Health Cards to Assist Decision Making in Consumer Health Search

Jimmy, MIS.[1,2], Guido Zuccon, PhD.[1], Gianluca Demartini, PhD.[1], Bevan Koopman, PhD.[3]
[1]University of Queensland, Brisbane, Australia; [2]University of Surabaya, Surabaya, Indonesia; [3]Australian E-Health Research Center, CSIRO, Brisbane, Australia.

Abstract

We investigate the effectiveness of health cards to assist decision making in Consumer Health Search (CHS). A health card is a concise presentation of a health concept shown along side search results to specific queries. We specifically focus on the decision making tasks of determining the health condition presented by a person and determining which action should be taken next with respect to the health condition. We explore two avenues for presenting health cards: a traditional single health card interface, and a novel multiple health cards interface. To validate the utility of health cards and their presentation interfaces, we conduct a laboratory user study where users are asked to solve the two decision making tasks for 8 simulated scenarios. Our study makes the following contributions: (1) it proposes the novel multiple health card interface, which allows users to perform differential diagnoses, (2) it quantifies the impact of using health cards for assisting decision making in CHS, and (3) it determines the health card appraisal accuracy in the context of multiple health cards.

Introduction

It is common practise for people to search the Web for health advice and information about conditions, treatments, experiences and health services – we refer to these search activities as Consumer Health Search (CHS). CHS is a challenging domain: effective search is hindered by vocabulary mismatch and the users' lack of domain expertise; these issues affect both query formulation and result appraisal[1]. A study by Zeng et al.[2] showed that, while the general public believes that they were effective in searching for medical advice online, 70% of the study's participants were relying on incorrect advice. Furthermore, Fox & Duggan[3] found that 38% of CHS users did not seek professional attention once they found medical advice online. This is problematic as in many cases incorrect medical diagnosis and treatment could lead to a fatal outcomes.

This study investigates the effectiveness of health cards to assist decision making in CHS (e.g., Figure 1). Health cards are a specific type of entity cards which have recently been introduced by major web search engines to provide quicker access to trusted information around a specific health concept[4]. In general Web search, entity cards are effective in supporting user search activities by presenting heterogeneous information in a coherent way[5]. Yet, no previous work has thoroughly investigated the effectiveness of health cards to assist decision making in CHS.

We specifically focus on the health decision making tasks of (A) determining the health condition presented by a person (self-diagnosis) and (B) determining which action should be taken next with respect to the health condition (e.g., consult a doctor, self-treat, etc.). To support users in making these decisions, we propose a novel interface that shows multiple health cards within a search engine result page (Figure 1 right) – we call this multi-cards – in place of traditional single health card interfaces (Figure 1 left). The multi-cards are inspired by interfaces for product comparison used within shopping websites[6] (e.g., to compare laptops, shoes, etc.) where features of the compared products are summarised and presented side-by-side to assist the user in their purchase decisions. We believe that the multi-cards would allow CHS users to perform *differential diagnoses* by quickly, and with less effort, comparing their health observations to several probable conditions at once. In this context, we aim to address the following questions:

RQ1: How do single and multiple cards influence CHS users when making health decisions? We measure the impact of health cards on CHS decisions based on the following measurements: (1) the *use of health cards* as a source of information; (2) the *correctness* of decisions made on their basis; (3) the *time* needed to make decisions; (4) the *number of web pages opened*; (5) the rate of *good abandonment*[7]; and (6) the *level of confidence* in the decisions.

The multi-cards solution allows to show a set of health cards rather than a single card. This is useful in contexts where the search systems is unsure about which one relevant card should be shown to the user. The ability of showing multiple

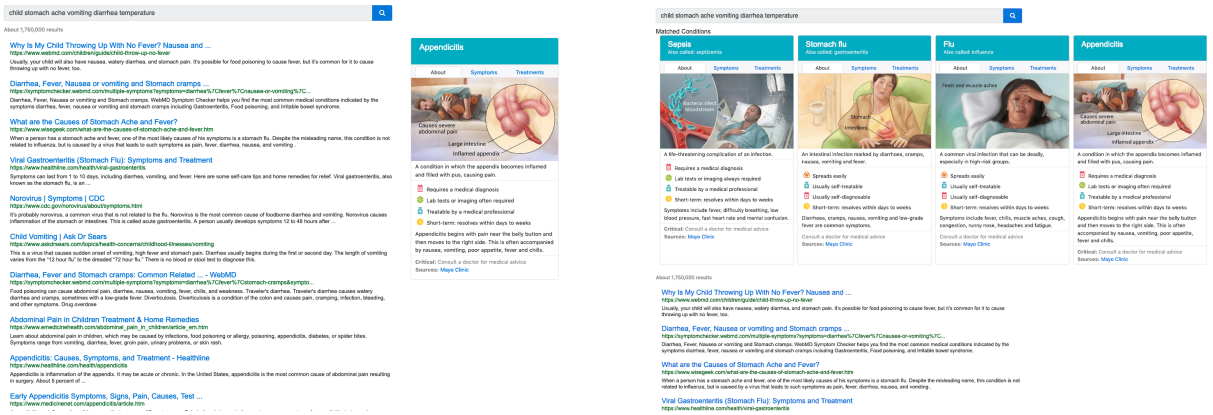


Figure 1: The middle and right panes of our search interface with a single health card (Left) and the health multi-cards (Right) for scenario 1. The left pane of both interfaces (not shown in the figure) contains the health scenario and the tasks to complete, including the request for reporting the confidence in the answers.

health cards increases the chance that the relevant (correct) health card is shown for the user's condition. However, it is unclear whether users would be capable of identifying the correct card for their condition. We investigate this in our second question:

RQ2: How accurate are CHS users in appraising the correctness of health cards? Search results appraisal is challenging in CHS, and is affected by medical terminology, lack of prior knowledge, and cognitive biases[8], among others. We investigate if this holds for the use of health cards by measuring how well CHS users identify the correct health cards to their health situation, in the context of a multi-cards interface.

To answer these questions, we conduct a study where 64 participants are presented with 8 health scenarios and pre-formulated queries, and are asked to consider a search engine result page (SERP) containing health cards and organic search result snippets. Participants are then left to interact with the SERP and are asked to make two decisions: (A) What is the most likely health condition for the scenario? (B) What would you do next? Participants are rotated across two interfaces: one with a single health card, and one with the multi-cards.

Methods

A within-subjects user study was set up to answer the research questions presented above. Figure 2 depicts the activity flow of the user study. Participants were requested to complete 8 health scenarios using four search interfaces: (SC) with a correct health single-card, (SN) with a incorrect health single-card, (MC) with health multi-cards, including a correct card, and (MN) with health multi-cards, where none of the cards are correct. A health card is correct when it matches the target, known diagnosis of the scenario (see Table 2).

The user study was performed in a usability laboratory with a PC equipped with eye tracking technology. To minimise fatigue bias, we rotated the 8 scenarios and the 4 search interfaces using a Graeco-Latin square rotation resulting in 32 scenario–search interface combinations. Participants were principally university students. The study received Human

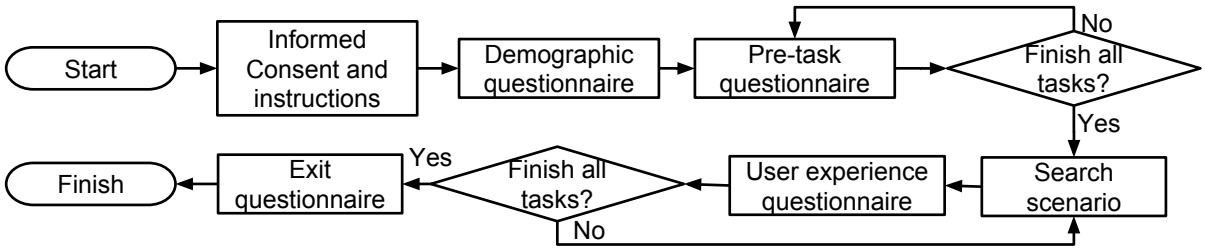


Figure 2: The user study activity flow.

Table 1: Pre-task questionnaire items.

Pre-task Questionnaire Items (options)
1. What is the most probable health condition for the scenario? (open answer)
2. What would you do now? (1=Self-treat, 2=Contact an health professional, 3=Use an emergency service)
3. How confident are you with your answers? (1=Very not confident to 5=Very confident)
4. How interested are you to learn more about the topic of this scenario? (1=Very uninterested to 5=Very interested)
5. How many times have you searched for information about the topic of this scenario? (1=Never, 2=1-2 times, 3=3-4 times, 4=$\geq$ 5 times)
6. How much do you know about the topic of this scenario? (1=Nothing, 2=Little, 3=Some, 4=A great deal)

Research Ethics Committee clearance (ref num 2018002115). The rest of this section details each part of the user study.

Informed 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, highest level of education, education background, English proficiency[1], and the frequency of use of Web search engines. We used the responses to determine the participant's eligibility.

Pre-task questionnaire

After completing the demographic questionnaire, participants completed pre-task questionnaires shown in Table 1 for all 8 scenarios. We used the first three items to understand participants' background knowledge for each health scenario and items 4 to 6 (adapted from Kelly et al.[9]) to understand the participant's interest and background knowledge for each health scenario.

Search scenarios

We selected 8 scenarios of the 45 standardised patient vignettes used in a survey of symptoms checkers[10]. The vignettes were compiled from various clinical sources such as education material for health professionals and a medical resource website. Each vignette contained age, gender, symptoms, correct diagnosis and correct category of triage urgency for a given condition. They include both common and uncommon diagnoses (based on prevalence) from three categories of triage urgency: requiring emergency care, requiring non-emergency care, and self-care appropriate. We ensured that each diagnosis in the 8 selected scenarios had a matching Google health card.

Then, we created a topic description based on each vignette. A topic description contains all symptoms as reported by the patient in the vignette, excluding clinical observations (since in a real setting, the user would not have such information). We also replaced medical terms with layman terms, where appropriate (e.g., "rhinorrhea" was replaced with "runny nose" and "acetaminophen" was replaced with "paracetamol" as "paracetamol" is a more commonly known term in Australia than "acetaminophen"). Finally, we asked research students in our team lab (who have no medical background and had English as first language) to formulate a search query based for each topic description. Table 2 reports the topic description, diagnosis, urgency category, and search query for each health scenario.

To complete each scenario, we asked participants to first make a diagnosis then copy and paste the condition mention — either from the snippets, linked documents, or from the health cards. This protocol allowed us to track where participants found the relevant diagnosis mention and evidence for making their health decision (i.e., they could have found it across different information objects, but they made their final decision based on the copied one). Second, we asked participants to select the urgency condition for the scenario: requires emergency care (e.g., calling 911 or immediately going to hospital), requires non-emergency care (e.g., contacting GP or nurse help line), or self-care appropriate (e.g.,

[1]We verified participants English proficiency by checking whether they: (1) spoke 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.

Table 2: Topic description, correct diagnosis, triage-urgency, and user query string for each health scenario.

Topic 1: Your 12-year-old daughter had a sudden severe abdominal pain with nausea, vomiting, and diarrhea. Her body temperature is 40C.	
Diagnosis (Urgency): Appendicitis (Emergency)	**Query**: child stomach ache vomiting diarrhea temperature
Topic 2: Your 18-year-old brother had sever headache and fever for the last 3 days. He also became very sensitive to lights and experienced neck stiffness.	
Diagnosis (Urgency): Meningitis (Emergency)	**Query**: migraine fever neck stiffness
Topic 3: Your 65-year-old aunt has had pain and swelling in the right leg for 5 days. She has a history of hypertension and recently hoptialised for pneumonia. After returning home from hospital, she had begun walking, but the right leg became painful, tender, red and swollen.	
Diagnosis (Urgency): Deep vein thrombosis (Emergency)	**Query**: hypertension pain swelling leg
Topic 4: Your 18-month-old toddler has had a runny nose, cough and nasal congestion for a week. She also became irritable, sleeping restlessly, and not eating well. She developed a fever overnight. She attends day care and both you and your partner smoke.	
Diagnosis (Urgency): Acute otitis media (Non-emergency)	**Query**: baby fever restless smoking
Topic 5: Your 35 year-old aunt experienced nasal congestion for the last 15 days. She also has had facial pain and green nasal discharge for the last 12 days. She has had no fever. She is otherwise healthy, except for mild obesity. She is on no medications, except for an over-the -counter decongestant. She has no drug allergies.	
Diagnosis (Urgency): Acute sinusitis (Non-emergency)	**Query**: obese nasal congestion over a week
Topic 6: Your 56-year-old aunt who has a history of smoking had shortness of breath and cough for several days. She also had runny nose since 3 days ago. Further, she mentioned to has a productive cough with white sputum. She denies getting chilled or weight-loss and has not received any relief from over-the-counter cough medicine.	
Diagnosis (Urgency): Chronic obstructive pulmonary disease (Non-emergency)	**Query**: white sputum coughing smoker
Topic 7: Your 61 year old mother has had a runny nose and cough productive of yellow sputum for 4 days. She initially had fever as high as 38C but those have now resolved. She is otherwise healthy except for high cholesterol. She has no drug allergies.	
Diagnosis: Acute bronchitis (Self-care)	**Query**: runny nose and fever yellow sputum
Topic 8: Your friend, a 30-year-old man, has had a painful, swollen right eye for the past day. He experienced minor pain on the eyelid but no any history of trauma, no crusting, and no change in vision. He has no history of allergies or any eye conditions and denies the use of any new soaps, lotions, or creams. His right eye had a localised tenderness and redness.	
Diagnosis: Stye (Self-care)	**Query**: swollen red tender eye

taking over the counter drug or home-remedy, resting, performing activities to mitigate the condition). Finally, we asked participants to rate their confidence on the responses (1=Very not confident to 5=Very confident), and the quality of the presented health card(s) with respect to three dimensions: relevancy, understandability and trustworthiness (1=[neg], 2=Partially [pos] 3=[pos]; where [neg], [pos] labels were contextualised to the items; the partial option was only shown when multi-cards were shown).

User experience questionnaire and exit questionnaire

A user experience questionnaire captured the participants' perceived difficulty, perception on system effectiveness, satisfaction and workload. Finally, after completing all 8 health scenarios (including the user experience questionnaire), participants reported 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. Due to space constrains, the analysis on the user experience questionnaire and exit questionnaire are out of the scope of this paper, and are left to future work.

Search Interfaces

The search interfaces contained three panes: left, middle, and right. The left pane displayed the topic, instructions, and tasks to be completed by participants. For the search interface with a health single-card (Figure 1 left), the middle

Table 3: Health cards for each scenario. [C] indicates the correct card.

Id	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
1	Septicemia	Stomach flu	Influenza	food poisoning	Appendicitis [C]
2	Migraine	Meningitis [C]	Encephalitis	Hypertensive disease	Pink eye
3	Hypertensive disease	Eclampsia	Type 2 diabetes	Venous ulcer	Deep Vein Thrombosis [C]
4	Anorexia	Common Cold	Roseola	Sinusitis	Otitis Media [C]
5	Sinusitis [C]	Upper respiratory infection	Chronic obstructive pulmonary disease	Perianal abscess	Common Cold
6	Acute Bronchitis	Pneumonia	bronchiolitis	Cheilitis	Chronic obstructive pulmonary disease [C]
7	Upper respiratory infection	Bronchitis [C]	Chronic obstructive pulmonary disease	Seasonal allergies	Common Cold
8	Blepharitis	Pink eye	Stye [C]	Chapped lips	Chalazion

pane displayed the query string (disabled so a new query could not be entered) and the top ten search results (title, url, and snippet). The right pane of the health single-card interface showed the health card. For the search interface with health multi-cards (Figure 1 right), we merged the middle and the right panes to display the query string, the four health cards, and the top 10 search results. We designed the snippet list and the health cards following the Google search interface as it was the most popular search engine in the country this study took place; thus, participants would be accustomed to the interface.

Search Results

To obtain the search results for each health scenario, we submitted the query to the Bing Web Search API [2] on February 2^{nd}, 2019 and acquired the top 50 search results. To avoid problems with possible web pages and SERP updates, as noted by Jimmy et al.[11], we archived all search results and source web pages. When a participant clicked on any link in the interface (either from the results or from a health card), we presented them with the archived web page. We then identified disease or syndrome concepts in the title and the snippets of each search result using QuickUMLS[12], a tool for extracting Unified Medical Language System (UMLS, version 2018AA) medical thesaurus entities from free-text. We ranked the identified health concepts (i.e., disease or syndrome) based on how many of the top 50 search results contained each concept and we kept the 5 most frequent health concepts. We ensured the selected 5 concepts contained 4 incorrect health concepts and 1 correct health concept (there could not be more than 1 correct). If the top 5 concepts were all incorrect, we exchanged the lowest ranked with the correct concept. This was so that in the multicard interface we could display either 4 incorrect or 3 incorrect and one correct health cards. We only considered health concepts that matched a Google health card. Table 3 lists the five cards of each health scenario. Finally, we selected the 2 search results with the highest rank for each health concept, to make up in total 10 search results for each health scenario, to display in the SERP. The 10 search results were ordered based on ranks from Bing.

Health Cards

Health cards were acquired from the Google search engine based on Table 3. 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 health cards that had no image from the Google health cards, we obtained an image from other medical web pages that discuss the same condition. We fixed the health cards height to 600px and summarised the health cards content to ensure that all information fit the height setting. This was done to provide a similar look & feel for all search interfaces in the study.

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

Capturing Interaction Data

Throughout the user study, we captured participants interactions with the search interfaces using the *Big Brother logging service*[3]. Big Brother records mouse movements (including anchored to `<div>` containers, e.g., enter and leave the container), clicks, scroll, page loading (start and end), cut/copy/ paste, scroll position (mainly to align and validate eye-tracking data).

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[13]. We implemented the *velocity-threshold identification* algorithm[13] to identify fixation points. We set the velocity radius threshold to 70 pixels following the size of the 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[14]. 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[14]. Then, we mapped the fixation points to three *areas-of-Interest* (AOIs): scenario description (left pane), list of snippets, and health cards.

The eye gaze data was used to determine whether participants noticed the health cards, 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 left to future work.

Participants

The study was advertised widely through The University of Queensland, a large public university in Australia, as well as through Facebook groups mainly tailored to students and alumni of this university. We did not enforce participants to be university students or affiliates (but we excluded research staff), 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 web search engine on a daily basis, and proficient English readers and writers. Participants were told that the study would last approximately one hour and were given a AUD$15 gift card for their participation.

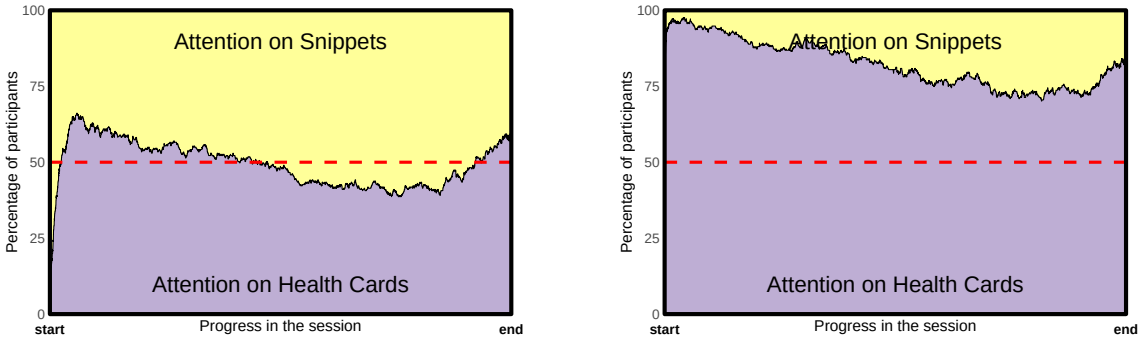
Experiment Results

From 64 participants, each performing 8 scenarios, we collection 512 interaction data points. This gave us enough power for statistical analysis (power > 0.90). Each of the sixteen sequences of scenarios-search interface pairs was performed by 3 participants. Participants comprised 38 females and 26 males: 31 between 18-24 y.o., 23 between 25-34 y.o., 9 between 35-44 y.o. and 1 between 45-54. Participant education background was: 20 Engineering/IT, 13 Business/Economics, 13 Science, 11 Public Health/Exercise and Well-being (not medical), 7 Humanities/Social Science. The highest level of completed education was: 13 high school, 13 diploma, 18 bachelor degree, 5 graduate diploma, and 15 postgraduate degree.

We started by identifying whether fatigue may have had a systematic effect on results. We do so by correlating the sequence of scenarios and the results from the six measurements used in RQ1 (as defined in Introduction). We found that there was a weak negative correlation between the scenario sequence and duration taken to complete a scenario (corr=-0.21): this may due to fatigue or acquired familiarity with the search tasks and the search interface. Further, we found very weak to no correlation between the scenario sequence and the other five measurements: health card usage rate (corr=0.02), correct diagnosis rate (corr=-0.05), correct urgency score (corr=0.06), number of page read (corr=0.04), good abandonment rate (corr=0.00), and confidence rate (corr=-0.13). This indicates that the results are comparable across scenario sequences.

We confirmed that the participants' level of interest, prior knowledge and prior search for the scenario had no systematic effect on results. Overall, participants considered the scenarios as neutral to interesting (Mean $(M) = 3.9$), had nothing to little knowledge of the scenarios ($M = 1.88$), never to 1 - 2 times previous search on the scenarios

[3] `https://github.com/hscells/bigbro`



(a) Single-card

(b) Multi-card

Figure 3: Percentage of participants' that paid attention to snippets vs. health cards overtime.

($M = 1.42$). We found very weak to no correlation between the participants' level of interest, prior knowledge and prior search for the scenario on all six measurements. Next, we investigated the impact of single-card and multi-cards on the participants' search behaviour.

The multi-cards interface was found to drive users' attention toward the health cards (as opposed to the search results). When the multi-cards were shown, the majority of participants spent more time on health cards ($M = 82\%$) than on snippets ($M = 18\%$). However, when only a single-card was shown, participants spent equal amounts of time on snippets and on health cards ($M = 51\%$ vs. $M = 49\%$). These findings were observed by measuring which AOI (i.e., snippets vs. health cards) participants paid attention to when health cards were displayed. (We removed eye tracker data associated to other display areas, e.g., the instructions pane.) Since the time taken by each session varied, we normalised durations, and present results with respect to the progress in the session. Figure 3 reports the percentage of participants that paid attention to each AOI throughout the session. These findings are understandable since when the single-card is shown, there is more information to process in the snippets and the display area containing the snippets is larger. On the other hand, the multi-cards occupy the majority of the initial display and only the top three snippets are visible without scrolling the screen. Participants may have also found most of the information they required among the multi-cards.

Participants often considered health cards earlier in their session: There is a moderate to strong negative correlation between attention on health cards (according to eye tracking) and time point in the session when both single-card (corr=-0.50) and multi-cards (corr=-0.90) were shown. Interestingly, regardless of the declining attention overtime, we found that the attention on health cards increased at the end of the session for both interfaces. We speculate that although prioritised health cards at the beginning of the session, they may have felt the health cards did not contain enough information to make the final decision, and went on examining snippets throughout the SERP. Participants may have felt that information from the snippets was still not sufficient to complete the scenario, hence, they re-considered the health cards at the end of the session.

Results for RQ1: How do single and multiple cards influence CHS users when making health decisions?

Table 4 reports the key metrics of how health cards impact CHS decisions. The participants perceived health cards as relevant (measure 7), trustworthy (measure 8) and easy to understand (measure 9) across search interfaces. We measured health cards' relevance as binary where 0=not relevant and 1=relevant. (Note that a health card that was considered as relevant, was not necessarily considered as correct.) We speculate that participants may have considered a incorrect health card as relevant since it helped participants to rule out a condition. Trustworthiness and understandability were measured on a 1 to 3 scale (3 being most trustworthy/understandable). Next, we will contrast the influence of single-card vs. multi-cards in CHS decision making.

Multi-cards were preferred as a source of information. When multi-cards were presented, in fact, most participants preferred to select information from one of the health cards to complete the scenario (68.75% of 256 scenarios), while

Table 4: The impact of single and multi cards on CHS users in making health decisions. Search interface: (SC) single correct card, (SN) single incorrect card, (MC) multi cards with a correct card, (MN) multi cards with incorrect card. The superscripts indicate statistical significance (unpaired t-test) between the result and the result from the condition associated with the superscript (lower cases refer to $\alpha < 0.05$ and upper cases refer to $\alpha < 0.01$). Higher results are better for measurements 1, 2, 5 and 6. Vice versa for measurements 3 and 4.

Measurements	Num. Cards Shown		Correct Card Present		Search Interface			
	Single^a	Multi^b	Yes^a	No^b	SC^a	SN^b	MC^c	MN^d
1. Health cards usage	0.3711^B	0.6875^A	0.5664	0.4922	0.4062^{CD}	0.3359^{CD}	0.7266^{AB}	0.6484^{AB}
2.A. Correct diagnosis	0.3789^B	0.2461^A	0.4922^B	0.1328^A	0.5781^{BCD}	0.1797^{AC}	0.4062^{ABD}	0.0859^{AC}
2.B. Urgency score	0.543^b	0.4453^a	0.543^b	0.4453^a	0.5977^d	0.4883	0.4883	0.4023^a
3. Duration (sec.)	157	158	150	165	148	166	152	164
4. Num. of page opened	2.4844^B	1.3438^A	1.7812	2.0469	2.2656^{Cd}	2.7031^{CD}	1.2969^{AB}	1.3906^{aB}
5. Good abandonment	0.2344^B	0.4727^A	0.3789	0.3281	0.2578^{CD}	0.2109^{CD}	0.5^{AB}	0.4453^{AB}
6. Confidence	3.6875	3.8359	3.8125	3.7109	3.7031	3.6719	3.9219	3.75
7. Perceived card's relevance	0.6875^B	0.9531^A	0.8516	0.7891	0.7578^{bCD}	0.6172^{aCD}	0.9453^{AB}	0.9609^{AB}
8. Perceived card's trustworthiness	2.6328	2.5195	2.6055	2.5469	2.6562	2.6094	2.5547	2.4844
9. Perceived card's understandability	2.8047^B	2.6016^A	2.6797	2.7266	2.7656	2.8438^{CD}	2.5938^B	2.6094^B

when the traditional single-card was presented, the organic search results were preferred more than the health card (only 37.11% of 256 scenario show information selected from the health card). This was regardless of the correctness of the health cards, and differences are strongly statistically significant.

The single-card was more effective than the multi-cards in leading participants to identify the correct diagnosis (37.5% vs. 24.61%): these differences are strongly statistically significant. Similarly, for the level of urgency correctness, we also found that the average correctness score of the submitted urgency was significantly higher when a single-card was shown (0.543) than when multi-cards were shown (0.4453). To determine the correctness score of the submitted urgency, we computed $score = 1 - (|\Delta urgency| * p)$ where $\Delta urgency = correct\,urgency - submitted\,urgency$ and the possible level of urgency are: 1= self-treat, 2= contact a health professional, and 3= use an emergency service. Lastly, p models the penalty for an incorrect urgency decision. We set the penalty for incorrect decisions so as to greatly penalise urgency decisions that put the well-being of the person at risk (e.g., decided to self-treat when the correct urgency was to use an emergency service). For this study, we set $p = 1$ when $\Delta urgency \geq 0$ and $p = 0.5$ when $\Delta urgency < 0$.

The multi-cards interface, on average, lead users to identify lower levels of urgency (not the correctness score). In fact, a significantly higher level of urgency was recorded when a single-card was shown ($M = 1.9258$) than when multi-cards were shown ($M = 1.7812$). Yet, the mean of the submitted level of urgency from both settings were lower than the mean of the correct level of urgency (2.125). This suggests that users were more likely to make incorrect "what to do" decisions that could have put the person in the scenario at risk.

The interface type did not influence the session duration (no significant difference, $M = 158seconds$); however the multi-cards required less efforts (clicking and browsing web pages) from participants than the single card interface (this difference is statistically significant). In fact, regardless of the presence of a correct card, users opened significantly less result pages when provided with multi-cards. Furthermore, results also show that multi-cards were more likely to lead participants to good abandonment [7], compared to single-card (difference is statistically significant). Remember that good abandonment occurs when a scenario is completed without clicking on links from the SERP. Interestingly, although not statistically significant, we found that the likelihood of users submitting a correct diagnosis was higher when good abandonment occurred (34.81%, 181 scenarios) than when clicking on links (29%). Of the 181 good abandonment occurrences, 83.43% of the decisions were made by selecting information from health cards.

Finally, we found that there were no significance differences in the level of confidence in the decisions made across all search interfaces. Interestingly, inline with Zeng et al.'s findings[2], we found no correlation between the level of

confidence in the user decisions and the user decisions' correctness (both diagnosis correctness and urgency score). The majority of the diagnosis decisions (66%) submitted with (very) confidence were incorrect.

Of the six measurements, only the diagnosis decisions' correctness suggests that the multi-cards are significantly less useful than the single-cards for assisting decision making in CHS. In fact, the highest probability of making a correct diagnosis (57.81%) and the highest average urgency correctness score (0.5977) were found when single correct cards were shown. Nevertheless, determining the single-card that is relevant to a user's query is not trivial for a search system. A search system may identify the single-card to show by relying on the top search result; alternatively it could rely on ranking cards according to the popularity of the underlying concept within the search results. However, if a system based on a single card interface was to follow any of these two methods, then it would display the correct card for only 1 of the 8 queries in Table 2 (thus, probability of single correct card shown: $P(SC) = 0.125$). This is unlike for a multi-cards interface, for which probability of correct card shown: $P(MC) = 0.5$. Overall, the probabilities that a correct diagnosis is made depending on the two interfaces can be compared:

$$
\begin{aligned}
&P(correct|S) \lessgtr P(correct|M) \Rightarrow \\
&P(SC) \cdot P(correct|SC) + P(SN) \cdot P(correct|SN) \lessgtr P(MC) \cdot P(correct|MC) + P(MN) \cdot P(correct|MN) \Rightarrow \\
&0.125 \cdot 0.5781 + 0.875 \cdot 0.1797 \lessgtr 0.5 \cdot 0.4062 + 0.5 \cdot 0.0859 \Rightarrow \\
&0.2295 \leq 0.2461
\end{aligned} \tag{1}
$$

The result above suggests that, if the probability of the system showing a correct health card was accounted for, then the multi-cards interface is more likely to lead to correct diagnoses.

Results for RQ2: How accurate are CHS users in appraising the correctness of health cards?

We tied the correctness of the appraisal of an information object (cards, snippets) to the diagnosis decision that participants made based on the information object, which is identified by the source of information measurement. That is, a correct appraisal of a health card, for example, occurs when a correct decision is made and information from that health card is selected in their answer.

We found that the majority of participants was not able to assess the correctness of a health card within the MC interface. In fact, of all decisions taken using this interface, 39.1% were taken based on non-correct health cards (and which lead to incorrect diagnosis), while 33.6% were taken based on correct health cards. This situation is however better than when users attempted to assess snippets. In this case, in fact, the majority of decisions taken using snippets lead to incorrect diagnoses (20.3%), while only 7.0% were correct. (In every SERP, a minimum of 2 out of 10 snippets contained the correct diagnosis).

In summary, the appraisal of health cards by users is still a challenge, but health cards are more likely to be correctly appraised than search results snippets.

Conclusion

We investigated the influence of health cards to assist decision making in consumer health search. Specifically, we presented 8 health scenarios and asked 64 participants to make two decisions: determining the health condition presented in the scenario and determining which action should be taken next with respect to the health condition. The experiment was conducted in a laboratory using two search interfaces: one with a traditional single health card and one with a novel multiple health cards. Regardless of the search interfaces, health cards were perceived as relevant, trustworthy and easy to be understood.

The novel multi-cards interface is more likely to lead to correct diagnosis than the single-card interface (see equation 1). Indeed, showing a correct single health card leads to the highest probability of submitting a correct diagnosis. However, determining the correct single health card that is relevant to a user's query is not trivial for a search system: an errors are made were an incorrect card is presented instead. If the probability of the system showing a correct health card is accounted for, then the multi-cards interface (which is more likely to show a correct card) leads to a higher number of correct diagnosis decisions than the single card interface.

The multi-cards interface enables users to make health decisions with significantly less effort than the one with the single card. When multi-cards were shown, in fact, participants clicked significantly less links, and completed sig-

nificantly more scenarios without clicking on links from the SERP (i.e., good abandonment), than when they were presented with a single card.

The multiple health cards were the most preferred source of information, compared to the traditional single health card and the organic search results snippets. From the participants' eye-gazing data, in fact, we found that when a single-card was shown, participants spent a comparable amount of time considering the health cards and the snippets. On the other hand, when multi-cards were shown, participants spent significantly more time considering the health cards, compared to the snippets.

Appraisal of search results is still an issue, even within the multi-cards interface. In fact, only 33.6% of decisions taken in this context and using the health cards lead to a correct diagnosis. However, this was better than when using snippets (7.0%)– and in fact the appraisal of health cards appeared to be more accurate than the appraisal of search engine results snippets.

References

1. Zuccon G, Koopman B, and Palotti J. Diagnose this if you can. In *ECIR'15*, pages 562–567, 2015.
2. Zeng QT, Kogan S, Plovnick RM, Crowell J, Lacroix EM, and Greenes RA. Positive attitudes and failed queries: an exploration of the conundrums of consumer health information retrieval. *IJMI*, 73(1):45–55, 2004.
3. Fox S and Duggan M. Health online 2013. Technical report, 2013.
4. Evgeniy Gabrilovich. Cura te ipsum: answering symptom queries with question intent. In *Second WebQA workshop, SIGIR 2016 (invited talk)*, 2016.
5. Bota H, Zhou K, and Jose JM. Playing your cards right: the effect of entity cards on search behaviour and workload. In *CHIIR'2016*, pages 131–140, 2016.
6. Michael Spenke, Christian Beilken, and Thomas Berlage. Focus: the interactive table for product comparison and selection. In *UIST'96*, pages 41–50. ACM, 1996.
7. Jane Li, Scott Huffman, and Akihito Tokuda. Good abandonment in mobile and pc internet search. In *SIGIR'09*, pages 43–50. ACM, 2009.
8. White RW and Horvitz E. Cyberchondria: studies of the escalation of medical concerns in web search. *TOIS*, 27(4):23, 2009.
9. Kelly D, Arguello J, Edwards A, and Wu W. Development and evaluation of search tasks for iir experiments using a cognitive complexity framework. In *ICTIR'15*, pages 101–110, 2015.
10. Semigran HL, Linder JA, Gidengil C, and Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ*, 351, 2015.
11. Jimmy, Zuccon G, and Demartini G. On the volatility of commercial search engines and its impact on information retrieval research. In *SIGIR'18*, pages 1105–1108, 2018.
12. Soldaini L and Goharian N. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, SIGIR*, 2016.
13. Blignaut P. Fixation identification: the optimum threshold for a dispersion algorithm. *Attention, Perception, & Psychophysics*, 71(4):881–895, 2009.
14. Diez M, Boehm-Davis DA, Holt RW, Pinney ME, Hansberger JT, and Schoppek W. Tracking pilot interactions with flight management systems through eye movements. In *ISAP'01*, volume 6, 2001.