framework to support single-interaction IR research. *Journal of Documentation*, *74*(2), 289–308. https://doi.org/10.1108/JD-09-2017-0128

van der Vegt, A., Zuccon, G., Koopman, B., & Deacon, A. (2019). Impact of a search engine on clinical decisions under time and system effectiveness constraints: Research protocol. *JMIR Research Protocols*, *8*(5), e12803.

Voorhees, E., & Harman, D. (2005). *TREC: Experiment and evaluation in information retrieval*. Cambridge, MA: MIT Press.

Westbrook, J. I., Coiera, E. W., & Gosling, A. S. (2005). Do online information retrieval systems help experienced clinicians answer clinical questions? *Journal of the American Medical Informatics Association*, *12*(3), 315–321.

How to cite this article: van der Vegt A, Zuccon G, Koopman B. Do better search engines really equate to better clinical decisions? If not, why not? *J Assoc Inf Sci Technol*. 2020;1–15. https://doi.org/10.1002/asi.24398