Search Interfaces for Biomedical Searching

How do Gaze, User Perception, Search Behaviour and Search Performance Relate?

YING-HSANG LIU, Department of Archivistics, Library and Information Science, Oslo Metropolitan University, Norway

PAUL THOMAS, Microsoft, Australia

TOM GEDEON, Optus Centre for AI, Curtin University, Australia

NICOLAY RUSNACHENKO, Bauman Moscow State Technical University, Russia

The objective of this controlled information retrieval (IR) user experiment is to gain an understanding of domain experts' interactions with novel search interfaces within the context of biomedical information search, with a goal of better search interface design. In this paper, we examine the relationships among user perception, gaze and search behaviour and user search performance. An eye-tracking study of biomedical domain experts' interactions with novel search interfaces was conducted. A total of thirty-two users participated and searched for documents answering eight complex exploratory search tasks, using four different search interfaces. The findings suggest that gaze behaviour in terms of fixation durations based measures of areas of interest (AOI), i.e., visual attention to the elements of title, author, abstract and MeSH (Medical Subject Headings) terms in document surrogates is correlated with search performance. Users are more likely to achieve better search performance by precision-based measures when 1) search tasks are perceived as difficult; 2) users attend to the element of abstract; and 3) users can recall using the per-query suggestions during the search processes. More importantly, our findings suggest that a user search interface design that displays contextual information between the suggested keywords and the document may better support users reformulating their queries for complex search tasks in the biomedical domain. We discuss implications for the design of search user interfaces for biomedical searching.

CCS Concepts: • **Information systems** → **Query suggestion**; **Search interfaces**; *Presentation of retrieval results*; • **Human-centered computing** → *User models.*

Additional Key Words and Phrases: Query suggestion, Search interfaces, Exploratory search, MeSH (Medical Subject Headings), Session-based evaluation

ACM Reference Format:

Ying-Hsang Liu, Paul Thomas, Tom Gedeon, and Nicolay Rusnachenko. 2022. Search Interfaces for Biomedical Searching: How do Gaze, User Perception, Search Behaviour and Search Performance Relate?. In *ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '22), March 14–18, 2022, Regensburg, Germany*. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3498366.3505769

1 INTRODUCTION

The significance of user characteristics in user interactions with information retrieval (IR) systems is generally recognized, but current IR systems are primarily designed for one style: specified search [e.g., 2, 17, 35, 50]. A user-centred approach to interface design that takes into account individual differences, search goals and tasks, has the potential to support users interacting with IR systems more efficiently and effectively. More importantly, the search user interfaces and results from current IR systems have not been optimized for domain experts by considering their domain expertise and querying behaviour. While ordinary searchers may be satisfied with the search results from short queries, effective search in specialized domains requires the domain knowledge and ability to formulate complex queries partly because of the technical nature and large sizes of biomedical information resources [41, 57].

Controlled vocabularies such as MeSH (Medical Subject Headings) have been extensively used to organize information resources in the biomedical domain. The practice of using controlled vocabularies costs millions of dollars (e.g., the use of MeSH terms by the US National Library of Medicine); however, the usefulness of these terms for information access has not been rigorously evaluated in interactive search environments [cf. 38]. Current search systems (e.g., PubMed[1] and MEDLINE based on MeSH) use various retrieval techniques (e.g., suggested term mapping and query expansion) to map user queries to potentially relevant documents. For example, in PubMed the user query "covid 19" is automatically mapped to several MeSH terms, such as "COVID-19", "COVID-19 Vaccines", "covid-19 testing" and "sars-cov-2". Yet automatic query expansion using MeSH terms and their weighting still have high variability for search effectiveness [60]. More importantly, user studies have suggested that domain experts will benefit the most from MeSH terms but that they also need search tools to support their reformulation of queries [38].

This paper reports the findings from a larger research project that investigates domain experts' interaction with novel search user interfaces by observing eye gaze, user perception, search behaviour and search performance when searching clinical search topics. In this study the search tasks are exploratory since they involve learning and investigation activities [42] for open-ended information problems. The effect of user characteristics, exemplified by domain knowledge, search experience and cognitive style on gaze behaviour and the relationship between eye gaze patterns and search behaviour in uncertain search environments have been reported in previous publications [37, 59].

In this paper, we focus on identifying the relationships among gaze, user perception, search behaviour and search performance, with particular references to user models and implications for gaze-based search interface design. To that end, our specific research questions are as follows:

- Where do people look when searching complex questions in biomedical searching?
- How do gaze, user perception, search behaviour and search performance relate?

This study contributes to our enhanced understanding of the relationships among user perception, gaze, search behaviour and search performance, with particular references to the design of search interfaces for biomedical searching.

2 RELATED WORK

2.1 User characteristics and perception in search interfaces and IR systems

From the cognitive perspective, IR researchers have been concerned with the usability and usefulness of search interfaces or system features that are designed to support various kinds of IR tasks, particularly query formulation and reformulation tasks [e.g., 3, 25]. Interactive IR research has focused on these search tasks, partly because query formulation/reformulation tasks are considered complex cognitive processes and users need further support during search processes [e.g., 23, 58, 59].

Some studies have revealed that the user characteristics of search expertise and domain knowledge affect user perception in search interfaces and system features in uncertain search environments. For example, domain experts found the term suggestion feature for unfamiliar search tasks [57]. Doctors were able to use the proposed semantic components for structuring queries and are likely to produce highly relevant documents [41]. Searchers with less search

[1] http://www.ncbi.nlm.nih.gov/pubmed

expertise used the query suggestion feature more frequently for search topics perceived as difficult [45]. Searchers without a medical background did not pay attention to the search interface designed for the exploration of the relationships among the retrieved documents when performing health information tasks [61].

Overall, these studies suggest that the usefulness of search features in support of various IR tasks depends on searchers' domain knowledge and search experience, as well as the perceived difficulty of or familiarity with search tasks. However, it's not certain in which contexts users pay attention to these system features, and whether the use of these systems features contribute to better search performance and efficiency. Further, searchers' levels of domain knowledge and expertise in user studies have rarely been measured by standardized tests [46] or user comprehension of documents [38]. More research on the design of search interfaces using eye-tracking techniques, as a window to the user cognitive processes, to study user characteristics and visual search behaviour in more detail is needed.

2.2 Eye gaze and search interface design

Recent human-computer interaction (HCI) and IR research has focused on users' cognitive aspects in search interactions by measuring the gaze patterns, an indicator of searcher attention and cognitive processes [e.g., 11, 13, 39, 49]. The use of eye-tracking equipment for capturing searchers' fixation patterns provided a rich set of data to understand whether searchers read document surrogates (e.g. summary and metadata) and more importantly, how searchers attend to different elements of search interfaces or search results [e.g., 30, 32].

For example, in a study that compared search interfaces with visible and collapsible facets [30], no significant difference was found in the user's areas of interest (AOI) on the facets panel. In another study of user interactions with a faceted search interface, users spent the most time looking at the search result items in which no distinction was made between title and abstract in terms of AOI [34]. Similarly, there's no significant effect of the interface (list vs. tabular) in total fixation time on search results [28].

Even though the elements of search results or search interfaces are characterized in different ways for research purposes, research suggests that users pay more attention to the elements related to the contents of search results pages than others such as search suggestions and URLs [e.g., 10, 13, 34]. Users' attention to the snippets of web pages increased when the length expanded [10]. In web search environments users paid more attention to top 3, next top 3 and top advertisements than other regions, such as related searches on the search engine results page (SERP) [13], while few abstracts in SERPs from Google and Yahoo were viewed in query reformulation [39]. However, the title of lower-ranked items was considered more important than the snippets of higher-ranked items [54].

These studies generally suggest that there's no significant difference in users' gaze on comparisons of search interface layouts, and users' attention to elements of interfaces depends on the length and quality of snippets on SERPs, as well as the characterization and, displayed position of AOI in search interfaces.

The modelling of user search behaviour using eye-tracking has been concerned with levels of domain knowledge, user interests, inference of search task types and relevance judgment [1, 4, 8, 22]. However, there is limited understanding of the effect of individual differences in user perceptions and patterns of gaze for the design of search interfaces in support of specific IR tasks, such as query formulation and reformulation [cf. 14] for exploratory and complex search tasks. Recent studies have extended this thread of research to building computational user models for predicting search success in information visualization tasks [56], perceived relevance of documents [4], as well as examining the relationship between eye gaze and work roles when users interact with textual and numerical information within the interfaces of a data-driven persona system [50]. Overall, more research on the understanding of user cognitive processes in support of specific IR tasks, and the integration of eye gaze data and search behaviour for predicting user search performance is needed.

3 METHODS

This is a user-centred eye-tracking study of gaze, user perception, search behaviour and search performance for exploratory search tasks in the biomedical domain, with particular reference to the user's attention to and use of the document surrogates. The study has been approved by the Human Research Ethics Committees at the Australian National University and at Charles Sturt University.

3.1 Participants

A total of 32 people participated in the study. Genders were balanced and most participants were students (13 undergraduate and 15 postgraduate), and young (19 were aged between 18 and 24, and 9 were between 25 and 34). A majority (27 of 32) had not used MeSH, but most had substantial experience using search engines (half reported daily use and 12 reported use search engines, such as Google and Bing several times a day or more). The participants had background knowledge in the domains of biology, biotechnology, medical science, neuroscience and biomedical engineering, or some knowledge of biology in their prior learning.

3.2 Experimental design

We used a $4 \times 4 \times 2$ factorial design with four search interfaces, controlled search topic pairs and cognitive styles (see Liu et al. [37] for more details about cognitive styles). A 4×4 Graeco-Latin square design was used [18, 33] to arrange the experimental conditions. Given the interest in the fixed, main effects and interactions of analysis of variance (ANOVA), a medium effect size of .25, $\alpha < .05$ and total sample size of 256 (32 subjects × 8 topics) provides very good statistical power ($\beta = .07$) [7, 15].

3.3 Search interfaces

Participants searched on four different search interfaces, with a single search system behind the scenes. The four search interfaces were distinguished by whether MeSH terms were presented and how the displayed MeSH terms were generated:

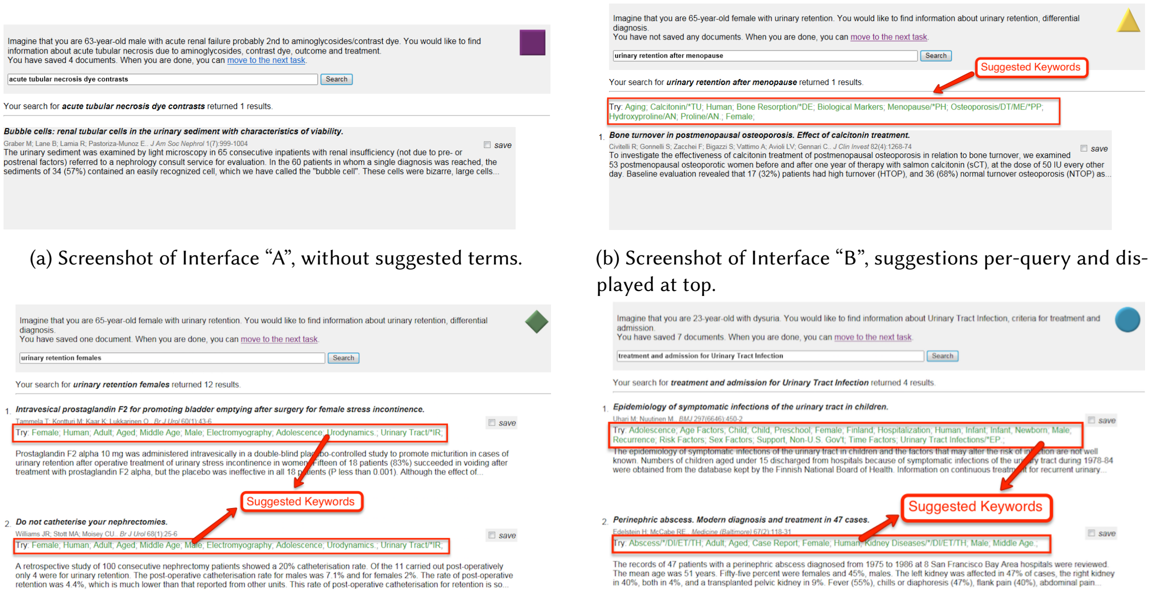
Interface "A" (Figure 1a) mimicked web search and other search systems with no controlled vocabulary. This interface had a brief task description at the top; a conventional search box and button; and each result was represented with its title, authors, publication details, and abstract where available.

The full text was not available, so the results were not clickable. Users judged their success on the titles and abstracts alone.

Interface "B" (Figure 1b) added MeSH terms to the interface. After the user's query was run, MeSH terms from all results were collated; the ten most frequent were displayed at the top of the screen. This mimics the per-query suggestions produced by systems like ProQuest[2].

MeSH terms were introduced with "Try:" and were clickable: if a user clicked a term, their query was refined to include the MeSH term and then re-run.

[2] For example, see http://www.proquest.co.uk/en-UK/products/brands/pl_pq.shtml



Imagine that you are 63-year-old male with acute renal failure probably 2nd to aminoglycosides/contrast dye. You would like to find information about acute tubular necrosis due to aminoglycosides, contrast dye, outcome and treatment.
You have saved 4 documents. When you are done, you can move to the next task.

acute tubular necrosis dye contrasts

Your search for ***acute tubular necrosis dye contrasts*** returned 1 results.

Bubble cells: renal tubular cells in the urinary sediment with characteristics of viability.
Graber M; Lane B; Lamia R; Pastoriza-Munoz E.. *J Am Soc Nephrol* 1(7):999-1004
The urinary sediment was examined by light microscopy in 65 consecutive inpatients with renal insufficiency (not due to pre- or postrenal factors) referred to a nephrology consult service for evaluation. In the 60 patients in whom a single diagnosis was reached, the sediments of 34 (57%) contained an easily recognized cell, which we have called the "bubble cell". These cells were bizarre, large cells...

(a) Screenshot of Interface “A”, without suggested terms.

Imagine that you are 65-year-old female with urinary retention. You would like to find information about urinary retention, differential diagnosis.
You have not saved any documents. When you are done, you can move to the next task.

urinary retention after menopause

Your search for ***urinary retention after menopause*** returned 1 results.

Try: Aging; Calcitonin/*TU; Human; Bone Resorption/*DE; Biological Markers; Menopause/*PH; Osteoporosis/DT/ME/*PP; Hydroxyproline/AN; Proline/AN.; Female;

1. ***Bone turnover in postmenopausal osteoporosis. Effect of calcitonin treatment.***
Civitelli R; Gonnelli S; Zacchei F; Bigazzi S; Vattimo A; Avioli LV; Gennari C.. *J Clin Invest* 82(4):1268-74
To investigate the effectiveness of calcitonin treatment of postmenopausal osteoporosis in relation to bone turnover, we examined 53 postmenopausal osteoporotic women before and after one year of therapy with salmon calcitonin (sCT), at the dose of 50 IU every other day. Baseline evaluation revealed that 17 (32%) patients had high turnover (HTOP), and 36 (68%) normal turnover osteoporosis (NTOP) as...

(b) Screenshot of Interface “B”, suggestions per-query and displayed at top.

Imagine that you are 65-year-old female with urinary retention. You would like to find information about urinary retention, differential diagnosis.
You have saved one document. When you are done, you can move to the next task.

urinary retention females

Your search for ***urinary retention females*** returned 12 results.

1. ***Intravesical prostaglandin F2 for promoting bladder emptying after surgery for female stress incontinence.***
Tammela T; Kontturi M; Kaar K; Lukkarinen O.. *Br J Urol* 60(1):43-6
Try: Female; Human; Adult; Aged; Middle Age; Male; Electromyography; Adolescence; Urodynamics.; Urinary Tract/*IR;
Prostaglandin F2 alpha 10 mg was administered intravesically in a double-blind placebo-controlled study to promote micturition in cases of urinary retention after operative treatment of urinary stress incontinence in women. Fifteen of 18 patients (83%) succeeded in voiding after treatment with prostaglandin F2 alpha, but the placebo was ineffective in all 18 patients (P less than 0.001). Although the effect of...

2. ***Do not catheterise your nephrectomies.***
Williams JR; Stott MA; Moisey CU.. *Br J Urol* 68(1):25-6
Try: Female; Human; Adult; Aged; Middle Age; Male; Electromyography; Adolescence; Urodynamics.; Urinary Tract/*IR;
A retrospective study of 100 consecutive nephrectomy patients showed a 20% catheterisation rate. Of the 11 carried out post-operatively only 4 were for urinary retention. The post-operative catheterisation rate for males was 7.1% and for females 2%. The rate of post-operative retention was 4.4%, which is much lower than that reported from other units. This rate of post-operative catheterisation for retention is so...

(c) Screenshot of Interface “C”, suggestions per-query and displayed alongside each document.

Imagine that you are 23-year-old with dysuria. You would like to find information about Urinary Tract Infection, criteria for treatment and admission.
You have saved 7 documents. When you are done, you can move to the next task.

treatment and admission for Urinary Tract Infection

Your search for ***treatment and admission for Urinary Tract Infection*** returned 4 results.

1. ***Epidemiology of symptomatic infections of the urinary tract in children.***
Uhari M; Nuutinen M.. *BMJ* 297(6646):450-2
Try: Adolescence; Age Factors; Child; Child, Preschool; Female; Finland; Hospitalization; Human; Infant; Infant, Newborn; Male; Recurrence; Risk Factors; Sex Factors; Support, Non-U.S. Gov't; Time Factors; Urinary Tract Infections/*EP.;
The epidemiology of symptomatic infections of the urinary tract in children and the factors that may alter the risk of infection are not well known. Numbers of children aged under 15 discharged from hospitals because of symptomatic infections of the urinary tract during 1978-84 were obtained from the database kept by the Finnish National Board of Health. Information on continuous treatment for recurrent urinary...

2. ***Perinephric abscess. Modern diagnosis and treatment in 47 cases.***
Edelstein H; McCabe RE.. *Medicine (Baltimore)* 67(2):118-31
Try: Abscess/*/DI/ET/TH; Adult; Aged; Case Report; Female; Human; Kidney Diseases/*/DI/ET/TH; Male; Middle Age.;
The records of 47 patients with a perinephric abscess diagnosed from 1975 to 1986 at 8 San Francisco Bay Area hospitals were reviewed. The mean age was 51 years. Fifty-five percent were females and 45%, males. The left kidney was affected in 47% of cases, the right kidney in 40%, both in 4%, and a transplanted pelvic kidney in 9%. Fever (55%), chills or diaphoresis (47%), flank pain (40%), abdominal pain...

(d) Screenshot of Interface “D”, suggestions per-document and displayed with the document.

Fig. 1. Four search interfaces developed in the study.

Interface “C” (Figure 1c) used the same MeSH terms as “B” but displayed them alongside each document, where they may have been more (or less) visible. It is a hybrid of Interfaces “B” and “D” for comparative purposes. There is no difference in the layout of Interfaces “C” and “D”. The difference lies in how the suggestions are generated: per-query suggestions in Interface “C” and per-document suggestions in Interface “D”.

Interface “D” mimicked EBSCOhost[3] and similar systems that provide indexing terms alongside each document. As well as the standard elements from Interface “A”, Interface “D” displayed the MeSH terms associated with each document, as part of that document’s surrogate (Figure 1d).

Each interface was labelled with a simple figure—square, circle, diamond, or triangle—which was referred to in the exit questionnaire.

3.4 Search topics

Search topics used here were a subset of the clinical topics from OHSUMED [24], created for batch-mode IR system evaluation. The topics were slightly rewritten so they read as instructions to the participants (See Figure 2 for an example and Appendix for eight selected topics).

Topics were selected to cover a range of difficulties with topics sorted according to the number of judged relevant documents with two topics selected at random, from each quartile. These eight topics were then randomly paired off to produce four pairs of topics. A final topic (See Figure 2), the same for all participants, was used for training.

[3] http://www.ebscohost.com/

Imagine that you are 63-year-old male with acute renal failure probably 2nd to aminoglycosides/contrast dye.

You would like to find information about acute tubular necrosis due to aminoglycosides, contrast dye, outcome and treatment.

Fig. 2. An example of OHSUMED search topic, reworded for the participants.

3.5 Procedure

Participants were given brief instructions about the search task and system features, followed by a practice topic and then the searches proper. They were informed the test collection is incomplete and out-of-date, an agreed condition for the dataset reuse, since the OHSUMED test collection was used [24], with MEDLINE data from 1987 to 1991. User interaction data recorded included: all queries, mouse clicks, retrieved and saved documents, time spent, and eye movements.

Background and exit questionnaires collected demographic information and asked participants about their perception of the search process. Participants' opinions of the tasks and the interfaces were sought. In particular, users were asked: (1) How difficult was the search task? (2) How useful was the system in completing the search task? (3) whether you noticed the keywords suggested by the systems? and (4) whether you used the keywords suggested by the systems? Finally, information on participants' cognitive styles was collected by a computerized test [47], which took a further 15 minutes to complete.

3.6 Hardware and software

The search system was built on Solr[4], which uses the Boolean logic in query specifications. The search results were ranked by default relevance score, based on the vector space model with the TF-IDF weighting [51]. The MeSH terms were not specifically weighted.

Eye gaze data was recorded from two Sony VFCB-EX480B infrared (IR) cameras which are controlled by Seeing Machines FaceLab 4.5 software[5] and attached to a dedicated machine running Windows 7. At the same time, EyeWorks Design and EyeWorks Record[6] were used to present instructions for the corresponding search tasks during the experiment. Gaze points were recorded at 60Hz, and the eye gaze data included the x and y coordinates of where the eye was looking on the screen, as well as the time that gaze point is recorded. A Windows 7 computer was dedicated to the cognitive styles test [47].

3.7 Data analysis

Where do people look when searching complex questions in biomedical searching? Recordings were analyzed to see how often there were fixations in different parts of document surrogates (i.e., different elements of the interfaces), and therefore how often people looked at each part in particular interfaces.

Four common areas of interest (AOI) were specified: title, author, abstract and MeSH (except for Interface A, without MeSH) to investigate which elements received attention. EyeWorks Analyze was used to specify the AOI and fixations were specified as gazes within a 5-pixel radius which lasted at least 75 ms [43]. We used the metric of fixation durations

[4] http://lucene.apache.org/solr/
[5] http://www.seeingmachines.com/product/facelab/
[6] http://www.eyetracking.com/Software/EyeWorks

per AOI as an indicator of user attention or interest since it has been extensively used for usability research and user interface design [e.g., 21, 39].

3.8 User perceptions

We used a logarithmic cross-ratio analysis [19, 53] to examine user perception of search task difficulty, system usefulness, and notice/use of keywords (MeSH terms) and their relationship to gaze, search behaviour and search performance. The correlation effect size *r* was calculated [7, 40] when the relationship is statistically significant.

3.9 Evaluation metrics

We measured participants' domain knowledge in biology as the number of undergraduate and postgraduate classes taken. The mean values (and therefore cut-points) were 11.5 and 2.2 courses, respectively.

User self-reported task difficulty and system usefulness variables were binarized into "high" and "low" classes, corresponding to cases above or below the mean; self-reported variables for noticing and using MeSH terms were already binary. Similarly, the time spent looking at each of the four AOI, measures of search behaviour and search performance were also binarized. This gave us sixteen 2 × 2 contingency tables, each with a total of 256 observations, from which to determine the correlations among gaze, user perception, search behaviour and search performance. We had a total of 250 observations for search performance because of user errors in using Boolean syntax.

User search performance was calculated from the search logs. User queries were sent to our experimental IR system. An example of user query *q* was as follows:

(abstract:"autoimmune" OR title:"autoimmune") AND (abstract:"symptoms" OR title:"symptoms")

The relevance judgment was based on the OHSUMED test collection [24][7]. Binary relevance judgments in which definitely and possibly relevant documents were merged were used, partly because not all the search performance measures are based on a graded relevance scale. We used the C implementation of the trec_eval tool[8] for calculating user search performance.

With regards to evaluation metrics for user search performance, we considered the factors of incomplete relevance judgments in the test collection [5, 48], models of user behavior [6, 26, 44], multiple-query sessions [27, 29, 52] and alignment between system and user performance [6, 20]. Since we were concerned about user performance in a search session, we chose the bpref of the best performing query (*bpref_bq*) [9] and the R-Precision of the best performing query (*Rprec_bq*). As suggested in previous research [6, 26], the evaluation measure of normalized Discounted Cumulative Gain (nDCG) better captures the user behaviour data than other measures. The maximum, minimum and mean value of nDCG (namely, *nDCG_max*, *nDCG_min* and *nDCG_mean*) for a query were used to capture user search behaviour. To simulate very patient users in our study, we set the *bq* value of 10 in calculating session-based DCG (sDCG) based on previous research [26, p. 6]. And the measure of sDCG divided by the number of queries (*sDCG/q*) was used to reflect the multiple-query search session and the agreement between system and user performance [6, 27, 52]. Overall, these measures were chosen to capture the dynamics of session-based user search behaviour and user performance in a search session.

[7] https://dmice.ohsu.edu/hersh/ohsumed

[8] https://github.com/usnistgov/trec_eval

4 RESULTS

4.1 Search interfaces and search performance

Our overall results reveal that there was no difference in search performance by interface (Figure 3). To be more specific, there was no statistically significant difference between the search interfaces and the various measures of search performance by the analysis of variance (ANOVA), including bpref_bq ($F(3, 246) = 0.49, p > .05$), Rprec_bq ($F(3, 246) = 0.94, p > .05$), nDCG_max ($F(3, 246) = 0.60, p > .05$), nDCG_min ($F(3, 246) = 1.92, p > .05$), nDCG_mean ($F(3, 246) = 1.50, p > .05$), and sDCG/q ($F(3, 246) = 1.50, p > .05$).

These results suggest that there was no direct relationship between proposed search interfaces and user search performance. That is, the search interface itself did not lead to better search performance. It's therefore worth examining further the relationship between gaze behaviour and search performance.

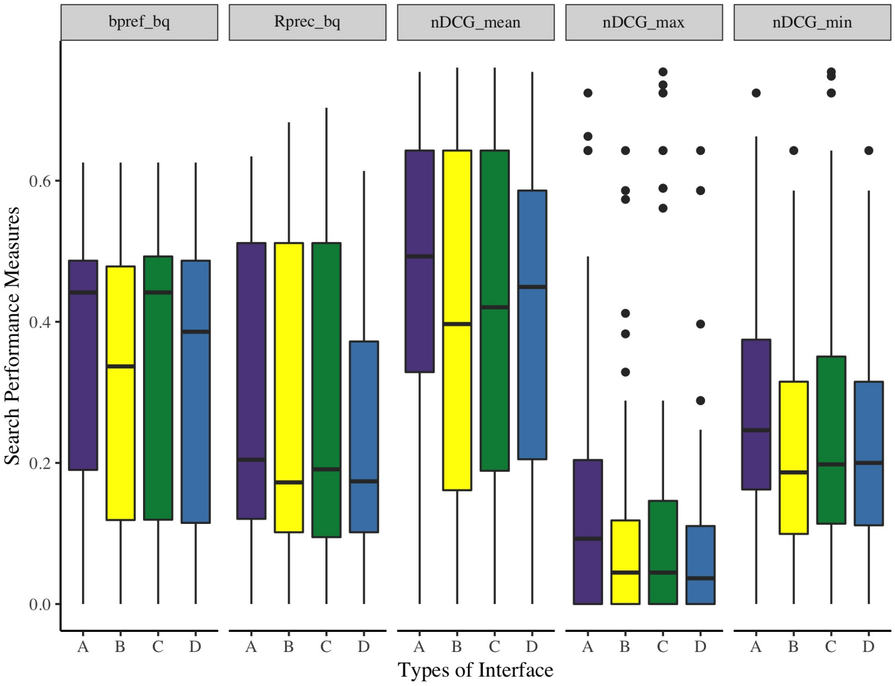


Fig. 3. Search performance by various measures for each interface.

4.2 Gaze and search performance

The results from logarithmic cross-ratio analysis show that there were some significant relationships between gaze and search performance in the elements of abstract and MeSH terms (Table 1). There was a positive relationship between the element of abstract and the search performance in terms of bpref_bq (bpref – one kind of precision measures – of the best query in a search session). The odds of users' attention to the element of abstract are 1.74 higher given the search performance by bpref_bq. In other words, when users attend to the element of abstract, it's 74% (1.74 - 1) more

	CutPoint	*Odds*	*Log*	*Stand.*	*t−*	*Stat.*	*r*
	(Mean)	*Ratio*	*Odds*	*Error*	*Value*	*Signif.*	ES
Title							
bpref_bq	0.34	0.62	-0.48	0.26	-1.88	No	
Rprec_bq	0.26	0.72	-0.33	0.26	-1.27	No	
nDCG_max	0.43	0.64	-0.44	0.25	-1.74	No	
nDCG_min	0.12	1.52	0.42	0.28	1.47	No	
nDCG_mean	0.25	0.94	-0.06	0.25	-0.22	No	
sDCG/q	1.72	1.31	0.27	0.26	1.01	No	
Author							
bpref_bq	0.34	1.09	0.09	0.26	0.32	No	
Rprec_bq	0.26	1.08	0.08	0.26	0.30	No	
nDCG_max	0.43	1.18	0.17	0.25	0.67	No	
nDCG_min	0.12	1.08	0.08	0.28	0.29	No	
nDCG_mean	0.25	1.03	0.03	0.25	0.12	No	
sDCG/q	1.72	1.22	0.20	0.26	0.76	No	
Abstract							
bpref_bq	0.34	1.74	0.56	0.26	**2.15**	**Yes**	**0.15**
Rprec_bq	0.26	1.09	0.09	0.26	0.34	No	
nDCG_max	0.43	1.38	0.32	0.25	1.28	No	
nDCG_min	0.12	1.05	0.05	0.28	0.18	No	
nDCG_mean	0.25	1.40	0.34	0.25	1.32	No	
sDCG/q	1.72	1.24	0.22	0.26	0.82	No	
MeSH							
bpref_bq	0.34	0.81	-0.21	0.26	-0.80	No	
Rprec_bq	0.26	0.78	-0.25	0.26	-0.98	No	
nDCG_max	0.43	0.77	-0.26	0.25	-1.01	No	
nDCG_min	0.12	0.54	-0.61	0.29	**-2.12**	**Yes**	**-0.17**
nDCG_mean	0.25	0.80	-0.22	0.25	-0.86	No	
sDCG/q	1.72	0.58	-0.54	0.27	**-2.02**	**Yes**	**-0.15**

Table 1. Summary of the relationship between gaze and search performance (N gaze = 250, N search performance = 250; statistical significance at 95%).

likely to produce more relevant search results in a search session. Users might be able to extract useful information when they attend to the element of an abstract, and thus contribute to better search performance. The correlation effect size *r* of 0.15 between abstract and bpref_bq was small.

On the other hand, there was a negative relationship between the element of MeSH terms and the search performance measured by nDCG_min (the minimum of nDCG in a search session) and sDCG/q (a measure of the multi-query performance in a session). When users paid attention to the elements of MeSH terms, it was less likely to obtain better search results by 46% (1-0.54) and 42% (1-0.58) for nDCG_min and sDCG/q respectively. This might reveal that users were more likely to attend to the MeSH terms when they were struggling in their searches since both measures are indicators of the overall success within a search session. The correlation effect size *r* of -0.17 between MeSH and nDCG_min, as well as between MeSH and sDCG/q of -0.15 were both small.

Nonetheless, the positive significant correlation between user's attention to the element of abstract and search performance might be useful for the design of system features for the enhancement of search performance. The negative significant correlation between user's attention to MeSH terms and search performance might suggest the struggling situation during the search process, and users might need additional support to use the MeSH terms more effectively.

4.3 User perception and search performance

The results reported in the following subsections include user-perceived search task difficulty, the usefulness of the whole search system, and whether users noticed or used the suggested keywords (i.e., MeSH terms) in search interfaces.

4.3.1 Search task difficulty and system usefulness. Table 2 reveals that there was a very significant relationship between the search task difficulty and the search performance in terms of bpref_bq, but there was no significant relationship between the perceived usefulness of the system and the user search performance. Specifically, when the search tasks were perceived difficult, they were 128% (2.28-1) more likely to obtain better search results in terms of bpref_bq. That is, when search tasks were considered difficult, users' extra efforts were reflected in the measure of bpref, a preference-based measure that considers the incomplete relevance judgements [9]. It might be an indication of motivated searchers in this user experiment. The correlation effect size of 0.20 between search task difficulty and bpref_bq was small.

	CutPoint	*Odds*	*Log*	*Stand.*	*t−*	*Stat.*	*r*
	(Mean)	*Ratio*	*Odds*	*Error*	*Value*	*Signif.*	ES
Search task difficulty							
bpref_bq	0.34	2.28	0.82	0.29	**2.80**	**Yes**	**0.20**
Rprec_bq	0.26	1.65	0.50	0.31	1.63	No	
nDCG_min	0.12	1.62	0.48	0.35	1.39	No	
nDCG_mean	0.25	1.77	0.57	0.30	1.90	No	
nDCG_max	0.43	1.77	0.57	0.29	1.95	No	
sDCG/q	1.72	1.24	0.21	0.31	0.70	No	
System usefulness							
bpref_bq	0.34	1.18	0.16	0.27	0.61	No	
Rprec_bq	0.26	1.18	0.16	0.27	0.61	No	
nDCG_max	0.43	1.14	0.14	0.26	0.51	No	
nDCG_min	0.12	1.42	0.35	0.29	1.22	No	
nDCG_mean	0.25	1.09	0.09	0.26	0.33	No	
sDCG/q	1.72	1.58	0.46	0.27	1.67	No	

Table 2. Summary of the relationship between search task difficulty, system usefulness and search performance (N search task difficulty and system usefulness = 250, N search performance = 250; statistical significance at 95%)

4.3.2 Notice of keywords. Table 3 shows that types of search interfaces make a difference in the relationship between users' notice of suggested keywords and search performance. That is, user perception was affected by the types of search interfaces proposed in the study. Specifically, when users noticed the keywords in Interface B (per-query keywords displayed at the top), they were more likely to obtain worse search results in terms of nDCG_min, nDCG_mean and sDCG/q. There may be quite a few ineffective queries (low nDCG_min) or overall unsuccessful search sessions (low nDCG_mean and sDCG/q). So when users interacted with Interface B, their additional attention to the per-query keywords may suggest that they were struggling in their searches or more contextual information of suggested per-query keywords was needed. Since the MeSH terms were assigned by professional indexers, with complex structures, it would be difficult for users to understand the relationship between the assigned MeSH terms and the retrieved documents. These results are consistent with the relationship between user's attention to the element of MeSH terms and search performance in Table 1.

By contrast, when users noticed the keywords in Interface D (per-document keywords displayed with each document), they were more likely to produce more relevant search results in terms of bpref_bq. The results by other measures were not statistically significant in Interface D. Similar to the potentially relevant terms in abstracts, the suggested keywords displayed with each document have been useful for obtaining more relevant documents. This may be partly because more contextual information could be inferred when MeSH terms were displayed with each document.

The results from Interface C show that the relationship between users' attention to the keywords and user performance was mixed. In line with the findings of Interface D, when users noticed the keywords in Interface C, they were more likely to obtain positive search results in terms of bpref_bq and nDCG_max, but they were more likely to get negative search

	CutPoint	*Odds*	*Log*	*Stand.*	*t−*	*Stat.*	*r*
	(Mean)	*Ratio*	*Odds*	*Error*	*Value*	*Signif.*	*ES*
Interface B							
bpref_bq	0.34	1.31	0.27	0.27	1.02	No	
Rprec_bq	0.26	1.12	0.11	0.27	0.41	No	
nDCG_max	0.43	1.24	0.21	0.26	0.81	No	
nDCG_min	0.12	0.29	-1.23	0.29	**-4.19**	**Yes**	**-0.31**
nDCG_mean	0.25	0.56	-0.58	0.26	**-2.19**	**Yes**	**-0.15**
sDCG/q	1.72	0.37	-0.99	0.27	**-3.62**	**Yes**	**-0.26**
Interface C							
bpref_bq	0.34	1.78	0.58	0.26	**2.25**	**Yes**	**0.16**
Rprec_bq	0.26	1.52	0.42	0.26	1.61	No	
nDCG_max	0.43	1.66	0.51	0.25	**2.01**	**Yes**	**0.14**
nDCG_min	0.12	0.50	-0.69	0.29	**-2.40**	**Yes**	**-0.19**
nDCG_mean	0.25	0.75	-0.28	0.25	-1.11	No	
sDCG/q	1.72	0.51	-0.68	0.27	**-2.54**	**Yes**	**-0.19**
Interface D							
bpref_bq	0.34	1.72	0.54	0.26	**2.10**	**Yes**	**0.15**
Rprec_bq	0.26	1.32	0.28	0.26	1.08	No	
nDCG_max	0.43	1.46	0.38	0.25	1.48	No	
nDCG_min	0.12	1.02	0.02	0.28	0.07	No	
nDCG_mean	0.25	1.06	0.06	0.25	0.25	No	
sDCG/q	1.72	0.68	-0.39	0.26	-1.46	No	

Table 3. Summary of the relationship between notice of keywords in search interfaces and search performance (N notice of keywords in search interfaces = 250, N search performance = 250; statistical significance at 95%)

results in terms of nDCG_min and sDCG/q. The mixed results can be explained by our design of search interfaces (see Figure 1) in which Interface C is a hybrid of Interfaces B and D. The results also validate the experimental design of the study in which the proposed search interfaces have intended effects on user perception.

	CutPoint	*Odds*	*Log*	*Stand.*	*t−*	*Stat.*	*r*
	(Mean)	*Ratio*	*Odds*	*Error*	*Value*	*Signif.*	*ES*
Interface B							
bpref_bq	0.34	1.34	0.29	0.28	1.06	No	
Rprec_bq	0.26	1.11	0.10	0.28	0.37	No	
nDCG_max	0.43	1.16	0.15	0.27	0.55	No	
nDCG_min	0.12	0.71	-0.35	0.31	-1.11	No	
nDCG_mean	0.25	0.74	-0.31	0.27	-1.12	No	
sDCG/q	1.72	0.72	-0.33	0.29	-1.16	No	
Interface C							
bpref_bq	0.34	1.92	0.65	0.28	**2.36**	**Yes**	**0.17**
Rprec_bq	0.26	1.40	0.34	0.27	1.25	No	
nDCG_max	0.43	1.57	0.45	0.27	1.70	No	
nDCG_min	0.12	0.79	-0.23	0.30	-0.77	No	
nDCG_mean	0.25	0.91	-0.09	0.27	-0.35	No	
sDCG/q	1.72	0.65	-0.44	0.28	-1.54	No	
Interface D							
bpref_bq	0.34	1.77	0.57	0.34	1.70	No	
Rprec_bq	0.26	1.19	0.17	0.32	0.54	No	
nDCG_max	0.43	1.46	0.38	0.25	1.48	No	
nDCG_min	0.12	1.46	0.38	0.34	1.10	No	
nDCG_mean	0.25	1.55	0.44	0.32	1.38	No	
sDCG/q	1.72	1.27	0.24	0.33	0.73	No	

Table 4. Summary of the relationship between use of keywords in search interfaces and search performance (N use of keywords in search interfaces = 250, N search performance = 250; statistical significance at 95%).

4.3.3 Use of keywords. Table 4 indicates that there was only one significant relationship between the users' perceived use of suggested keywords in Interface C and the search performance in terms of bpref_bq. It means that when users recognized and used the suggested keywords in Interface C where per-query suggestions were displayed alongside each document, they were by a factor of 1.92 (or 92%) more likely to obtain better search results. That is, when users were able to use the per-query suggestions with less contextual information than the per-document suggestions, they would produce better search results. The correlation effect size of 0.17 between the use of keywords and bpref_bq was small.

As mentioned in § 4.1, there was no significant relationship between gaze behaviour in terms of AOI and search performance. However, we found that there is a significant relationship between users' perception of their attention to the suggested keywords displayed in search interfaces and their search performance. Users would be more attentive to per-document suggestions since they provide contextual information between the MeSH terms and the document. And when searchers were able to recall using the per-query suggestions in Interface C, they were more likely to obtain better search results. Given that EBSCOhost and similar search systems provide indexing terms alongside each document (corresponding to Interfaces C and D), our findings suggest that a user search interface design with contextual information between the suggested keywords and the document may better support users for precision oriented exploratory search tasks in the biomedical domain.

4.4 Search behaviour and search performance

Table 5 shows that there were significant relationships between search behaviour and search performance. To be more specific, the number of queries and typed queries issued, as well as the number of pages viewed in a search session were all negatively associated with search performance by nDCG_min, nDCG_mean and sDCG/q. For instance, when users issued more queries, it was 84% (1-0.16) less likely to obtain better search results in terms of sDCG/q. And when users viewed more pages in search results, it was 66% (1-0.34) less likely to get better sDCG/q scores.

Since the search performance measures of nDCG_mean and sDCG/q are used to summarize the overall performance with multi-query sessions by applying a query discount [e.g., 26, 27, 29], it's not surprising that either more queries or typed queries are correlated with session-based metrics. The low minimum of nDCG in some sessions reveals some unsuccessful searches, which may reflect the fact that users were struggling during the search process, or additional support was needed to make the best use of the suggested keywords. This interpretation is also supported by the findings that the number of pages viewed was negatively correlated with search performance in terms of nDCG_min, nDCG_mean and sDCG/q.

Overall, our proposed search interfaces have intended effects on user perception, gaze and search behaviour by motivated participants in the study. Our results reveal that users will pay attention to the suggested keywords with a search interface design that displays the contextual information between the suggested keywords and the document, similar to the proposed Interfaces C and D. When users can recall using the per-query suggested keywords in Interface C, they are more likely to obtain better search results. The implications for user search interface design, with particular references to the display and method of suggestions, will be discussed in the next section.

What is the relationship between gaze and search performance? In this study, we found that user's attention to the element of abstract is positively correlated with search performance in terms of bpref (one kind of precision measures), while fixation on MeSH terms is negatively correlated with search performance in terms of session-based measures (nDCG_min and sDCG/q). The previous finding also suggested that searchers look at the abstracts more often than other elements of documents in the proposed search interfaces [37]. These results support the hypothesis that search interfaces have significant impact on gaze behaviour when users have complex questions. Our findings have confirmed

	CutPoint	*Odds*	*Log*	*Stand.*	*t−*	*Stat.*	*r*
	(Mean)	*Ratio*	*Odds*	*Error*	*Value*	*Signif.*	*ES*
Number of queries							
bpref_bq	0.34	1.45	0.37	0.26	1.44	No	
Rprec_bq	0.26	0.75	-0.29	0.26	-1.10	No	
nDCG_max	0.43	1.27	0.24	0.25	0.93	No	
nDCG_min	0.12	0.11	-2.20	0.40	**-5.57**	**Yes**	**-0.53**
nDCG_mean	0.25	0.30	-1.20	0.27	**-4.44**	**Yes**	**-0.31**
sDCG/q	1.72	0.16	-1.80	0.32	**-5.71**	**Yes**	**-0.45**
Number of typed queries							
bpref_bq	0.34	0.78	-0.24	0.26	-0.94	No	
Rprec_bq	0.26	1.38	0.32	0.26	1.22	No	
nDCG_max	0.43	1.05	0.05	0.26	0.21	No	
nDCG_min	0.12	0.27	-1.33	0.34	**-3.93**	**Yes**	**-0.34**
nDCG_mean	0.25	0.43	-0.83	0.27	**-3.12**	**Yes**	**-0.22**
sDCG/q	1.72	0.40	-0.93	0.29	**-3.23**	**Yes**	**-0.25**
Number of pages viewed							
bpref_bq	0.34	0.87	-0.14	0.26	-0.53	No	
Rprec_bq	0.26	1.61	0.48	0.27	1.79	No	
nDCG_max	0.43	1.06	0.06	0.26	0.23	No	
nDCG_min	0.12	0.28	-1.27	0.35	**-3.66**	**Yes**	**-0.33**
nDCG_mean	0.25	0.47	-0.74	0.27	**-2.75**	**Yes**	**-0.20**
sDCG/q	1.72	0.34	-1.07	0.30	**-3.57**	**Yes**	**-0.28**

Table 5. Summary of the relationship between search behavior and search performance (N search behavior = 250, N search performance = 250; statistical significance at 95%).

the importance of users' attention to the elements related to subject topics (i.e., abstract of documents or snippets of SERPs) for extracting relevant information when they perform searches for exploratory search tasks [34].

Studies that compared gaze behaviour on search interfaces, such as list vs. tabular and visible vs. collapsible facets [e.g., 28, 30], have indicated that there is no significant difference in terms of fixation measures. Since users' attention to elements of interfaces is affected by, at least, the quality of snippets on SERPs, as well as the region and position of AOI at a micro-level analysis, it's expected that no significant differences can be found for different kinds of search interfaces in terms of fixation measures.

From perspectives of search interface design, eye tracking can be used as an input device, or to infer user cognitive state, such as level of attention [36]. Previous research has also demonstrated attentional mechanisms in the eye movement [12], and identified the relationship between the types of search tasks and the transitions in eye movement [8]. Given the findings that user's attention to specific elements of search interfaces is correlated with search performance in both positive and negative directions, eye tracking data should be interpreted in view of elements of search interfaces and characteristics of search tasks [13, 39]. Nonetheless, computational models based on eye gaze data have demonstrated that it is feasible to use the first ten seconds of interaction data to predict visual search task success in information visualization systems [56]. Future research on whether computational models based on gaze data can successfully predict user search performance in complex biomedical search tasks is suggested.

What is the relationship between user perception and search performance? Our findings reveal that user's notice of suggested keywords in Interface C and D is positively correlated with search performance in terms of one kind of precision measure (bpref of the best query). In both search interfaces, indexing terms are displayed alongside each document (See Figure 1c and 1d). We also found that the relationship between users' perception of the use of keywords in Interface C and search performance is significant. Previous research shows that the perceived quality of index terms affects the use of index terms in online search databases [16, 55]. Further, there is evidence that the perceived usefulness of terms in abstracts is correlated with better search performance by domain experts in terms of the precision measure [38]. As such, one way of supporting query formulation/reformulation tasks for precision-oriented exploratory search in the biomedical domain is to provide quality index terms that are displayed alongside each document in search interfaces. Since Interfaces C and D are distinguished by how the suggested keywords are generated, the perceived quality of index terms and how to provide more contextual information between the assigned keywords and the document deserve further research.

Toward an integration of user perception, gaze, search behaviour and search performance. Based on research findings from this study and the research literature, we can draw several generalizations about the relationships among user perception, gaze, search behaviour and search performance, with particular references to the design of search interfaces for biomedical searching: 1) User perception, search behaviour and gaze behaviour all affect search performance significantly; 2) User perception and search behaviour affect gaze behaviour significantly. For example, types of search interfaces and user-perceived search task difficulty have significant effects on gaze behaviour [37]. Search topic familiarity significantly affects search behaviour in biomedical searching [57]. Perceived usefulness of terms in abstracts is correlated with better search performance by domain experts [38]. And search behaviours, such as issuing queries and MeSH terms that involve notable mental efforts are correlated with changes in eye gaze patterns [59]. As elucidated in a recent review of user search interface design [35], an enhanced understanding of these variables and their relationships in interactive IR studies will shed light on how search interfaces can be designed to better support specific IR tasks.

Limitation of the study. In the IR user experiment, we can observe the user behaviour in detail under controlled environments, with a high level of internal validity. One limitation of the design is that participants are self-selected and they may not be representative of the population. The interaction effects of selection biases and the experimental variable, i.e., search interfaces, are another factor that may limit the generalizability of this study [33]. The re-purposing of clinical search topics and outdated test collection for the interactive search environment may pose threat to external validity, since there may be a mismatch between the searcher's topic knowledge and the test collection. Finally, since we measured user domain knowledge by the number of biology undergraduate and postgraduate classes taken, users' levels of domain expertise may not be sufficient to make the best use of MeSH terms for complex clinical search topics. The ecological relevance of the experimental tasks and the external validity of the study could be enhanced by studies of biomedical professionals [31].

5 CONCLUSIONS

This eye-tracking study of the biomedical domain experts' interactions with novel search interfaces was designed to better understand the relationship among the gaze, user perception, search behaviour and search performance. The findings suggest that user perception, search behaviour and gaze behaviour are all correlated with search performance. Specifically, users are more likely to achieve better search performance by precision-based measures when 1) search tasks are perceived as difficult; 2) users attend to the element of abstract; and 3) users can recall using the per-query suggestions during the search processes. Search behaviours, such as issuing queries and the pages viewed in search engine results page (SERP) that involve expending mental efforts and exploitation of resources are negatively correlated with search performance. These findings suggest that gaze behaviour is affected by both user perception during the search process and search behaviour. Our findings that user's attention to the element of abstract and fixation on suggested keywords of MeSH terms are correlated with search performance has implications for the design of eye gaze-based search user interfaces for biomedical searching.

6 ACKNOWLEDGEMENT

This research project was in part funded by the 2014 ALIA (Australian Library and Information Association) Research Grant Award. The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of ALIA or any author affiliated organisation.

APPENDIX

Eight selected topics

- ID: 4 Imagine that you are 88-year-old with subdural. You would like to find information about reviews on subdurals in elderly.
- ID: 9 Imagine that you are 30-year-old with fever, lymphadenopathy, neurologic changes and rash. You would like to find information about t-cell lymphoma associated with autoimmune symptoms.
- ID: 78 Imagine that you are 42-year-old black man with hypertension. You would like to find information about beta blockers and blacks with hypertension, utility.
- ID: 105 Imagine that you are 68-year-old woman with anemia of chronic illness. You would like to find information about review of anemia of chronic illness.
- ID: 94 Imagine that you are 23-year-old with dysuria. You would like to find information about Urinary Tract Infection, criteria for treatment and admission.
- ID: 47 Imagine that you are 65-year-old female with urinary retention. You would like to find information about urinary retention, differential diagnosis.
- ID: 16 Imagine that you have chronic fatigue syndrome. You would like to find information about chronic fatigue syndrome, management and treatment.
- ID: 58 Imagine that you 65-year-old female with a breast mass. You would like to find information about diagnostic and therapeutic work up of breast mass.

REFERENCES

[1] Oswald Barral and et al. 2015. Exploring peripheral physiology as a predictor of perceived relevance in information retrieval. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI '15)*. ACM, New York, 389–399. https://doi.org/10.1145/2678025.2701389

[2] Nicholas J. Belkin. 2008. Some(what) grand challenges for information retrieval. *SIGIR Forum* 42, 1 (2008), 47–54. https://doi.org/10.1145/1394251.1394261

[3] N. J. Belkin, P. G. Marchetti, and C. Cool. 1993. Braque: design of an interface to support user interaction in information retrieval. *Information Processing & Management* 29, 3 (1993), 325–344. https://doi.org/10.1016/0306-4573(93)90059-m

[4] Nilavra Bhattacharya, Somnath Rakshit, Jacek Gwizdka, and Paul Kogut. 2020. Relevance prediction from eye-movements using semi-interpretable convolutional neural networks. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (Vancouver BC, Canada) *(CHIIR '20)*. Association for Computing Machinery, New York, NY, USA, 223–233. https://doi.org/10.1145/3343413.3377960

[5] Chris Buckley and Ellen M. Voorhees. 2004. Retrieval Evaluation with Incomplete Information. In *Proceedings of the ACM SIGIR Conference (SIGIR '04)*. Association for Computing Machinery, New York, NY, USA, 25–32. https://doi.org/10.1145/1008992.1009000

[6] Ben Carterette. 2011. System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. In *Proceedings of the ACM SIGIR Conference (SIGIR '11)*. Association for Computing Machinery, New York, NY, USA, 903–912. https://doi.org/10.1145/2009916.2010037

[7] Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences*. L. Erlbaum Associates, Hillsdale, NJ. https://doi.org/10.4324/9780203771587

[8] Michael J. Cole, Jacek Gwizdka, Chang Liu, Nicholas J. Belkin, and Xiangmin Zhang. 2013. Inferring user knowledge level from eye movement patterns. *Information Processing & Management* 49, 5 (2013), 1075–1091. https://doi.org/10.1016/j.ipm.2012.08.004

[9] Nick Craswell. 2009. Bpref. In *Encyclopedia of Database Systems*, Ling Liu and M. Tamer Özsu (Eds.). Springer, Boston, MA, 266–267. https://doi.org/10.1007/978-0-387-39940-9{_}489

[10] Edward Cutrell and Zhiwei Guan. 2007. What are you looking for?: An eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI Conference (CHI '07)*, Vol. 25. ACM, New York, 407–416. https://doi.org/10.1145/1240624.1240690

[11] Masoud Davari, Daniel Hienert, Dagmar Kern, and Stefan Dietze. 2020. The role of word-eye-fixations for query term prediction. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (Vancouver BC, Canada) *(CHIIR '20)*. Association for Computing Machinery, New York, NY, USA, 422–426. https://doi.org/10.1145/3343413.3378010

[12] Heiner Deubel and Werner X. Schneider. 1996. Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research* 36, 12 (1996), 1827–1837. https://doi.org/10.1016/0042-6989(95)00294-4

[13] Susan T. Dumais, Georg Buscher, and Edward Cutrell. 2010. Individual Differences in Gaze Patterns for Web Search. In *Proceedings of the Symposium on Information Interaction in Context (IIiX '10)*, Vol. 3. Association for Computing Machinery, New York, NY, USA, 185–194. https://doi.org/10.1145/1840784.1840812

[14] Carsten Eickhoff, Sebastian Dungs, and Vu Tran. 2015. An Eye-Tracking Study of Query Reformulation. In *Proceedings of the ACM SIGIR Conference (SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 13–22. https://doi.org/10.1145/2766462.2767703

[15] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39, 2 (2007), 175–191. https://doi.org/10.3758/bf03193146

[16] Raya Fidel. 1991. Searchers' selection of search keys: Ii. Controlled vocabulary or free-text searching. *Journal of the American Society for Information Science* 42, 7 (1991), 501–514. https://doi.org/10.1002/(sici)1097-4571(199108)42:7<501::aid-asi5>3.0.co;2-v

[17] Raya Fidel. 2012. *Human Information Interaction*. The MIT Press. https://doi.org/10.7551/mitpress/9780262017008.001.0001

[18] Ronald Aylmer Fisher. 1935. *The design of experiments* (9^{th} ed.). Hafner Press, New York. https://doi.org/10.2307/2343406

[19] Joseph L. Fleiss, Bruce A. Levin, and Myunghee Cho Paik. 2003. *Statistical methods for rates and proportions* (3^{rd} ed.). John Wiley, Hoboken, NJ. https://doi.org/10.1002/0471445428

[20] Norbert Fuhr. 2017. Some common mistakes in IR evaluation, and how they can be avoided. *SIGIR Forum* 51, 3 (2017), 32–41. https://doi.org/10.1145/3190580.3190586

[21] Joseph H Goldberg and Xerxes P Kotval. 1999. Computer interface evaluation using eye movements: Methods and constructs. *International Journal of Industrial Ergonomics* 24, 6 (1999), 631–645. https://doi.org/10.1016/s0169-8141(98)00068-7

[22] Jacek Gwizdka. 2014. Characterizing Relevance with Eye-Tracking Measures. In *Proceedings of the Information Interaction in Context Symposium (IIiX '14)*, Vol. 5. Association for Computing Machinery, New York, NY, USA, 58–67. https://doi.org/10.1145/2637002.2637011

[23] Marti Hearst. 2009. *Search user interfaces*. Cambridge University Press, New York. https://doi.org/10.1017/cbo9781139644082

[24] William Hersh, Chris Buckley, T. J. Leone, and David Hickam. 1994. OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research. In *Proceedings of the ACM SIGIR Conference (SIGIR '94)*, Vol. 17. Springer-Verlag, Berlin, Heidelberg, 192–201. https://doi.org/10.1007/978-1-4471-2099-5_20

[25] Peter Ingwersen. 1996. Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation* 52, 1 (1996), 3–50. https://doi.org/10.1108/eb026960

[26] Kalervo Järvelin, Susan L. Price, Lois M. L. Delcambre, and Marianne Lykke Nielsen. 2008. Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions. In *Proceedings of the ECIR '08*. Springer Berlin Heidelberg, Berlin, Heidelberg, 4–15.

[27] Jiepu Jiang and James Allan. 2016. Correlation Between System and User Metrics in a Session. In *Proceedings of the ACM CHIIR Conference* (Carrboro, North Carolina, USA). Association for Computing Machinery, New York, NY, USA, 285–288. https://doi.org/10.1145/2854946.2855005

[28] Yvonne Kammerer and Peter Gerjets. 2012. Effects of search interface and Internet-specific epistemic beliefs on source evaluations during Web search for medical information: An eye-tracking study. *Behaviour & Information Technology* 31, 1 (2012), 83–97. https://doi.org/10.1080/0144929x.2011.599040

[29] Evangelos Kanoulas, Ben Carterette, Paul D. Clough, and Mark Sanderson. 2011. Evaluating Multi-Query Sessions. In *Proceedings of the ACM SIGIR Conference (SIGIR '11)* (Beijing, China). Association for Computing Machinery, New York, NY, USA, 1053–1062. https://doi.org/10.1145/2009916.2010056

[30] Max Kemman, Martijn Kleppe, and Jim Maarseveen. 2013. Eye Tracking the Use of a Collapsible Facets Panel in a Search Interface. *Lect. Notes Comput. Sci.* 8092 (2013), 405–408. https://doi.org/10.1007/978-3-642-40501-3_47

[31] Suzanne Kieffer. 2017. ECOVAL: Ecological validity of cues and representative design in user experience evaluations. *AIS Transactions on Human-Computer Interaction* 9, 2 (2017), 149–172. https://doi.org/10.17705/1thci.00093

[32] Jaewon Kim, Paul Thomas, Ramesh Sankaranarayana, Tom Gedeon, and Hwan-Jin Yoon. 2015. Eye-tracking analysis of user behavior and performance in web search on large and small screens. *Journal of the Association for Information Science and Technology* 66, 3 (2015), 526–544.

[33] Roger Kirk. 2012. *Experimental Design: Procedures for the Behavioral Sciences* (3^{rd} ed.). SAGE. https://doi.org/10.4135/9781483384733

[34] Bill Kules, Robert Capra, Matthew Banta, and Tito Sierra. 2009. What do exploratory searchers look at in a faceted search interface?. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '09)*. ACM, New York, 313–322. https://doi.org/10.1145/1555400.1555452

[35] Chang Liu, Ying-Hsang Liu, Jingjing Liu, and Ralf Bierig. 2021. Search interface design and evaluation. *Foundations and Trends in Information Retrieval* 15, 3-4 (2021), 243–416. https://doi.org/10.1561/1500000073

[36] Ying-Hsang Liu and Ralf Bierig. 2014. A Review of Users' Search Contexts for Lifelogging System Design. In *Proceedings of the Information Interaction in Context Symposium (IIiX '14)*, Vol. 5. Association for Computing Machinery, New York, NY, USA, 271–274. https://doi.org/10.1145/2637002.2637040

[37] Ying-Hsang Liu, Paul Thomas, Marijana Bacic, Tom Gedeon, and Xindi Li. 2017. Natural search user interfaces for complex biomedical search: An eye tracking study. *Journal of the Australian Library and Information Association* 66, 4 (2017), 364–381. https://doi.org/10.1080/24750158.2017.1357915

[38] Ying-Hsang Liu and Nina Wacholder. 2017. Evaluating the impact of MeSH (Medical Subject Headings) terms on different types of searchers. *Information Processing & Management* 53, 4 (2017), 851–870. https://doi.org/10.1016/j.ipm.2017.03.004

[39] Lori Lorigo, Maya Haridasan, Hrönn Brynjarsdóttir, Ling Xia, Thorsten Joachims, Geri Gay, Laura Granka, Fabio Pellacini, and Bing Pan. 2008. Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science & Technology* 59, 7 (2008), 1041–1052. https://doi.org/10.1002/asi.20794

[40] Daniel Lüdecke. 2017. *Esc: Effect size computation for meta analysis*. R package version 0.4.0.

[41] Marianne Lykke, Susan Price, and Lois Delcambre. 2012. How doctors search: A study of query behaviour and the impact on search results. *Information Processing & Management* 48, 6 (2012), 1151–1170. https://doi.org/10.1016/j.ipm.2012.02.006

[42] Gary Marchionini. 2006. Exploratory search: From finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46. https://doi.org/10.1145/1121949.1121979

[43] Sandra P Marshall. 2000. Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity. US Patent 6,090,051.

[44] Alistair Moffat and Justin Zobel. 2008. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM Transactions on Information Systems* 27, 1, Article 2 (2008), 27 pages.

[45] Xi Niu and Diane Kelly. 2014. The Use of Query Suggestions during Information Search. *Information Processing & Management* 50, 1 (2014), 218–234.

[46] Miranda Lee Pao, Suzanne F. Grefsheim, Mel L. Barclay, James O. Woolliscroft, Mark McQuillan, and Barbara L. Shipman. 1993. Factors affecting students' use of MEDLINE. *Computers and Biomedical Research* 26, 6 (1993), 541–555. https://doi.org/10.1006/cbmr.1993.1038

[47] Elizabeth R Peterson, Ian J Deary, and Elizabeth J Austin. 2003. The reliability of Riding's Cognitive Style Analysis test. *Personality and Individual Differences* 34 (2003), 881–891.

[48] Tetsuya Sakai. 2007. Alternatives to Bpref. In *Proceedings of the ACM SIGIR Conference (SIGIR '07)* (Amsterdam, The Netherlands). Association for Computing Machinery, New York, NY, USA, 71–78.

[49] Joni Salminen, Bernard J. Jansen, Jisun An, Soon-Gyo Jung, Lene Nielsen, and Haewoon Kwak. 2018. Fixation and Confusion: Investigating Eye-Tracking Participants' Exposure to Information in Personas. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 110–119. https://doi.org/10.1145/3176349.3176391

[50] Joni Salminen, Ying-Hsang Liu, Sercan Şengün, João M. Santos, Soon-gyo Jung, and Bernard J. Jansen. 2020. The effect of numerical and textual information on visual engagement and perceptions of AI-driven persona interfaces. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 357–368. https://doi.org/10.1145/3377325.3377492

[51] Gerard Salton and Michael J. McGill. 1986. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

[52] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. 2010. Do User Preferences and Evaluation Measures Line Up?. In *Proceedings of the ACM SIGIR Conference (SIGIR '10)* (Geneva, Switzerland). Association for Computing Machinery, New York, NY, USA, 555–562.

[53] Tefko Saracevic, Paul Kantor, Alice Y. Chamis, and Donna Trivison. 1988. A study of information seeking and retrieving. I. Background and methodology. *Journal of the American Society for Information Science* 39, 3 (1988), 161–176.

[54] Denis Savenkov, Pavel Braslavski, and Mikhail Lebedev. 2011. Search Snippet evaluation at Yandex: Lessons learned and future directions. *Lect. Notes Comput. Sci.* 6941 (2011), 14–25.

[55] Ali Shiri and Crawford Revie. 2006. Query expansion behavior within a thesaurus-enhanced search environment: A user-centered evaluation. *Journal of the American Society for Information Science & Technology* 57, 4 (2006), 462–478.

[56] Moritz Spiller, Ying-Hsang Liu, Md Zakir Hossain, Tom Gedeon, Julia Geissler, and Andreas Nürnberger. 2021. Predicting visual search task success from eye gaze data as a basis for user-adaptive information visualization systems. *ACM Transactions on Interactive Intelligent Systems* 11, 2, Article 14 (2021), 25 pages. https://doi.org/10.1145/3446638

[57] Muh-Chyun Tang, Ying-Hsang Liu, and Wan-Ching Wu. 2013. A study of the influence of task familiarity on user behaviors and performance with a MeSH term suggestion interface for PubMed bibliographic search. *International Journal of Medical Informatics* 82, 9 (2013), 832–843. https://doi.org/10.1016/j.ijmedinf.2013.04.005

[58] Nina Wacholder. 2011. Interactive query formulation. *Annual Review of Information Science and Technology* 45 (2011), 157–196.

[59] Peter Wittek, Ying-Hsang Liu, Sándor Darányi, Tom Gedeon, and Ik Soo Lim. 2016. Risk and ambiguity in information seeking: Eye gaze patterns reveal contextual behavior in dealing with uncertainty. *Frontiers in Psychology* 7 (2016), 1790.

[60] Theodore B Wright, David Ball, and William Hersh. 2017. Query expansion using MeSH terms for dataset retrieval: Ohsu at the bioCADDIE 2016 dataset retrieval challenge. *Database* 2017 (2017).

[61] Yan Zhang, Ramona Broussard, Weimao Ke, and Xuemei Gong. 2014. Evaluation of a scatter/gather interface for supporting distinct health information search tasks. *Journal of the Association for Information Science & Technology* 65, 5 (2014), 1028–1041.