Health Cards for Consumer Health Search

Jimmy
University of Queensland, Brisbane, Australia
University of Surabaya (UBAYA), Surabaya, Indonesia
jimmy@uqconnect.edu.au

Guido Zuccon
University of Queensland
Brisbane, Australia
g.zuccon@uq.edu.au

Bevan Koopman
Australian E-Health Research Center, CSIRO
Brisbane, Australia
bevan.koopman@csiro.au

Gianluca Demartini
University of Queensland
Brisbane, Australia
g.demartini@uq.edu.au

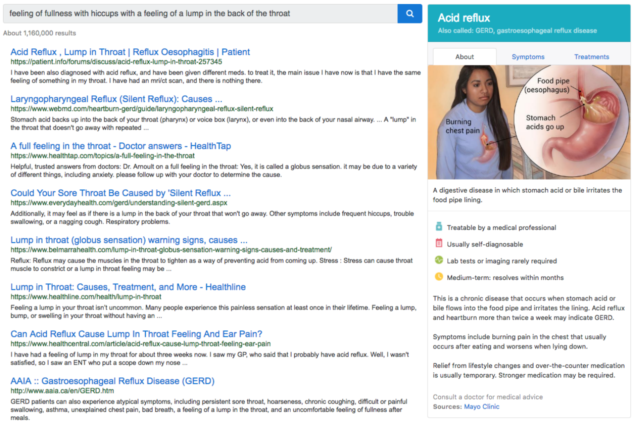
ABSTRACT

This paper investigates the impact of health cards in consumer health search (CHS) — people seeking health advice online. Health cards are a concise presentations of a health concept shown along side search results to specific health queries; they have the potential to convey health information in easily digestible form for the general public. However, little evidence exists on how effective health cards actually are for users when searching health advice online, and whether their effectiveness is limited to specific health search intents. To understand the impact of health cards on CHS, we conducted a laboratory study to observe users completing CHS tasks using two search interface variants: one just with result snippets and one containing both result snippets and health cards. Our study makes the following contributions: (1) it reveals how and when health cards are beneficial to users in completing consumer health search tasks, and (2) it identifies the features of health cards that helped users in completing their tasks. This is the first study that thoroughly investigates the effectiveness of health cards in supporting consumer health search.

ACM Reference Format:
Jimmy, Guido Zuccon, Bevan Koopman, and Gianluca Demartini. 2019. Health Cards for Consumer Health Search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3331184.3331194

1 INTRODUCTION

An entity card is an information object within a Search Engine Result Page (SERP) which contains summarised information about entities associated with the user's query. They are intended to support user search activities by presenting various types of factual information that relate to the user's query in a coherent way [6]. Presenting relevant cards increases user engagement with the search results and reduces the number of queries issued to complete the user's tasks [6]. A specific type of entity cards are the *Health Cards*, which present information around a specific health concept in an enhanced and easily digestible way [26]. Figure 1 shows a health card for "Acid reflux" displayed on the right pane.



feeling of fullness with hiccups with a feeling of a lump in the back of the throat

About 1,160,000 results

Acid Reflux , Lump in Throat | Reflux Oesophagitis | Patient
https://patient.info/forums/discuss/acid-reflux-lump-in-throat-257345
I have been also diagnosed with acid reflux, and have been given different meds. to treat it, the main issue I have now is that I have the same feeling of something in my throat. I have had an mri/ct scan, and there is nothing there.

Laryngopharyngeal Reflux (Silent Reflux): Causes ...
https://www.webmd.com/heartburn-gerd/guide/laryngopharyngeal-reflux-silent-reflux
Stomach acid backs up into the back of your throat (pharynx) or voice box (larynx), or even into the back of your nasal airway. ... A "lump" in the throat that doesn't go away with repeated ...

A full feeling in the throat - Doctor answers - HealthTap
https://www.healthtap.com/topics/a-full-feeling-in-the-throat
Helpful, trusted answers from doctors: Dr. Arnoult on a full feeling in the throat: Yes, it is called a globus sensation. it may be due to a variety of different things, including anxiety. please follow up with your doctor to determine the cause.

Could Your Sore Throat Be Caused by 'Silent Reflux ...
https://www.everydayhealth.com/gerd/understanding-silent-gerd.aspx
Additionally, it may feel as if there is a lump in the back of your throat that won't go away. Other symptoms include frequent hiccups, trouble swallowing, or a nagging cough. Respiratory problems.

Lump in throat (globus sensation) warning signs, causes ...
https://www.belmarrahealth.com/lump-in-throat-globus-sensation-warning-signs-causes-and-treatment/
Reflux: Reflux may cause the muscles in the throat to tighten as a way of preventing acid from coming up. Stress : Stress can cause throat muscle to constrict or a lump in throat feeling may be ...

Lump in Throat: Causes, Treatment, and More - Healthline
https://www.healthline.com/health/lump-in-throat
Feeling a lump in your throat isn't uncommon. Many people experience this painless sensation at least once in their lifetime. Feeling a lump, bump, or swelling in your throat without having an ...

Can Acid Reflux Cause Lump In Throat Feeling And Ear Pain?
https://www.healthcentral.com/article/acid-reflux-cause-lump-throat-feeling-ear-pain
I have had a feeling of lump in my throat for about three weeks now. I saw my GP, who said that I probably have acid reflux. Well, I wasn't satisfied, so I saw an ENT who put a scope down my nose ...

AAIA :: Gastroesophageal Reflux Disease (GERD)
http://www.aaia.ca/en/GERD.htm
GERD patients can also experience atypical symptoms, including persistent sore throat, hoarseness, chronic coughing, difficult or painful swallowing, asthma, unexplained chest pain, bad breath, a feeling of a lump in the throat, and an uncomfortable feeling of fullness after meals.

Acid reflux
Also called: GERD, gastroesophageal reflux disease

A digestive disease in which stomach acid or bile irritates the food pipe lining.

- Treatable by a medical professional
- Usually self-diagnosable
- Lab tests or imaging rarely required
- Medium-term: resolves within months

This is a chronic disease that occurs when stomach acid or bile flows into the food pipe and irritates the lining. Acid reflux and heartburn more than twice a week may indicate GERD.

Symptoms include burning pain in the chest that usually occurs after eating and worsens when lying down.

Relief from lifestyle changes and over-the-counter medication is usually temporary. Stronger medication may be required.

Consult a doctor for medical advice
Sources: Mayo Clinic

Figure 1: User interface for a search scenario, when the SERP is set to display the health card.

This study focuses on the impact of health cards in consumer health search (CHS). CHS is a challenging domain: effective search is hindered by vocabulary mismatch and lack of domain expertise by users; these issues affect both query formulation and result interpretation [36, 39, 40]. The appearance of a health card on a SERP is currently triggered for a number of health related queries issued to major commercial search engines such as Google and Bing. The rationale is that health cards may support users searching for health advice by presenting coherent, understandable and trustworthy health information relevant to the user's query [8].

Are health cards beneficial to CHS users? Are they generally beneficial or only in limited and specific scenarios; e.g., for self-diagnosis v.s. for gathering information about living with a chronic disease? We have already highlighted that CHS is a challenging domain; the factors that make it so may well also impact the use of health cards. In general web search, for example, users are able to accurately discern an entity card's relevance to the query [15]. In CHS, this may not be as easy: even when a health card is relevant, a lack of medical expertise may mean users do not recognise it to be so. For example, when searching information for "feeling of fullness with hiccups with a feeling of a lump in the back of the throat" (query 200 in the CLEF 2018 dataset), a user might not know that the relevant health card for this query is "Acid reflux" and thus, may decide to ignore the important information found in this card.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331194

No previous work has thoroughly investigated the benefits of health cards in consumer health search. In this context, we aim to address the following research questions:

RQ1: Are health cards beneficial to users in completing health search tasks? They are beneficial if they (1) are used as a *source of information* to complete health search tasks, (2) enable users to *correctly* complete health search tasks, (3) reduce the *time* needed to complete the health search tasks, (4) reduce the *effort* required to complete the health search tasks, (5) reduce the user's *perceived workload*, and (6) improve the user's *satisfaction.*

RQ2: How does the benefit vary across search intents? As with RQ1, the same 6 measurements (*source of information*, *time*, *effort*, etc.) are used to measure benefit.

RQ3: What are the health card features that help users? Health cards are composed of a number of features, including the parts of a card (e.g., symptoms & treatments), and the fields of a card part (e.g., a description, the possible treatments, & synonyms of a condition). We considered a health card feature as helpful if it is used to *answer* health search tasks.

To answer these research questions, we conducted a study where 48 participants were presented with 8 CHS scenarios (thus, resulting in 384 interaction data points). Participants were not asked to formulate the query; instead, queries from the CLEF 2018 eHealth collection were used. Participants were left to interact with the SERP (i.e., search result snippets and relevant health cards) and they were asked to collect evidence that helped them solve each CHS scenario. All SERP interactions were recorded and participant's submissions were measured. This was done in a within-subject design for two different search interfaces: the first displaying just the search results and the second displaying both search results and health cards, so that the benefit of using health cards could be measured.

The primary contributions of our study are (1) quantify the impact of using health cards in consumer health search; and (2) identify the features of health cards that helped users in addressing their CHS tasks.

2 RELATED WORK

Health has been recently become an important focus for web search research. Recent work has looked at how to use web search data to identify users suffering from a certain disease [31] and the use of web search query logs [38], blogs [21] or social media data [19] to build models for disease surveillance. Machine learning models have been designed to create keyword search engines over medical literature [22]. In the following, we overview related work in the area of consumer health search which we focus on in our work and on recent research performed on the creation, use, and evaluation of entity cards in SERPS.

2.1 Consumer Health Search

Studies on user experience in CHS show that most users find it difficult to formulate effective queries, to select appropriate results from SERPs and to interpret information within the search results (including discerning whether the health advice is trustworthy/correct) [1, 25, 32, 33, 36, 40]. Query expansion and query reformulation have been found, at times, to be beneficial [11, 24, 30]. For example, expanding CHS queries by adding the correct medical expression related to a query expressed in layman's terms, led to improved retrieval effectiveness and improved completion of health search tasks [1] [30]. However, this may also introduce results that are less familiar to the user and more difficult to understand for non-experts [20]. Another avenue to support CHS users in formulating effective queries is by recommending *alternative* query terms; high quality query recommendations can significantly improve the rates of successful queries issued by CHS users [37]. In this work, we depart from previous attempts that focused on querying aspects; instead, our focus is on the search result appraisal and information acquisition. In particular, we investigate whether the use of health cards could assists users with completing their health search task.

As for problems regarding the discovering and understanding of search results, Alpay et al. [1] suggested that these are caused by the gap between the informational context of the search results and the user's personal context (e.g., lack of medical knowledge). Lau and Coiera [16] and later White [34] further found that people seeking health advice online are affected by all sorts of cognitive biases, including anchoring (prior belief), results presentation/access order effect, and exposure effect (length of time taken to process a result).

To overcome this gap, search technologies need to contextualise the relevant medical information to suit the user's knowledge and awareness about the medical condition/situation they are searching. A number of leading web search engines have taken the initiative to display health cards along with search results when identifying the user has issued a health query. These cards may convey medical information in a context that can be understood by the general public. To evaluate the benefit of health cards, in this study we devised an empirical, user-centred exploration displaying health cards to address various CHS intents.

2.2 Entity Cards

Health cards more generally, and outside the health domain, can be referred to as entity cards or information cards [29]. An *entity card* presents a rich and coherent set of information about a specific entity; this commonly includes the entity's name and type, a textual summary, a factual summary, key features, relationships, and links to related entities [3, 29]. Entity cards are now an integral part of the SERP in commercial search engines like Google, Bing, and Yandex. Studies show their use improves user engagement, attracts user attention, and enhances user experience [3, 6].

An entity card is often displayed as an additional item along with the list of search results and is usually placed in the centre or right pane of a SERP (see Figure 1). The idea of an entity card is somewhat similar in spirit to what was achieved in aggregated/vertical search [2]; i.e., information from different sources and related to different aspects of the query is brought together in the results. However, in aggregated/vertical search, results from different specialised services (e.g., image, video, news, etc.) are blended within the SERP, while an entity card involves the creation of a new information object (the card) which integrates and summarises the information obtained from one or more sources.

Authors of [6] showed how entity cards help users navigate SERPs and summarised the relevant information by influencing

[1] Increased number of relevant results for medical-related queries.

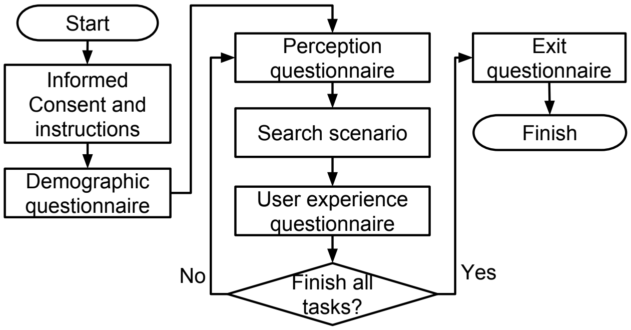


Figure 2: The user study flowchart.

their search behaviour. In [9], authors showed how to automatically generate and how to evaluate the effectiveness of entity cards in SERPs. Compared to this body of work, we perfomed a study focussed on in-lab user studies as compared to crowdsourcing and on CHS as compared to general web search.

A specific specialisation of an entity card is the health card [8]: cards regarding a health-related entity (typically a medical condition, but also tests, treatments, services, etc.). While previous work has shown the impact of entity cards on user experience and overall task effectiveness, to date, there has been no thorough analysis on the effectiveness of health cards, including their impact on search behaviour and task completion when seeking health information or advice online. Our study takes the first steps to address this gap.

3 METHODS

A user study was set up to answer our three research questions. Figure 2 depicts the flowchart of the user study. In a within-subject design, participants were requested to complete eight health scenarios (Section 3.3) using two search interfaces (one with health cards and the other without; detailed in Section 3.6) in a usability laboratory with a PC equipped with eye tracking technology. To minimise bias with fatigue, we rotated the eight scenarios and the two search interfaces using a Graeco-Latin square rotation [13]. Participants were recruited principally amongst a university's population (Section 3.10). The study has received Human Research Ethics Committee clearance (*ref num 2018002115*). The rest of this section details each part of the user study.

3.1 Consent and demographic questionnaire

After consenting to participate, each participant was given a set of instructions presenting the elements of the interface and rules for the collection of evidence to answer the scenarios. Next, a demographic questionnaire collected information on the participant's age group (grouped by ten-years intervals[2]), highest level of completed education, English proficiency[3], and the frequency of use of general-purpose search engines. We used the responses to determine the participant's eligibility, as described in Section 3.10.

3.2 Perception questionnaire

After completion of the demographic questionnaire, participants moved to consider each of the 8 health scenarios assigned to them, one at the time. Before undertaking a scenario in the search interface, participants were presented with the scenario and asked to complete a perception questionnaire.

The perception questionnaire was adapted from Kelly et al. [14] and served to understand the participant's interest and background knowledge on each health scenario. Furthermore, it allowed us to capture the complexity of the scenario, as perceived by participants. Table A in the online appendix[4] lists the perception questionnaire items and the available response options.

3.3 Search scenarios

After completing the perception questionnaire for a scenario, participants were asked to complete the assigned search scenario. While the artificial search scenario may not represent the participants' information need, yet, we selected this approach as this is a common approach (e.g., [14, 18, 23, 27]) which enables control over the experiment conditions and comparison of results across participants [5, 13]. Each scenario consisted of a topic, a task, a *given* user query, the top ten search results for the user query (Section 3.8), and a health card (Section 3.7) (if using the search interface with a heath card). We asked participants to complete the task by copying and pasting relevant evidence from one or many parts of the presented information (i.e. the search results, documents themselves or from the health card) that allowed them to solve the task. This protocol allowed us to track where participants found the relevant evidence needed to solve the search scenario.

Search scenarios were selected from the CLEF 2018 collection [12], a collection used for evaluating search engines tailored to consumer health search. The collection contains 50 topics, each composed of a query issued to the Health-On-the-Net search service[5] (along with other query variations manually derived) and a topic narrative manually created by the organisers of CLEF based on the query.[6]

We selected the scenarios based on the "product" and "task complexity" facets used by Li and Belkin [17, 18]. For the "product" facet, we considered the factual (F) and intellectual (I) values. Factual scenarios consider tasks seeking health information related to a given condition, whereas Intellectual scenarios consider seeking health information based on general observations (i.e. symptoms). For the "task complexity" facet, in line with prior work [18], we considered low complexity (L) as scenarios with only one sub-task and high complexity (H) as scenarios with multiple sub-task.

We combined the values of "product" and "task complexity" facets to produce four search-task types: FL, FH, IL, and IH. We selected two scenarios for each search task type, thus, resulting in eight scenarios in total. Table 1 lists the eight scenarios.

3.4 User experience questionnaire

The user experience questionnaire was used to capture the participants experienced difficulty, perception on system effectiveness, satisfaction and workload. This questionnaire was also adapted

[2] Following the guidelines for age-group data anonymisation from the Australian Bureau of Statistics.

[3] We verified participants English proficiency by checking whether they: (1) speak English as first language, or (2) achieved IELTS overall test score of at least 5.0 with a score of at least 4.5 in each of the four test components. These are the minimum English proficiency to work in Australia.

[4] https://bit.ly/2W8IT0R

[5] https://hon.ch/en/

[6] Note, queries may contain typos (one used in this study does).

Table 1: User study scenarios for each search-task type.

(Topic Id, Type) Topic and Task of the Scenario
(176, FL_1) **Topic**: Your physiotherapist has mentioned you may have pelvic inflammatory disease and suggested you to go to a doctor. **Task**: Find out more information about how this disease can be treated. (195, FL_2) **Topic**: Your son was bitten by a tick and his exams suggest that he has Lyme disease. Before speaking with a doctor, you want to get information on possible treatments for this disease. **Task**: Find out more information on effective treatments for Lyme disease.
(154, FH_1) **Topic**: Your elderly father has just been diagnosed with high blood pressure (HBP). **Task**: Find some information that discusses living with high blood pressure and its effects on daily living, including which food and activities he should avoid. (170, FH_2) **Topic**: You have been diagnosed with rheumatoid arthritis by your doctor. **Task**: Find out more information on this disease and what its likely course is.
(152, IL_1) **Topic**: A colleague from work who was very social suddenly became withdrawn and has shown various mood alterations. You think there might be something wrong with her mental health. **Task**: Find out more information on diseases that might be causing this change in her behaviour. (163, IL_2) **Topic**: You have been feeling a bit anxious recently, and are considering going to a doctor for a consultation. **Task**: Find information on possible strategies for day to day coping with your anxiety problem.
(172, IH_1) **Topic**: Yesterday you noticed that your mum was trembling and quivering. She did not do this on purpose, and when you asked her, she said she felt fine. **Task**: Find out what may have caused this, and whether it is something serious. (200, IH_2) **Topic**: It's few days now that you have been getting hiccups after eating. You felt you eat enough every time, in fact, you felt full. At the same time, you feel something in the back of your throat: like if you had a bump or lump. **Task**: Find out what you may have and when its time to make an appointment with a doctor.

from Kelly et al. [14]. Table B in the online appendix lists the user experience questionnaire items and the available responses.

3.5 Exit questionnaire

After completing all 8 search scenarios, we asked our participants to express their overall experience in completing the tasks and their previous experiences in searching online for health information with specific attention to the use of health cards. Table C in the online appendix shows the questionnaire items and available options in the exit questionnaire.

3.6 Search Interfaces

The search engine result page contained three panes (Figure 1 shows the middle and right panes only). On the left pane, the system displayed the topic, the task, instructions to complete the task and a text box for participants to paste selected evidence. The middle pane showed the query string (disabled so they could not enter a new query) and the top ten search results (title, url, and snippet). A health card was displayed on the right pane when the experimental condition required health cards.

Table 2: For each topic, the health cards displayed and the initial query used to trigger it.

(Topic) Health Card Title	Initial Query
(176) Pelvic inflammatory disease	pelvic inflammatory disease
(195) Lyme disease	affective treatments for chronic lyme disease
(154) High blood pressure	high blood pressure
(170) Rheumatoid arthritis	rheumatoid arthritis prognosis
(152) Clinical depression	emotional and mental disorders
(163) Anxiety disorder	Anxiety coping skills
(172) Essential tremor	involuntary trembling or quivering
(200) Acid reflux	feeling of fullness with hiccups with a feeling of a lump in the back of the throat

We designed the middle and the right panes following the Google SERP. We followed Google as it was the most popular search engine in the country this study took place; thus, participants would be accustomed to the interface.

3.7 Health Cards

Health cards were acquired from the Google search engine. For each scenario, we submitted the initial query from the CLEF 2018 collection to Google. If a heath card was displayed, then we scraped it, including any image and link. If there was no health card, then a physician examined the scenario to determine the target condition relevant to the scenario (also aided by the relevance assessments from CLEF 2018). After examination, the physician provided a diagnosis relevant to the scenario — we then queried Google with the diagnosis and scraped the health card for that diagnosis. Note that, later on in the study, the physician assessed every scenario in a similar manner to determine the health diagnosis for analysis of the results; this confirmed that the health cards acquired through the original query, matched the target diagnosis. Table 2 lists the topic id, health card title and initial query for each scenario.

Each health card contained a title, aliases (i.e., "also called"), if any, an image, a summary tab (i.e., about), a symptoms tab, and a treatments tab. Each tab contained a URL that linked to the source information for the health card. For the health card, "Essential tremor" we found no image in the Google card; thus we obtained the image from the source URL presented in the card. This was done to provide a similar look & feel for all health cards in the study.

3.8 Search Results

The original CLEF 2018 queries for each of the considered scenarios were used to acquire search results. To ensure that the search results were on the same topic as the corresponding health card, we further expanded the query by adding words from the health card's title that were not in the query.

For each query, we retrieved the top ten search results for each extended query using the Bing Web Search API [7] on October 5th,

[7] https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api/

2018. Finally, we archived all search results and source web pages to avoid problems with possible web pages and SERP updates, as noted by Jimmy et al. [10]. When a participant clicked on any link in the interface (either from the results or from the health card), we presented them with the archived web page.

3.9 Capturing Interaction Data

Throughout the user study, we captured participants interactions with the search interfaces using the *Big Brother logging service*[8]. This allowed us to record mouse movements (including anchored to `<div>` containers, e.g., enter and leave the container), clicks, scroll, page loading (start and end), cut/copy/ paste, screen resize (mainly to align and validate eye-tracking data).

In addition, we used the Tobii Pro Spectrum eye tracker to acquire eye gaze data, set to operate at the frequency of 300Hz. The eye tracker was connected to a monitor with a resolution of 1920 x 1080 pixels. The eye tracker was calibrated for each participant at the start of the study using the method described by Blignaut [4]. We used the *velocity-threshold identification* algorithm [4, 28] to identify fixation points. We set the velocity radius threshold to 70 pixels following the size of eye gazing point visualisation from the Tobii Pro Eye Tracker Manager. We set the minimum fixation duration threshold to 700ms following the highest average fixation duration recorded by Diez et al.'s experiments [7]. We selected this fixation duration (as opposed to shorter durations, e.g. 100ms, used in other studies to measure gaze) because we were interested in analysing fixation points when participants were looking with attention for information to complete a scenario: fixation points for such activities are longer than fixation points for other activities that do not require in-depth processing [7]. Then, we mapped the fixation points to three *Area-of-Interest*s (AOIs): scenario description (left pane), list of snippets (middle pane), and health card (right pane, if displayed).

The eye gaze data was used to determine whether participants noticed the health card displayed on the interface, and how much time they spent on the health card, compared to the rest of the SERP or actual result web pages. Other analyses of the collected eye tracking data was regarded as being out of scope of this paper, and is left for future work.

3.10 Participants

The study was advertised widely through the University of Queensland and the Queensland University of Technology, two large public universities in Australia, as well as through Facebook groups mainly tailored to students and alumni of these universities. Note that we did not enforce participants to be university students or affiliates, and we allowed any member of the public to take part in the study. Nevertheless, the majority of the participants were university students.

The following eligibility criteria for participation in the study were set and enforced: aged 18 years or above, no specific prior medical studies, experienced with using a general-purpose search engine on a daily basis, and proficient reading and writing of English. Participants were told that the study would last approximately one hour and were given a $15 gift card for their participation.

Table 3: Perception questionnaire: interest & knowledge.

Task	Interest	Previous search frequency	Previous knowledge
FL_1^a	3.58±0.92	$1.23^{cef} \pm 0.56$	$1.25^{cef} \pm 0.53$
FL_2^b	3.67±0.93	$1.31^{cef} \pm 0.66$	$1.27^{cef} \pm 0.57$
FH_1^c	4.08^{h}±0.82	$1.9^{abgh} \pm 1.08$	$2.23^{abdg} \pm 0.9$
FH_2^d	3.56±1.05	$1.48^{f} \pm 0.71$	$1.58^{cf} \pm 0.68$
IL_1^e	3.88±0.84	$1.81^{abgh} \pm 0.82$	$2.02^{abh} \pm 0.76$
IL_2^f	3.85±0.87	$2.02^{abdg} \pm 1$	$2.17^{abdg} \pm 0.72$
IH_1^g	3.96±0.9	$1.31^{cef} \pm 0.59$	$1.67^{cf} \pm 0.81$
IH_2^h	3.46^{c}±1.01	$1.19^{cef} \pm 0.45$	$1.38^{cef} \pm 0.57$

We suggested a time limit of 60 minutes for the overall experiment but did not enforce it. Participants were allowed to complete a task without successfully identifying any relevant information: this happened on one occasion.

In total, we collected 384 results and interaction data from 48 participants[9] which give us enough power to make statistical analysis (power > 0.90). Each of the sixteen sequences of scenarios-search interface pairs as produced by the Graeco-Latin square rotation was performed by 3 participants. Participants consisted of 27 females and 21 males in the following age groups: 20 between 18-24 y.o., 21 between 25-34 y.o. and 7 between 35-44 y.o.. Participants were from various education backgrounds, with the following highest level of education completed: 8 high school, 5 diploma, 11 bachelor degree, 5 graduate diploma, and 19 postgraduate degree.

4 EXPERIMENTAL RESULTS

In the following, we report the findings for each research question considered in this work. In all experiments, for statistical significance analysis, we used the repeated-measures ANOVA with Bonferroni as follow-up test. In all result tables, superscripts refer to statistical significance between the result and the result associated with the superscript ($p < 0.05$).

4.1 Prior Knowledge, Interest, and Fatigue

We start by analysing our results to identify whether the participants' level of interest, prior knowledge on the scenarios, and fatigue may have had a systematic effect on results.

Table 3 shows that all scenarios were perceived as moderate to highly interesting (Mean (M)=3.76; Standard Deviation (SD)=0.94), although FH_1 and IH_2 were found to be significantly more interesting (FH_1) and less interesting (IH_2), respectively. As noted in Table 3, participants responses in terms of past experience varied significantly across scenarios, however, the past search experience was bound between never to a couple of times (M=1.53; SD=0.81). In terms of prior knowledge on the scenarios, differences across scenarios were significant; however, on average, participants reported to have no or little prior knowledge (M=1.70; SD=0.79).

Then we investigated the participants' level of understanding of the scenarios (Table 4). All scenarios were perceived as moderate to well defined in terms of types of information needed (M=3.80; SD=0.76) and the expected solution (M=3.80; SD=0.77). There are no significant differences between scenarios, with the exception of

[8] https://github.com/hscells/bigbro

[9] Each participant performed 8 scenarios.

Table 4: Perception questionnaire: how defined each task is.

Task	Information needed	Expected solution
FL_1^a	3.79± 0.71	3.81± 0.76
FL_2^b	3.94^e± 0.7	4^e± 0.65
FH_1^c	4.06± 0.56	4.06± 0.63
FH_2^d	3.79± 0.71	3.69± 0.8
IL_1^e	3.48^b± 0.9	3.52^b± 0.85
IL_2^f	3.96± 0.65	3.83± 0.66
IH_1^g	3.65± 0.89	3.73± 0.89
IH_2^h	3.73± 0.82	3.77± 0.75

FL_2 and IL_1 that were significantly different between each other (FL_2 was more defined, while IL_1 was somewhat less defined).

These results indicate that the scenarios were homogeneous in terms of participants interest, prior knowledge, and task definition.

We then turned to investigate participants fatigue by correlating the sequence of scenarios and the results from the six measurements used in RQ1 and RQ2 (defined in Section 1). We found that there is a significant negative correlation between scenario sequence and duration taken to complete a scenario (Pearson=-0.30, p<0.001): this may be due to fatigue or acquired familiarity with task and interfaces. On the contrary, we found no significant correlation between scenario sequence and the other five measurements: health card usage rate (Pearson=-0.03, p=0.54), correctness (Pearson=-0.02, p=0.76), effort (i.e., the number of links opened when completing a scenario) (Pearson=-0.05, p=-0.28), perceived workload (Pearson=-0.02, p=0.66), and perceived satisfaction (Pearson=-0.05, p=0.35). These suggest that the results are comparable across scenario sequences; in addition, the experiment's Graeco-Latin design further mitigates the effect of the possible fatigue or acquired familiarity.

4.2 Analysis of Search Interface

We then analysed the overall user experience after completing all 8 search scenarios as recorded in the exit questionnaire. Regardless of the search interface, participants, on average, agreed or strongly agreed that the system was easy to use (91%), provided useful information (91%), displayed results of similar quality to general-purpose search engines (76%), and were satisfied (87%). When asked about whether they noticed the health cards in our interface, 93% of the participants answered positively. Note that at the start of the experiment, participants were given a set of instructions and a description of the search interface. This included advising the presence of both snippet items and health cards.

4.3 Analysis of Search Behaviour

We analysed search behaviour by evaluating to which AOI (i.e. snippets vs. health cards[10]) participants paid attention to through each scenario (session), when health cards were displayed. Since the time taken by each session varied, we normalised durations, and present results with respect to the progress of the session. Figure 3 shows the percentage of participants that paid attention to each AOI through the session. Overall, we found that the majority of participants spent more time on snippets (M=55.40%) than on health cards (M=44.60%). This is understandable as there is more information in the snippets to process and the display area containing the snippets is larger.

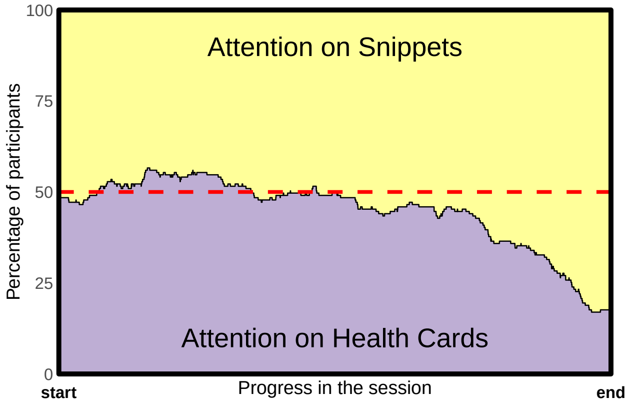


Figure 3: Percentage of participants paying attention to snippets vs. health cards throughout a session. This analysis considers only data obtained when health cards were displayed.

We found a strong negative correlation (Pearson=-0.83) between giving attention to health cards and time in the session (and vice-versa for snippets). That is, participants tend to consider health cards earlier in the session. In particular, 48% of participants start a session by giving attention to the health card vs. 18% end a session on the health card. We speculate that although health cards are consulted and are considered with as much attention as the (probably top) snippets to start with, participants may have felt the health cards did not contain enough information to complete the scenarios, and went on examining snippets throughout the SERP.

4.4 RQ1: The Benefits of Health Cards

As mentioned in Section 1, we considered health cards being of benefit to consumer health search based on six measurements. First, we investigated whether health cards are used as a source of information to complete health search scenarios. Of the 192 scenarios completed with health cards displayed, the majority were completed without selecting health cards as a source of information (51.04%). Of the 48 participants, 35 (72%) selected information from the cards at least once across the four tasks with displayed health cards. These results suggest that most participants *perceived* the health cards as beneficial to complete some of the search scenarios. Nevertheless, overall, the organic search results were *perceived* by the participants as more beneficial than the health cards.

Second, we assessed the selected evidence based on a scoring guide adapted from Wilson and Draney [35] (Table 5) and guidelines from a physicians, when in doubt, we further confirmed individual cases with a physician. We found that the average correctness of the selected evidence did not significantly differ across conditions: no health cards were displayed (M=2.38; SD=1.03), cards were displayed but not selected (M=2.26; SD=1.16), and cards were displayed and selected as a source of information (M=2.54; SD=0.95). Figure 4 depicts these distributions.

[10] We removed eye tracker recordings associated to other display areas.

Table 5: Scoring guide to determine the correctness of selected evidence.

Score	Definition
3	Complete and correct response.
2	Partially correct response missing some minor elements.
1	Contains a small fraction of the expected answer.
0	Contains no correct response.

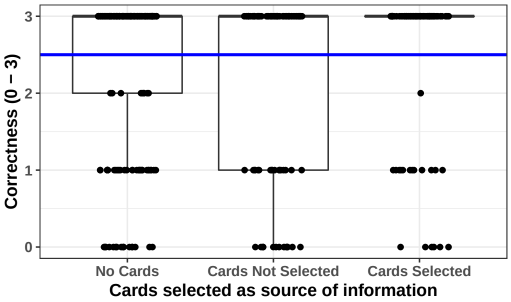


Figure 4: The average correctness of the submitted evidence (higher is better). The horizontal line shows the average correctness that would be achieved if, for all scenarios, participants selected only evidence from the health cards.

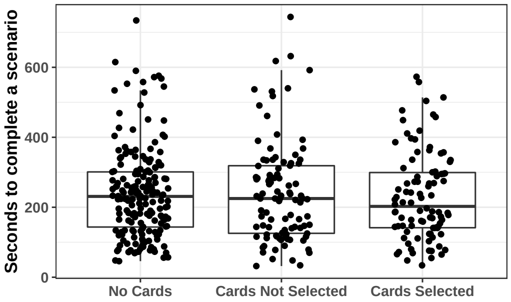


Figure 5: Time required by participants to complete a search scenario. The lower value, the quicker a participant completed the scenario.

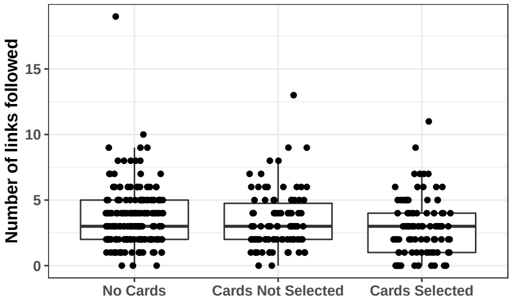


Figure 6: Average effort spent by participants in completing the scenarios, measured as the number of web pages opened (links followed): the lower the less effort was spent. Note that the number of links followed includes both clicks on search results and on links in health cards. Further, participants may have clicked multiple times on the same link.

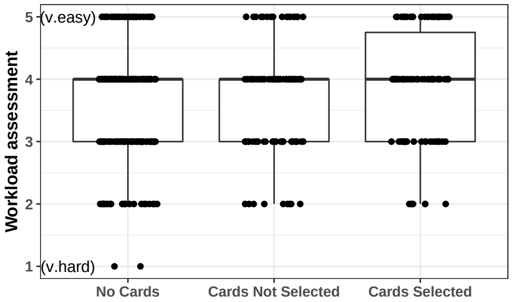


Figure 7: Perceived workload when completing scenarios.

We then compared the correctness of the submitted evidence to the correctness an hypothetical user would have achieved if all scenarios were completed by selecting only information from the health cards. The horizontal line in Figure 4 suggests that most participants performed better than this hypothetical user, by gathering information beyond what displayed in the health cards.

Third, we measured whether health cards reduced the time needed to complete the health scenarios. Figure 5 shows that, on average, there were no significant differences in the amount of time (in seconds) required to complete scenarios in all three conditions: no cards were displayed (M=240s, SD=128), cards were displayed but were not selected (M=242s, SD=145), and finally, cards were displayed and selected (M=231s, SD=126).

Fourth, we measured the effort required to complete the health scenarios. We estimated effort as the number of links followed by participants. Figure 6 shows that, on average, participants spent significantly less effort when selecting information from health cards (M=2.91; SD=2.21) compared to when no health cards were displayed (M=3.61; SD=2.21).

Fifth, we measured the participants' perceived workload after completing the health scenarios. Figure 7 shows that there were no significant differences in the level of perceived workload when completing scenarios in all three conditions: no cards (M=3.59; SD=0.93), displayed cards not selected (M=3.69; SD=0.87) and displayed cards selected (M=3.86; SD=0.89).

Sixth, we compared the participants' overall satisfaction with their submitted evidence. Figure 8 shows that, on average, participants were significantly more satisfied with their submissions when selecting information from health cards (M=3.80; SD=0.91) compared to when no health cards were displayed (M=3.45; SD=1.06).

Finally, we examined the interaction between prior knowledge and the six measurements of benefit. To this aim, we used a repeated-measures ANOVA with Bonferroni as follow up test. We found that there was positive significant interaction ($p < 0.01$) between prior knowledge and correctness of selected evidence, and between prior knowledge and workload. We further analyse the interaction between prior knowledge and correctness with regard to the following conditions: (1) no cards where displayed, (2) cards displayed but not selected, and (3) cards displayed and selected. We found that significant interactions occurred in the first two conditions, but not in the third. This implies that health cards may help bridge the gap between knowledgeable and less knowledgeable users.

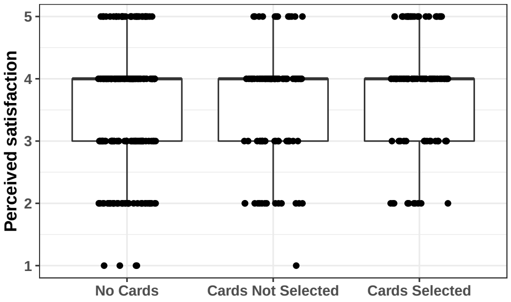


Figure 8: Average satisfaction (1 = very unsatisfied, 5 = very satisfied).

4.5 How does the Benefit Vary across Search Intents? (RQ2)

To answer RQ2, we analysed results based on the two scenario facets: "product" (Factual vs. Intellectual) and "complexity" (Low vs. High complexity). When comparing results across "product" values, we found that health cards were more beneficial to Factual than Intellectual scenarios based on all six measurements. We speculate this may be because in the factual scenarios, the health cards clearly match the scenarios and thus users easily infer the health card's relevance. On the other hand, the health cards for the intellectual scenarios loosely match the scenarios and thus users may not easily infer their relevance, or may be unsure about it (e.g., "acid reflux" for scenario IH_2).

First, the majority of participants selected health cards as a source of information when completing Factual scenarios (53.12%). On the contrary, most Intellectual scenarios were completed based only on information from the search results (see Figure 9 A).

Second, we found that participants submitted *statistically significantly* more correct answers when they selected information from health cards to complete Factual scenarios. Interestingly, although not statistically significant, selecting information from health cards to complete Intellectual scenarios lead to lower correctness than using information only from the search results (see Figure 9 D).

Third, using health cards as a source of information *statistically significantly* reduced the amount of time required to complete Factual scenarios. On the other hand, we found that Intellectual scenarios were completed faster using only information from the traditional search results, though not significantly (see Figure 9 B).

Fourth, health cards benefited participants by *statistically significantly* reducing the amount of effort (i.e., the number of links opened when completing a scenario) needed to complete Factual scenarios. This benefit also occurred for Intellectual scenarios but with less (and not significant) difference (see Figure 9 C).

Fifth and sixth, when selected as a source of information, health cards were perceived as *statistically significantly* reducing the level of workload needed to complete Factual scenarios and significantly improved the level of satisfaction in the participants' own solution. These benefits were also perceived for Intellectual scenarios, but with less and not significant differences (see Figure 9 E & F).

We further analysed these six measures across different scenario complexities. We found that participants are more likely to use health cards as a source of information when completing Low complexity scenarios rather than High complexity scenarios: When health cards were shown, 52.08% of the Low complexity scenarios were completed by selecting information from health cards vs. 45.83% of the High complexity ones. Next, we analysed the effect of selecting health cards as a source of information in completing scenarios of different complexity. Table 6 shows that, regardless of the complexity, selecting health cards as a source of information improved performance on all five remaining measures: increased correctness, reduced duration, reduced effort (i.e., the number of links visited), reduced workload[11], and increased satisfaction. Nevertheless, we found that these improvements were not significant (with the exception of workload and satisfaction for high complexity scenarios).

Table 6: The effect of health cards on Low and High complexity scenarios. The * and ** indicate significant differences, measured by ANOVA with p < 0.05 and p < 0.01, respectively.

	Low Complexity		High Complexity	
	S.Result	H.Card	S.Result	H.Card
Correctness	2.74 ± 0.84	2.88 ± 0.59	1.95 ± 1.14	2.16 ± 1.12
Duration	227 ± 139	222 ± 139	253 ± 128	241 ± 110
Effort	3.56 ± 2.51	2.82 ± 2.40	3.52 ± 1.88	3.02 ± 1.99
Workload	3.72 ± 0.92	3.88 ± 0.90	3.53^{*} ± 0.89	3.84^{*} ± 0.89
Satisfaction	3.65 ± 1.00	3.80 ± 0.95	3.35^{**} ± 1.01	3.80^{**} ± 0.88

4.6 Health card features that help users (RQ3)

To answer RQ3, we investigated health card features that were selected by participants to complete search scenarios. Of the 94 user-scenarios completed using health cards as a source of information, evidences were selected from all three parts of the health cards, with the following proportions[12]: "About" (70%), "Symptoms" (18%), and "Treatment" (50%).

We further analysed which fields of each parts were selected. For the "About" part, the health card contained a list of factual summaries ("treatment", "diagnosable by", "required lab tests", "duration", and "spread") and a more verbose textual summary. We found that 17% of all 384 cases contain evidence selected from the "About" part of the health cards, all contain at least some portion of the textual summary. As for the factual summary, we found that "diagnosable by" was selected in 17% of cases, "treatment" (15%), "diagnostic test" (14%), "duration" (12%), and "spread" (3%). The "Symptoms" part contained a textual summary and a list of symptoms. We found that the textual summary was selected in 71% of cases and the list of symptoms in 65% of cases. Finally, for the "Treatment" part, the textual summary was selected in 61% of cases and the usage rates of the factual summaries were: "medication or treatment" (63%), "specialist" (40%), and "self-treatment" (34%).

5 DISCUSSION AND LIMITATIONS

In this section, we further investigate the impact of presenting health cards on user behaviour and contrasted the effect they have on health search tasks with that general entity cards have for web search (as reported by Bota et al. [6]).

Overall, health search tasks required statistically significantly less effort when health cards were shown, regardless of whether

[11] Workload scores ranged between 1 (very hard) to 5 (very easy).

[12] Note, a participant may have selected from multiple parts.

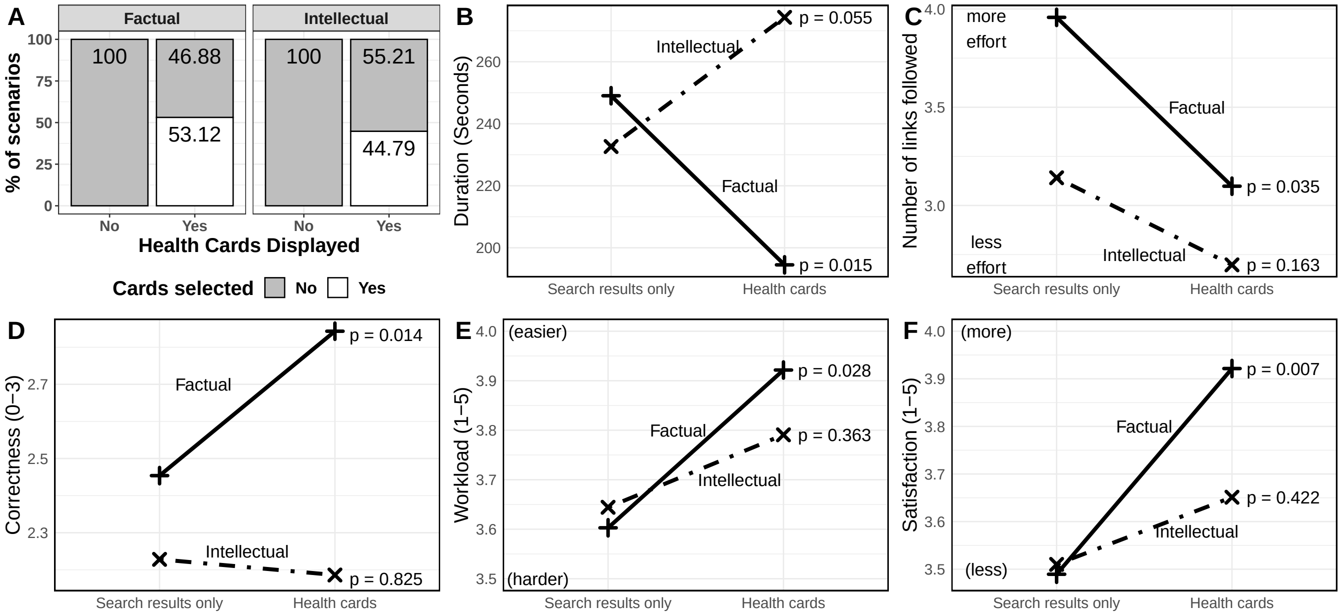


Figure 9: The effects of health cards on Factual and Intellectual scenarios based on six measurements: (A) percentage of scenarios completed using health cards as source of information, (B) correctness, (C) session duration to complete a scenario, (D) effort, (E) workload, (F) satisfaction. The horizontal axis for plots B to F refers to whether information on health cards were selected ("Health cards") or not ("Search results only"). Statistical significant differences are annotated in the plots.

they were used or not. Specifically, less links were clicked when health cards were shown (3.156 vs. 3.615, p=0.043), in line with the results of Bota et al.'s [6] for general entity cards, although they did not report statistically significant differences.

When examining workload, participants perceived workload to complete a health search scenario as statistically significant less when a health card was shown: the average workload to completed a search scenario was 3.776 when health cards were present and 3.589 when health cards were absent (p=0.043; 1: "very hard"; 5: "very easy"). This is in contrast with the results reported by Bota et al. [6], which showed entity cards attracted more workload, although the differences were not statistically significant. We also found that, regardless of whether health cards were used or not, participants felt statistically significantly more satisfied with their submission when health cards were shown (mean satisfaction=3.693) compared to when health cards were absent (3.453, p=0.018).

Interestingly, while the benefits of presenting health cards were apparent, participants seemed to prefer to engage with the organic search results rather than with the health cards. Many of the scenarios (51%), in fact, were completed without selecting health cards as a source of information, and 28% of the participants never selected information from health cards to complete any of the four scenarios where health cards were shown.

These results suggest that health cards led (on average) to higher user benefit in consumer health search than general entity cards in general web search. Nevertheless, such positive effects may be left unreaped. While our results did not undercover why users did not rely more on health cards, we posit that multiple reasons may be responsible for this, including the perceived completeness of the information in the health cards, the trustworthiness of the information and of the match between the card and the scenario. In addition, there is the bias that users who are habituated to seeing search results as a list of links have toward this type of SERP.

In our study, we forbade participants from formulating queries as we focused on measuring the impact of presenting health cards in a controlled manner, without polluting the results with differing query capabilities across users and differing health card to query matching effectiveness. Another limitation of our study is that only relevant health cards were shown: this was done to focus on the effect relevant cards had on user behaviour and decisions, without letting the relevance of a card influence the analysis. Future work will consider end-to-end consumer health search experiments, considering health cards in the context of user-formulated queries, the impact of non relevant health cards and the presentation of multiple candidate health cards for a query.

6 CONCLUSION

In this study, we investigated the impact of health cards on consumer health search. We conducted a laboratory study with 48 participants to complete 8 health scenarios using two search interfaces: one with search result snippets only and one with both result snippets and health cards.

Health cards were used most in Factual scenarios, where they provided significant benefits over using only search results, in terms of more correct answers, faster task resolution, decreased effort and workload, and higher user satisfaction, regardless of the scenario's complexity. However, health cards provided no significant benefits in Intellectual scenarios. These results suggest that health cards are best suited to well-defined health search tasks (i.e., Factual scenarios), rather than "exploratory" tasks.

As for the health card features that most helped users, we found that the condition's summary (the "About" part of the health card) was the most used to select evidence from. The condition's summary contains a textual summary and a factual summary ("treatment", "diagnosable by", etc.) of the condition. In our experiments, all participants that selected evidence from health cards did so principally from the textual summary of the "About" part.

With regard to the effect of health cards on search behaviour, we found that participants generally considered health cards early on in their search session, and then considered the search results afterwards. This may be because participants needed more information to complete their tasks than that provided in the health cards, or that they examined search results to confirm or contrast the information in the health cards.

Finally, we also found that the use of health cards helped the less knowledgeable users to perform effectively as the more knowledgeable users (in term of correctness). Despite this, we found that, of the recruited participants, a considerable portion of those that had searched online for health advice before (93.6% of 48 participants), never noticed health cards in their previous search experiences (40.9% of 44 participants). While the reasons for this behaviour were unclear (e.g., their query may have not triggered the display of a health card, or they may have ignored the card because they did not know it existed, etc.) and are worth exploring in future work, these results highlight that the lack of user engagement with health cards may leave the benefits of health cards unreaped.

ACKNOWLEDGMENTS

Jimmy is sponsored by the Indonesia Endowment Fund for Education (Lembaga Pengelola Dana Pendidikan / LPDP) (20151022014644). Guido Zuccon is the recipient of an Australian Research Council DECRA Research Fellowship (DE180101579) and a Google Faculty Award; both these grants funded this study.

REFERENCES

[1] Laurence Alpay, John Verhoef, Bo Xie, Dov Te'eni, and JHM Zwetsloot-Schonk. 2009. Current challenge in consumer health informatics: Bridging the gap between access to information and information understanding. *Journal of BII* Vol. 2 (2009), BII–S2223.

[2] Jaime Arguello, Fernando Diaz, Jamie Callan, and Ben Carterette. 2011. A methodology for evaluating aggregated search results. In *ECIR'11*. Springer, 141–152.

[3] Krisztian Balog. 2018. *Utilizing Entities for an Enhanced Search Experience. In: Entity-Oriented Search*. Vol. 39. Springer.

[4] Pieter Blignaut. 2009. Fixation identification: The optimum threshold for a dispersion algorithm. *Journal of APP* 71, 4 (2009), 881–895.

[5] Pia Borlund. 2000. Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation* 56, 1 (2000), 71–90.

[6] Horatiu Bota, Ke Zhou, and Joemon M Jose. 2016. Playing your cards right: The effect of entity cards on search behaviour and workload. In *CHIIR'16*. ACM, 131–140.

[7] Melanie Diez, Deborah A Boehm-Davis, Robert W Holt, Mary E Pinney, Jeffrey T Hansberger, and Wolfgang Schoppek. 2001. Tracking pilot interactions with flight management systems through eye movements. In *ISAP'01*, Vol. 6. Citeseer.

[8] Evgeniy Gabrilovich. 2016. Cura Te Ipsum: answering symptom queries with question intent. In *Second WebQA workshop, SIGIR 2016 (invited talk)*.

[9] Faegheh Hasibi, Krisztian Balog, and Svein Erik Bratsberg. 2017. Dynamic Factual Summaries for Entity Cards. In *SIGIR'17*. ACM, 773–782.

[10] Jimmy, Guido Zuccon, and Gianluca Demartini. 2018. On the Volatility of Commercial Search Engines and its Impact on Information Retrieval Research. In *SIGIR'18*. ACM, 1105–1108.

[11] Jimmy, Guido Zuccon, and Bevan Koopman. 2018. Payoffs and pitfalls in using knowledge-bases for consumer health search. *IRJ* (2018), 1–45.

[12] Jimmy, Guido Zuccon, Joao Palotti, Lorraine Goeuriot, and Liadh Kelly. 2018. Overview of the CLEF 2018 Consumer Health Search Task. In *CLEF 2018 Evaluation Labs and Workshop: Online Working Notes (CEUR Workshop Proceedings)*.

[13] Diane Kelly et al. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval* 3, 1–2 (2009), 1–224.

[14] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-ching Wu. 2015. Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework. In *ICTIR'15*. ACM, 101–110.

[15] Dmitry Lagun, Chih-Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. 2014. Towards better measurement of attention and satisfaction in mobile search. In *SIGIR'14*. ACM, 113–122.

[16] Annie YS Lau and Enrico W Coiera. 2007. Do people experience cognitive biases while searching for information? *JAMIA* 14, 5 (2007), 599–608.

[17] Yuelin Li and Nicholas J Belkin. 2008. A faceted approach to conceptualizing tasks in information seeking. *Journal of IPM* 44, 6 (2008), 1822–1837.

[18] Yuelin Li and Nicholas J Belkin. 2010. An exploration of the relationships between work task and interactive information search behavior. *JASIST* 61, 9 (2010), 1771–1789.

[19] Sunghoon Lim, Conrad S Tucker, and Soundar Kumara. 2017. An unsupervised machine learning model for discovering latent infectious diseases using social media data. *JBI* 66 (2017), 82–94.

[20] Carla Teixeira Lopes and Cristina Ribeiro. 2015. Effects of terminology on health queries: An analysis by user's health literacy and topic familiarity. In *Current issues in libraries, information science and related fields*. 145–184.

[21] Edward Alan Miller and Antoinette Pole. 2010. Diagnosis blog: checking up on health blogs in the blogosphere. *AJPH* 100, 8 (2010), 1514–1519.

[22] Sunil Mohan, Nicolas Fiorini, Sun Kim, and Zhiyong Lu. 2018. A Fast Deep Learning Model for Textual Relevance in Biomedical Information Retrieval. In *WWW'18*. 77–86.

[23] Patrick Cheong-Iao Pang, Shanton Chang, Karin Verspoor, and Jon Pearce. 2016. Designing Health Websites Based on Users' Web-Based Information-Seeking Behaviors: A Mixed-Method Observational Study. *JMIR* 18, 6 (2016).

[24] Robert M Plovnick and Qing T Zeng. 2004. Reformulation of consumer health queries with professional terminology: a pilot study. *JMIR* 6, 3 (2004).

[25] Frances A Pogacar, Amira Ghenai, Mark D Smucker, and Charles LA Clarke. 2017. The Positive and Negative Influence of Search Results on People's Decisions about the Efficacy of Medical Treatments. In *SIGIR'17*. ACM, 209–216.

[26] Filip Radlinski, Nick Craswell, Bodo Billerbeck, Milad Shokouhi, Sanaz Ahari, Nitin Agrawal, Timothy Hoad, Song Zhou, and Muhammad Aatif Awan. 2015. Entity detection and extraction for entity cards. US Patent 9,158,846.

[27] David Robins, Jason Holmes, and Mary Stansbury. 2010. Consumer health information on the Web: The relationship of visual design and perceptions of credibility. *JASIST* 61, 1 (2010), 13–29.

[28] Dario D Salvucci and Joseph H Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *ETRA'00*. ACM, 71–78.

[29] Milad Shokouhi and Qi Guo. 2015. From queries to cards: Re-ranking proactive card recommendations based on reactive search history. In *SIGIR'15*. ACM, 695–704.

[30] Luca Soldaini, Andrew Yates, Elad Yom-Tov, Ophir Frieder, and Nazli Goharian. 2016. Enhancing web search in the medical domain via query clarification. *IRJ* 19, 1-2 (2016), 149–173.

[31] Luca Soldaini and Elad Yom-Tov. 2017. Inferring Individual Attributes from Search Engine Queries and Auxiliary Information. In *WWW'17*. 293–301.

[32] Isabelle Stanton, Samuel Ieong, and Nina Mishra. 2014. Circumlocution in diagnostic medical queries. In *SIGIR'14*. ACM, 133–142.

[33] Elaine G Toms and Celeste Latter. 2007. How consumers search for health information. *Health informatics journal* 13, 3 (2007), 223–235.

[34] Ryen White. 2013. Beliefs and biases in web search. In *SIGIR'13*. ACM, 3–12.

[35] Mark Wilson and Karen Draney. 2004. Some Links Between Large-Scale and Classroom Assessments: The Case of the BEAR Assessment System. *Yearbook of the National Society for the Study of Education* 103, 2 (2004), 132–154.

[36] Qing Zeng, S Kogan, N Ash, RA Greenes, AA Boxwala, et al. 2002. Characteristics of consumer terminology for health information retrieval. *Methods of Information in Medicine-Methodik der Information in der Medizin* 41, 4 (2002), 289–298.

[37] Qing T Zeng, Jonathan Crowell, Robert M Plovnick, Eunjung Kim, Long Ngo, and Emily Dibble. 2006. Assisting consumer health information retrieval with query recommendations. *JAMIA* 13, 1 (2006), 80–90.

[38] Bin Zou, Vasileios Lampos, and Ingemar Cox. 2018. Multi-Task Learning Improves Disease Models from Web Search. In *WWW'18*. 87–96.

[39] Guido Zuccon and Bevan Koopman. 2018. SIGIR 2018 Tutorial on Health Search (HS2018): A Full-day from Consumers to Clinicians. In *SIGIR'18*. ACM, 1391–1394.

[40] Guido Zuccon, Bevan Koopman, and Joao Palotti. 2015. Diagnose this if you can. In *ECIR'15*. Springer, 562–567.