Search Engines vs. Symptom Checkers: A Comparison of their Effectiveness for Online Health Advice

Sebastian Cross
sebastian.cross@uq.net.au
University of Queensland
Brisbane, Australia

Ahmed Mourad
a.mourad@uq.edu.au
University of Queensland
Brisbane, Australia

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

Bevan Koopman
bevan.koopman@csiro.au
CSIROBrisbane, Australia
Brisbane, Australia

ABSTRACT

Increasingly, people go online to seek health advice. They commonly use the symptoms they are experiencing to identify the health conditions they may have (self-diagnosis task) as well as to determine an appropriate action to take (triaging task); e.g., should they seek emergent medical attention or attempt to treat themselves at home? This paper investigates the effectiveness of two of the most common methods people use for self-diagnosis and triaging: online symptom checkers and traditional web search engines. To this end, we conducted a user study with 64 real-world users performing 8 simulated self-diagnosis tasks. Participants were exposed to both a representative symptom checker and a search engine. The results of our study provides empirical evidence for whether using a search engine for health information improves people's understanding of their health condition and their ability to act on them, compared to interacting with a symptom checker, which bases its interaction model on a question-answering process. Additionally, recorded answers to qualitative questionnaires from study participants provide insights into which style of interaction and system they prefer to use for obtaining medical information, and how helpful they thought each system was. These findings can help inform the development of better search engines and symptom checkers that support people seeking health advice online.

CCS CONCEPTS

• **Information systems** → *Users and interactive retrieval*; **Retrieval models and ranking**; **Environment-specific retrieval**; *Retrieval tasks and goals*; *Web search engines.*

KEYWORDS

Consumer Health Search, Symptom Checkers, Self-diagnosis, Evaluation, User Studies

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3450140

ACM Reference Format:
Sebastian Cross, Ahmed Mourad, Guido Zuccon, and Bevan Koopman. 2021. Search Engines vs. Symptom Checkers: A Comparison of their Effectiveness for Online Health Advice. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3442381.3450140

1 INTRODUCTION

A national survey in the US reported that one in three American adults have gone online to diagnose their medical condition [24]. This search activity, however, comes with a number of problems. Zeng et al. [29] reported that while most users believe they were effective when searching for medical advice online, 70% of the study's participants relied on incorrect medical advice. In addition, studies by Lau and Coiera [14] first, and White and colleagues [2, 25, 27] later, have also shown how these searches are often affected by a number of biases, including anchoring and presentation bias. Incorrect advice and biases can lead to potentially dire outcomes, creating the need for accurate health information that users can reliably trust, understand and easily obtain.

This study investigates how users obtain health advice online and how they follow this advice. Consumer Health Search (CHS) tasks include: self-diagnosing, triaging, seeking treatment for a known diagnosis, finding first-hand experiences from other sufferers, and others. We investigate two specific tasks: that of self-diagnosing (identifying the medical condition based on observations, signs and symptoms) and triaging (deciding the urgency of the health condition and what to do, e.g., seek emergency medical attention vs. self-treat). On one hand, self-diagnosis is the most popular search activity related to health. According to Google, trending Coronavirus search in July 2020 has been mainly focused on the symptoms of the virus (e.g., *Is a sore throat a sign of coronavirus?, what are the signs of coronavirus?, Is vomiting a symptom of coronavirus?*).[1] On the other hand, self-diagnosis is also a task for which current search technology is often ineffective [30]. This is either because the retrieval system itself fails to identify relevant information, or because the users fail to formulate effective queries or correctly understand the presented search results.

Online health advice is obtained from a number of different avenues. In this study, we investigate two specific technological solutions self-diagnosing users often rely upon: *search engines* and

[1] https://trends.google.com/trends/story/US_cu_D3EmPHEBAABYzM_en

symptom checkers. Search engines have long been used by the general public to access online health advice and information for self-diagnosing [24]; however, in recent years symptom checkers have been developed specifically to assist in this endeavour [7]. Symptom checkers are designed to provide insight into a person's condition and provide advice on triaging and, at times, how to treat the condition. Currently, the investigation into how effective and usable these systems are is limited. For example, different symptom checkers have been evaluated and compared only to other symptom checkers [7]. Similarly search engines have been evaluated for CHS tasks, including for self-diagnosis [9, 20], but never compared with symptom checkers.

This study specifically focuses on comparing search engines against symptom checkers in an attempt to determine users preferences between the two modes of accessing online health advice for self-diagnosing tasks. This study also aims to investigate if and how the information provided by the systems influences users decisions to seek medical attention, and at what level of urgency (self-treat vs. general practitioners vs. emergency). In fact, how the user processes and utilizes the information is an important factor in how symptom checkers operate, as many symptom checkers only provide a recommendation of the action to undertake, rather than a specific diagnosis with associated explanation of why that diagnosis applies [7]. Through this study, we aim to answer the following research questions:

RQ1: Does the use of search engines and symptom checkers alter people's decisions and confidence on self-diagnosis and triaging?
In order to determine if a user is influenced by the considered systems, their self-diagnosis and triaging will be recorded both before and after usage of the systems along with their confidence in the answers, and then compared.

RQ2: Which one is more *effective* for self-diagnosis and triaging — search engines or symptom checkers?
Effectiveness is measured by comparing the accuracy of the decisions made by participants before and after using both systems.

RQ3: Which one requires less *effort* in interactions — search engines or symptom checkers?
Effort is measured as time taken and self-perceived ease-of-use.

RQ4: Which one do people *prefer* more — search engines or symptom checkers?
Participants are asked to rate which system they prefer to use based on satisfaction, ease-of-use and usefulness.

2 RELATED WORK

Search engines are the first port-of-call for many people seeking health advice: a study by Pew Research [24] reported over 70% of online health seekers begin at a commercial web search engine. The most common intents related to online health search include seeking information on a specific health condition or medical problem (55%), treatments for a health condition (43%), and body weight control (27%) [24]. Similarly, White and Horvitz reported that about 42% of web information seekers have searched at least once for self-diagnose [27]. People do report success in health searches online [29]. Access to high quality online health information can help people better understand their health conditions, make informed decisions about health services access and treatment options, and aid them in hypothesis testing and differential diagnosis [13].

Searching online for health information, however, does not come without difficulties and risks. The quality, completeness, trustworthiness, presentation and accessibility (understandability) of online health information varies widely across websites [3, 4, 19]. People searching in this context often formulate ambiguous and underspecified queries [30], which in turn are not effective. People's prior medical knowledge impacts their effectiveness in formulating health queries [8, 16, 22]. In addition, this prior knowledge may not be necessarily correct [13], although they do rely on it, along with previous illness experience, as a guide for self-diagnosis [17]. Finally, health information seekers are often subject to many biases, e.g., anchoring, availability [14, 25, 27]; these biases lead to incorrect interpretation of health search results [13, 25, 26].

An alternative to search engines is *symptom checkers*, which implement dedicated algorithms that, through a series of question-answer interactions with users, attempts to determine the medical condition that may affect the user or provide triaging information, e.g., whether they should seek urgent medical attention.[2] A key advantage of using a symptom checker over a search engine is that people do not need to formulate queries or interpret web pages, tasks for which they often perform poorly in the context of health information seeking [29, 30]. In addition, the number of interactions required by the system, and the amount of time required to arrive at a recommendation, are typically lower than those needed when interacting with a search engine.

Nevertheless, symptom checkers come with their own drawbacks. At times people are uncertain about their answer to one of the symptom checker questions, yet systems typically do not support uncertain answers, or simply saying 'I don't know'. They are also unable to confirm their understanding about a question and the medical terminology used, e.g., they may confuse and misinterpret the expressions "chronic headache" and "severe headache". People may also struggle to interpret the answers provided by the symptom checkers, or may not trust the recommended diagnosis and triaging if no explanation regarding how the system arrived at the particular conclusion.

There is little empirical evaluation of existing symptom checkers with regard to the correctness of the diagnosis and triaging, and the coverage of the conditions. The little evidence available is not promising. A study by Semigran et al. [7] that compared a wide array of symptom checkers reported that systems lack accuracy for both the diagnosis and triaging tasks. Similar results have been reported by other studies in the literature, e.g., [5, 18, 28]. For triaging in particular, Semigran et al. [7] found symptom checkers err on the side of caution and recommend users consult a medical professional when self-treatment is instead reasonable (thus are risk averse). A limitation of that study, however, is that the symptom checkers were evaluated by one of the authors of the study rather than a sample of the intended general public users. Thus it is unclear

[2] Note that different symptom checkers provide different information to their users based on their design goals; however, the key functions they implement often include providing assistance with condition diagnosis as well as triage advice.

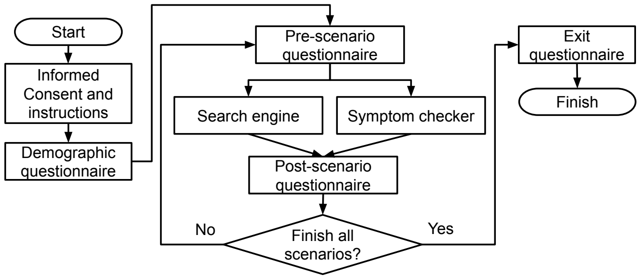

Start
Informed Consent and instructions
Demographic questionnaire
Pre-scenario questionnaire
Search engine
Symptom checker
Post-scenario questionnaire
No
Finish all scenarios?
Yes
Exit questionnaire
Finish

Figure 1: The user study flowchart.

how the general public would have interacted with these systems in the simulated health scenario used for evaluation. For example: would they have been able to understand and answer the questions prompted by the symptom checkers? And would they have been capable of entering the correct answer? What answer would they have provided in uncertain situations? And how would they have interpreted the output of the symptom checker? Would they have trusted it and use it for their health decisions? Our study addresses these limitations by empirically evaluating the effectiveness of symptom checkers in simulated health scenarios with users from the general public.

3 METHODS

A user study was set up to answer our four research questions. A total of 64 participants from the general public were recruited. The study was advertised through local Facebook groups, including those related to local universities, local community groups and local job seeking groups. The study required users to be above the age of 18, have no prior medical study, and to be proficient in English. Participants were told that the study would last approximately one hour, they were given a $15 gift card at the end of their participation.[3]

Figure 1 depicts the flowchart of the user study. Participants completed eight simulated health scenarios representing self diagnoses tasks. The tasks were to be completed using two different systems: 1) a search engine based on the Bing Search API; and 2) a symptom Checker based on the HealthDirect Symptom Checker tool. Each participant completed four tasks using one system and four tasks using the other. To minimise bias with fatigue, we rotated the eight scenarios and the two systems using a Graeco-Latin square rotation [10]. All 64 participants completed all 8 scenarios resulting in 512 scenario data points. The health scenarios, systems interfaces and results of this study are made available online. [4]

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 answering the scenarios. Next, a demographic questionnaire collected information on the participant's age group (grouped by ten-year intervals[5]), highest level of completed education, English proficiency[6], and the frequency of use of general-purpose search engines. We used the responses to determine the participant's eligibility.

Table 1: Demographics of participants.

Age Group	
18-24	37(57.81%)
25-34	19(29.68%)
35-44	3(4.68%)
45-54	1(1.56%)
Highest level of education	
Pre-high school	1(1.5%)
High school	20(31.25%)
Certificate III/IV	4(6.25%)
Advanced diploma & Diploma	2(3.12%)
Bachelor degree	22(3.374%)
Postgraduate degree	13(20.31%)
Graduate diploma & Graduate certificate	2(3.12%)
Prior experience *searching* for health advice online	
Never	5(7.8%)
Once	2(3.12%)
Seldom	14(21.87%)
Frequently	43(67.18%)
Prior experience using *search engines* for self-diagnosis	
Never	8(12.5%)
Seldom	30(46.87%)
Frequently	26(40.62%)
Prior experience using *symptom checkers*	
Yes	17(26.56%)
No	47(73.43%)

Results of the demographic questionnaire are shown in Table 1. Participants were all over the age of 18. Age was distributed as 37(57.81%) participants between 18-24, 19(29.68%)between 25-34, 3(4.68%) between 35-44 and 1(1.56%) between 45-54. All were proficient in English. The highest level of education achieved by participants was: 1(1.5%) pre-high school, 20(31.25%) high school, 4(6.25%) Certificate III/IV, 2(3.12%) Advance diploma & diploma, 22(3.374%) Bachelor Degree, 2(3.12%) Graduate Diploma & Graduate Certificate, 13(20.31%) Postgraduate Degree. In terms of prior experience searching for health advice online, 59(92.18%) of the 64 participants stated having prior experience with 2(3.12%) recalling doing so only once, 14(21.87%) seldom, while the remaining 43(67.18%) reported they search for health advice frequently. When asked specifically if they use search engines for self-diagnosis (as opposed to searching

[3] The study has received Human Research Ethics Committee clearance (*ref num 2018002115*).

[4] http://ielab.io/publications/cross-2021-search

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

[6] 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.

for other health information or advice), 8(12.5%) reported never doing so, 30(46.87%) reported doing this sporadically, and 26(40.62%) stated they often search with a self-diagnosis intent. When asked if they have experience using symptom checkers, 17(26.56%) participants reported having used these systems in the past, while the remaining 47(73.43%) said they never used them.

3.2 Pre-scenario Questionnaire

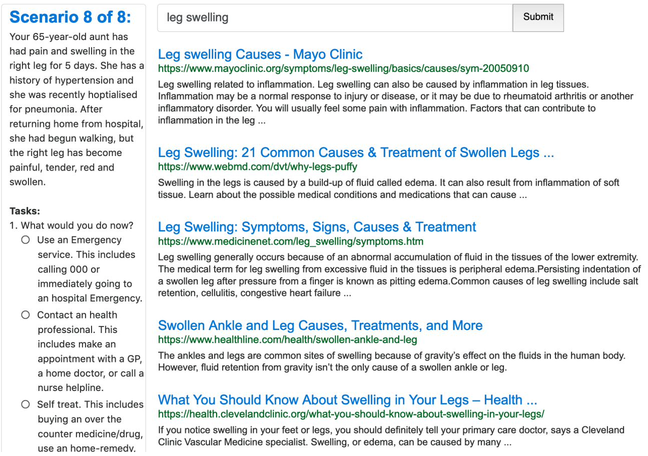
After completion of the demographic questionnaire, participants moved to consider each of the 8 health scenarios assigned to them, one at a time. Before being assigned to a system (symptom checker or search engine), participants had to read the scenarios and answer a series of questions shown in Table 3. These would be repeated at the end of each scenarios to compare the effect each system had on the user's ability to correctly diagnose and act on the medical condition provided in the scenario.

3.3 Self-diagnosis Scenarios

We selected eight scenarios of the 45 standardised patient vignettes (short description of the simulated illness) used in a survey of symptoms checkers [7]. 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. Diagnoses were provided by a panel of clinicians in the original study. Vignettes include both common and uncommon diagnoses (based on prevalence) from four categories of triage urgency: requiring emergency care, requiring non-emergency care, self-care appropriate [9], and not needing medical attention. We ensured that each diagnosis in the eight selected scenarios had a matching Google health card [9].

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 for each topic description. The eight scenarios used are provided in Table 2. Each scenario consisted of narrative describing a fictitious person experiencing a number of symptoms. Each scenarios also has a ground truth diagnosis and triage level. While the hypothetical scenario may not represent the participants' actual information need, this is a common approach (e.g., [11, 15, 21, 23]) which enables control over the experiment conditions and comparison of results across participants [1, 10].

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).



Scenario 8 of 8:

Your 65-year-old aunt has had pain and swelling in the right leg for 5 days. She has a history of hypertension and she was recently hoptialised for pneumonia. After returning home from hospital, she had begun walking, but the right leg has become painful, tender, red and swollen.

Tasks:

1. What would you do now?

○ Use an Emergency service. This includes calling 000 or immediately going to an hospital Emergency.

○ Contact an health professional. This includes make an appointment with a GP, a home doctor, or call a nurse helpline.

○ Self treat. This includes buying an over the counter medicine/drug, use an home-remedy,

leg swelling

Leg swelling Causes - Mayo Clinic
https://www.mayoclinic.org/symptoms/leg-swelling/basics/causes/sym-20050910
Leg swelling related to inflammation. Leg swelling can also be caused by inflammation in leg tissues. Inflammation may be a normal response to injury or disease, or it may be due to rheumatoid arthritis or another inflammatory disorder. You will usually feel some pain with inflammation. Factors that can contribute to inflammation in the leg ...

Leg Swelling: 21 Common Causes & Treatment of Swollen Legs ...
https://www.webmd.com/dvt/why-legs-puffy
Swelling in the legs is caused by a build-up of fluid called edema. It can also result from inflammation of soft tissue. Learn about the possible medical conditions and medications that can cause ...

Leg Swelling: Symptoms, Signs, Causes & Treatment
https://www.medicinenet.com/leg_swelling/symptoms.htm
Leg swelling generally occurs because of an abnormal accumulation of fluid in the tissues of the lower extremity. The medical term for leg swelling from excessive fluid in the tissues is peripheral edema.Persisting indentation of a swollen leg after pressure from a finger is known as pitting edema.Common causes of leg swelling include salt retention, cellulitis, congestive heart failure ...

Swollen Ankle and Leg Causes, Treatments, and More
https://www.healthline.com/health/swollen-ankle-and-leg
The ankles and legs are common sites of swelling because of gravity's effect on the fluids in the human body. However, fluid retention from gravity isn't the only cause of a swollen ankle or leg.

What You Should Know About Swelling in Your Legs – Health ...
https://health.clevelandclinic.org/what-you-should-know-about-swelling-in-your-legs/
If you notice swelling in your feet or legs, you should definitely tell your primary care doctor, says a Cleveland Clinic Vascular Medicine specialist. Swelling, or edema, can be caused by many ...

Figure 2: Screenshot of the search engine interface. Left-hand panel contains the scenario and questions for the participant to answer.

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 general practitioner or nurse help line), or self-care appropriate (e.g., taking over the counter drug or home-remedy, resting, performing activities to mitigate the condition). Finally, we asked participants to rate their confidence of the responses (1=Very not confident to 5=Very confident).

3.4 Search Engine

The search engine based system is developed to closely mimic that of the Bing search engine, as the system is built of the Bing Search API.[7] This user interface screenshot shown in Figure 2 is broken into two columns: query input box and results on the right; form containing the scenario and questions for the participant to answer on the left. This form contains an explanation of the scenario, to rate the triage urgency of the condition, what influence their decision on this specific question and confidence.

3.5 Symptom Checker

The symptom checker system was a clone of that deployed on HealthDirect.gov.au.[8] A screenshot of the user interface is show in Figure 3. The reasons for considering a single symptom checker, Health Direct in specific, are:

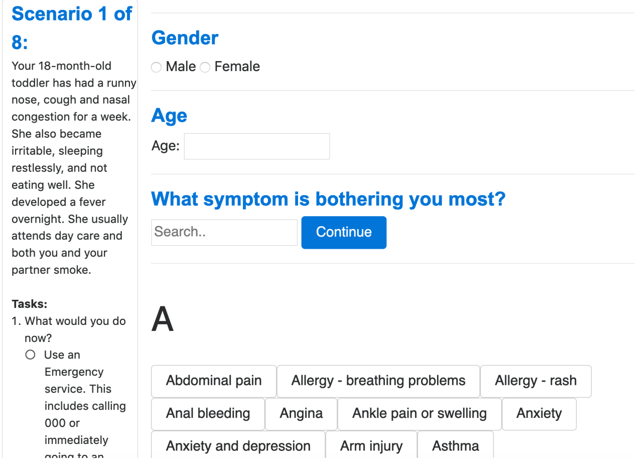
- **Representative.** It uses a question-answering process to determine the outcome. Symptom checkers either ask users to answer a series of questions or enter a list of symptoms [9]. The most popular symptom checkers on the market such as WebMD and Mayo Clinic use this approach.
- **Triage and Diagnosis focused.** Existing popular symptom checkers are either diagnosis focused, such as WebMD and Mayo Clinic, or triage focused, such as NHS [9]. Health Direct provides one or more diagnosis and a triage decision.

[7] https://www.bing.com/

[8] https://www.healthdirect.gov.au/symptom-checker/tool

Table 2: Health scenarios provided to participants. Included are the correct diagnosis and correct triage level.

#	Scenario	Diagnosis	Triage
1	Your 12-year-old daughter had a sudden severe abdominal pain with nausea, vomiting, and diarrhea. Her body temperature is 40C.	Appendicitis	Emergency
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.	Meningitis	Emergency
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 hospitalised for pneumonia. After returning home from hospital, she had begun walking, but the right leg became painful, tender, red and swollen.	Deep vein thrombosis	Emergency
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.	Acute otitis media	Non-emergency
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.	Acute sinusitis	Non-emergency
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 have 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.	Chronic obstructive pulmonary disease	Non-emergency
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 38°C but those have now resolved. She is otherwise healthy except for high cholesterol. She has no drug allergies.	Acute bronchitis	Self-care
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.	Stye	Self-care



Scenario 1 of 8:

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 usually attends day care and both you and your partner smoke.

Tasks:
1. What would you do now?
○ Use an Emergency service. This includes calling 000 or immediately going to an

Gender
○ Male ○ Female

Age
Age:

What symptom is bothering you most?
Search..

Figure 3: Screenshot of the symptom checker interface. Left-hand panel contains the scenario and questions for the participant to answer.

- **Accessibility.** It is a public service, which allows acquiring the questions and answers, unlike commercial symptom checkers, such as WebMD and Mayo Clinic.
- **Public funding.** A government-owned service that is endorsed by Australia's department of health and not driven by private organisations with misaligned objectives [6].
- **Quality of performance.** It's proven to perform with the highest accuracy against other symptom checkers on the market [6]. In addition, Health Direct is contextualized to the local community needs of the Australian health market, where this user study was conducted. For example, some diseases are more common in Australia (e.g., Dengue Fever, covered by Health Direct) than in the US (where the most popular symptom checkers are). Another example is related to the medical care system, which would affect the triage decision.
- **User experience and convenience.** This user study required each participant to complete the health scenarios using two systems (a search engine and a symptom checker) for a fair comparison. There is a trade-off between the statistical power required by the study, number of health scenarios, number of participants, the time allocated to complete the scenario by each participant, and the budget. Hence, we decided to limit it to two systems, 8 scenarios, 64 participants, and approximately 1 hour per study.

The symptom checker works by asking participants a series of multiple choice questions. On completion, the participant is presented with further information about their symptoms and a triage level (e.g., visit your doctor). Importantly, the symptom checkers

Table 3: Questionnaire items — pre, post and exit.

Pre and Post Questionnaire Items (options)

(1) What would you do now? (1=Use an emergency service, 2=Contact a health professional, 3=Self-treat, 4=Re-assess later)
(2) How confident are you with your answers? (1=Very not confident to 5=Very confident)
(3) What Medical Condition do you believe to have? (open answer)
(4) How confident are you with this Medical Condition? (1=Very not confident to 5=Very confident)

Pre-only Questionnaire Items (options)

(1) How interested are you to learn more about the topic of this scenario? (1=Very uninterested to 5=Very interested)
(2) 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=≥ 5 times)
(3) How much do you know about the symptoms observed in this scenario? (1=Nothing, 2=Little, 3=Some, 4=A great deal)

Exit Questionnaire Items (options)

(1) The Symptom Checker was easy to use. (1=Strongly disagree to 5=Strongly agree)
(2) The Search Engine was easy to use. (1=Strongly disagree to 5=Strongly agree)
(3) The Symptom Checker provided me with useful information. (1=Strongly disagree to 5=Strongly agree)
(4) The Search Engine provided me with useful information. (1=Strongly disagree to 5=Strongly agree)
(5) I am satisfied with the Symptom Checker. (1=Strongly disagree to 5=Strongly agree)
(6) I am satisfied with the Search Engine. (1=Strongly disagree to 5=Strongly agree)
(7) Overall System Preference. (I prefer the Search Engine, I prefer the Symptom Checker, I have no Preference, I would not use either)

(in general and in our case) do not give an explicit diagnosis — this is left to the user to infer.

3.6 Post-scenario questionnaire

On the completion of a scenario, the participant was once again presented with questions regarding her self-diagnosis and her triage level. These questions are once again found in Table 3. The participant had to provide a response for each question before moving to the next scenario.

3.7 Exit questionnaire

On the completion of all 8 scenarios the participant was presented with an exit questionnaire. These solicited qualitative feedback on ease-of-use, satisfaction and overall system preference. It also asks if the user has used Symptom checkers before, if so, what kind. Questions are detailed in Table 3.

Table 4: Participants prior knowledge, interest, and search experience.

	Mean	Std Deviation
Interest	3.72	0.85
Search Experience	1.44	0.76
Prior Knowledge	1.91	0.84

3.8 Data Gathering

The data gathering for this study is done using two components, a standard SQL database along with a logging service called Big Brother [9]. The Big Brother logging service captures the interactions of participants as they use both the symptom checker and the search engine. Logged actions include mouse moments, clicks, scrolls, page loading, cut/copy/paste as well as window scroll position. The data captured through Big Brother was to both understand how participants interacted with the systems and the different efforts this involved.

4 RESULTS

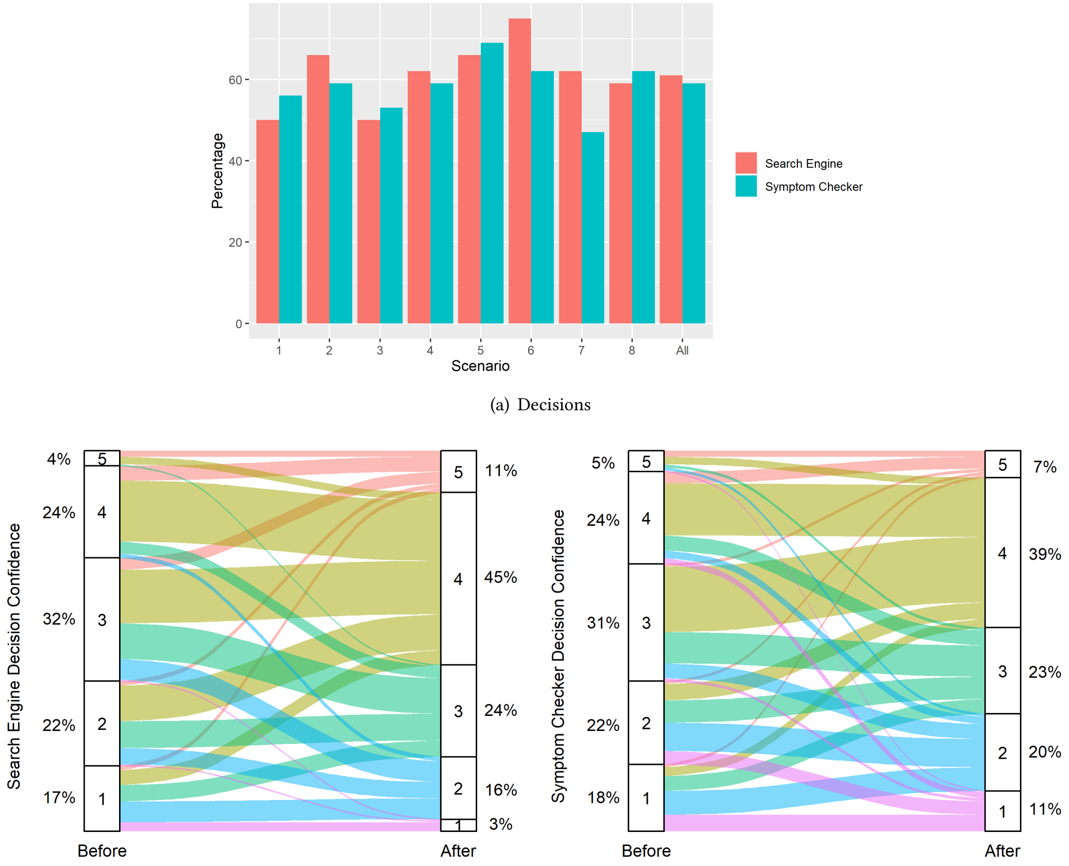
4.1 Prior Knowledge, Interest, Search Experience

We start by analysing results of the pre-only questionnaire (Table 3) to identify whether the participants' level of interest (Q1), search experience (Q2), and prior knowledge on the scenarios (Q3) may have had a systematic effect on results. Table 4 shows that all scenarios were perceived as moderate to highly interesting (Mean (M)=3.72; Standard Deviation (SD)=0.85). As noted in Table 4, 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=1.44; SD=0.76). 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=1.91; SD=0.84). These results indicate that the scenarios were homogeneous in terms of participants interest, prior knowledge, and task definition.

4.2 RQ1 — Change in Decisions

First we examine if search engines or symptom checkers influenced self-diagnosis decisions. Figure 4(a) shows the percentage of changed vs. unchanged decisions after using each system. We observe that using either system causes participants to change their decision roughly 60% of the time. A binomial statistical significance test was performed for three different situations. First, participants will change their initial diagnosis randomly (50%). Second, only participants with incorrect self-diagnosis will change their decisions (96%). Third, participants with incorrect diagnosis will change their decisions randomly (48%). First situation generated a statistically insignificant result with a p-value of 1, while the other two generated statistically significant results with p-values of 4e−142 and 5e−3, respectively. These tests were performed with the results of both the search engine and symptom checker , which indicates that these systems do indeed impact people's decision making. Some

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



(a) Decisions

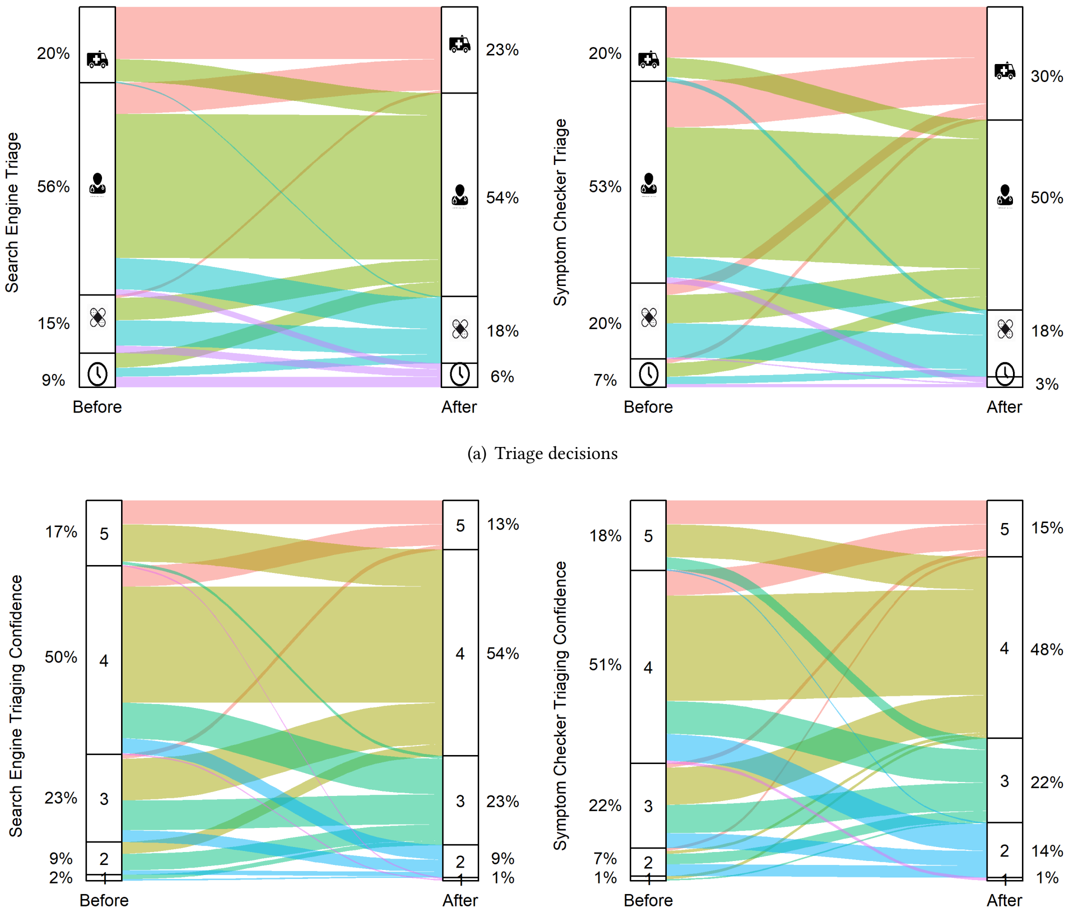
(b) Confidence

Figure 4: Change of self-diagnosis decisions after using searching engine or symptom checker (a) and change in confidence in their decisions (b). SE=search engine, SC=symptom checker. Confidence levels are from 5 (very confident) to 1 (very not confident).

small variation exists between scenarios, but the overall trend remains: most people change their decisions after using either system. Remember however that the symptom checker did not explicitly suggest a diagnosis to participants. Thus changes in diagnosis decisions when using the symptom checker may be ascribed to (1) the questions the symptom checker prompted the user, (2) the triaging recommendation provided by the symptom checker.

Figure 4(b) shows the confidence of participants decisions regarding self-diagnosis for the search engine (left) and the symptom checker (right). Overall confidence increased after using both systems. However, there were some cases of medium confidence before using the system that resulted in low confidence after using the system. The use of search engines also provided people with more confidence than the symptom checker, albeit statistically insignificant with p-value of 0.64 (based on a paired t-test). We posit that being able to query, read documents and interact with the search engine, rather than follow the rigid steps of the symptom checker, may have helped in raising confidence.

Now we examine if search engines or symptom checkers influenced triage decisions. Figure 5 shows the change in triage decisions (a) and change in triage decision confidence (b) before and after use of the two systems. In Figure 5(a), only small differences in triage decisions are observed for scenarios that are initially perceived as not needing medical attention, regardless of the system intervention participants were exposed to. When participants used the symptom checker, however, they often escalated their triaging decisions from requiring non-emergency care to requiring emergency care: a trend that is also observed, but with less strength, for when they used the search engine. A binomial statistical significance test is performed for three different situations. First, participants will change their triage decision randomly (50%). Second, only participants with incorrect triaging will change their decisions (52%). Third, participants with incorrect triaging will change their decisions randomly (26%). No statistical significant differences were found between triaging decisions taken with the symptom checker or with the search engine, with p-values of 1 for all scenarios.



(a) Triage decisions

(b) Triage confidence

Figure 5: Change in triage decisions (a) and change in triage decision confidence (b) before and after use of the two systems. Triage decisions (top to bottom) are: requiring emergency care, requiring non-emergency care, self-care appropriate, and not needing medical attention. Confidence levels are from 5 (very confident) to 1 (very not confident).

In Figure 5(b), we observe a little difference between the two systems in terms of the participant's confidence in their triage decisions, albeit statistically insignificant with p-value of 0.79 (based on a paired t-test). Both systems exhibit a regression to the mean after using the system; that is, highly confident participants lower their confidence after using the system, while highly unconfident participants raise their confidence.

In answer to RQ1, the use of search engines and symptom checkers (1) heavily alters peoples' self-diagnosis decisions while increasing their confidence in such decisions; (2) only slightly alters their triage decisions, with symptom checkers generally tending to escalate them.

4.3 RQ2 — System Effectiveness

We now compare the search engine and the symptom checker for accuracy on self-diagnosis. Figure 6 presents the number of correct answers for each scenario. First, the number of correct answers is generally very low, highlighting the challenge for self-diagnosis online and how people may struggle at this task or make poor decisions based on their online interactions. A two sample proportions z-test was considered to measure the statistical significance of self-diagnosis accuracy between the two systems. A p-value of 2e−3 indicates that the accuracy of the two systems was statistically different, with the search engine being more accurate in 7 out of 8 scenarios, and equal in the remaining scenario.

Figure 7 shows the percentage of correct triage decisions using each system, as well as the percentage of under-estimated and over-estimated triage decisions. Using the search engine, participants made slightly more correct decisions. When users make an incorrect triage decision, they are slightly more likely to underestimate the triage level (e.g., choose self-treat instead of go to the doctor) regardless of the underlying system.

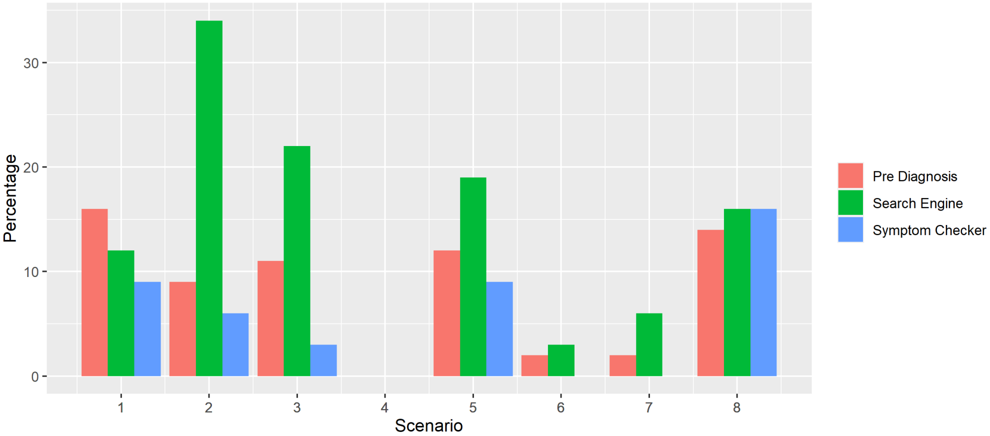


Figure 6: How using a search engine and a symptom checker impacts self-diagnosis decision correctness. SE=search engine, SC=symptom checker.

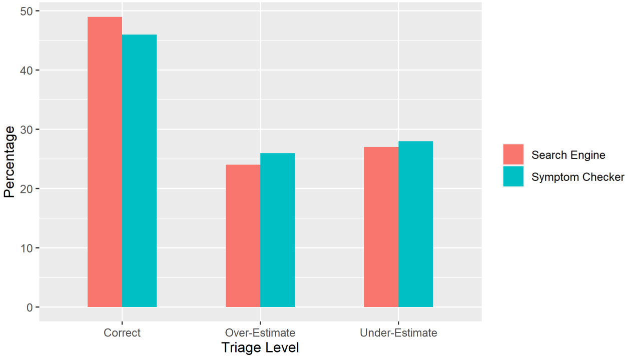


Figure 7: Correctness of triaging. SE=search engine, SC=symptom checker.

Overestimating the triage level is more likely with the symptom checker. A two sample proportions z-test was considered for participants who have chosen either the correct or higher triage urgency level. Although previous research reported that overestimation is a characteristic of symptom checker and can be intentional in their design [7], a p-value of 0.5 indicates that there was no statistically significant difference between the symptom checker and the search engine.

In answer to RQ2, the search engine was statistically more effective than the symptom checker for self-diagnosis. On the other hand, the search engine and the symptom checker were not statistically different for triaging.

4.4 RQ3 — Effort in use

The effort involved when using each system was analysed next. Effort was measured in terms of the time taken to complete a scenario using each system and the self-reported ease-of-use. Considering time, a paired t-test showed that there was no statistically significant difference between the time taken using the search engine versus the symptom checker with a p-value of 0.39.

Ease-of-use was self-reported on a 1–5 scale (5=strongly agree the system was easy to use). The search engine had a mean score of 3.6 while the symptom checker had a mean score of 3.25, indicating that participants found the search engine easier to use as a whole. The variance in ease of use was higher for the symptom checker: SD=1.2 versus SD=1.05 for the search engine. A paired t-test showed that the search engine was statistically easier to use with a p-value of 3e−7. Ease-of-use may be biased by the fact that many people are intimately familiar with using a search engine but may have never used a symptom checker.

In answer to RQ3, the search engine required less effort to use from a self-reported perspective but did not differ from the symptom checker in terms of the time taken to complete a scenario.

4.5 RQ4 — Participant preference

Next we examine the personal preference of participants with regard to each system. When asked which system they preferred, 32/64 (50%) answered the search engine, 22/64 (43%) the system checker, 5 (8%) no preference and 5 (8%) prefer to not use either.

Table 5 provides an overview of the qualitative feedback provided by participants. In terms of ease-of-use, the search engine was easier to use as previously discussed with regard to effort. Considering usefulness, both systems were marked as marginally in the Agree range, with the search engine slightly preferred. Similarly, satisfaction was marginally in the Agree range for both systems, but participants still found the search engine more satisfying. Paired t-tests on both the usefulness and satisfaction for each of the systems indicated the preference of the search engine with statistically significant p-values of 2e−5 and 6e−7, respectively.

To answer RQ4, participants were lukewarm with respect to the usefulness and the their satisfaction of both systems. This may also be representative of the challenge of the self-diagnosis as people struggle to use either system in this complex task. When, however, participants were directly asked which systems they preferred, there was a slight preference for the search engine.

Table 5: Participant qualitative feedback on ease-of-use, usefulness of each systems and satisfaction of each system. Figures show the number of participants

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)	Mean Answer	Std. dev.
	Number of participants						
Ease of use Symptom Checker	7	12	10	28	7	3.25	1.2084
Ease of use Search Engine	3	9	8	35	9	3.5937	1.0498
Useful Symptom Checker	4	11	16	27	6	3.3125	1.0671
Useful Search Engine	4	6	16	26	12	3.5625	1.0965
Satisfaction of Symptom Checker	6	12	16	23	7	3.2031	1.1571
Satisfaction of Search Engine	4	4	21	24	11	3.5312	1.0536

5 LIMITATIONS

Participants used our systems through the web and were not required to attend our usability laboratory. This remote setup allowed us to recruit beyond the convenience sample of a university population (more than half of our participants are not affiliated to the university where this research took place). Thus our participants included people of different cultural background, age, experience (with technology and health matters), education and walk of life, resembling a realistic sample of the local community of an urban area of a developed country. However, this setup reduces our ability to verify whether participants genuinely performed the tasks assigned to them. We note that any malicious behaviour by participants (e.g., randomly entering answers to the scenarios) would have equally affected both the search engine and the symptom checker because of the Graeco-Latin square rotation study design we adopted.

Our study employed two specific systems: a search engine which used the Bing Search API, and a symptom checker that replicated the one provided by HealthDirect.gov.au. Our experimental findings may be highly influenced by the specific systems used, and thus do not generalise to other search engines and symptom checkers. While differences in quality among top commercial web search engines may not be highly demarcated (e.g., a study in the context of health search did not find large difference in effectiveness between Bing and Google [30]), differences between symptom checkers may be more prominent [7]. The symptom checker we used was developed by a national public health service, and thus we believe it to be of high quality. Nevertheless, that symptom checker does not provide a suggestion for the possible diagnosis, but only offers triaging recommendations. In future work we will consider alternative symptom checkers, investigating the effect of system quality and features (e.g., provision of both diagnosis and triaging versus only triaging).

6 DISCUSSION & CONCLUSIONS

Going online for seeking health advice has become common practice. This is achieved through numerous avenues: the major share is taken by search engines, but health-specific services like symptom checkers are increasingly being consulted. We set out to study how effective each of these systems were with respect to self-diagnosis and triaging tasks.

The first finding to note is that both these systems strongly influence people: from an initial diagnosis decision, approximately 60% of people change that decision after using one of these systems (RQ1). Their triage decision was also altered by such systems. Triage decisions, rather than diagnosis decisions, may be a more significant measure in terms of someone's health impact: making an incorrect diagnosis, but consulting a doctor who proceeds to determine the correct diagnosis is a far better outcome than staying at home when medical attention is needed.

On actually choosing the correct diagnosis, the case for the self-diagnoser is not good: only a small percentage of people were able to use either system to find correct diagnoses (RQ2). Where people did get it right, it was more likely achieved using a search engine than a symptom checker. Indeed, there were a number of cases where people's decisions were worse after using a symptom checker. People's confidence in their decisions also changed using these systems — typically increasing. This also has implications: it is far worse to be made more confident in your wrong diagnosis after going online.

Overall, the search engine was more effective in almost every factor evaluated. More correct diagnostic decisions were obtained using the search engine. Given that the motivation for symptom checkers is to constrain users and make them more accurate, addressing many of the perceived shortcomings of search engines, this may be a surprising finding: one would have expected them to be more accurate than search engines. The unconstrained nature of interactions with search engines may well be their very advantage, particularly for explorative hypothesis testing and differential diagnosis [13].

Qualitatively, people preferred the general search engine over the specific health symptom checker. They found the search engine easier to use (even though the time taken to complete the tasks was on par with the symptom checker). Usefulness and satisfaction were also rated higher for the search engine. The exit questionnaire provided to the users was broad and did not ask in-depth questions about the aspects they preferred in the search engine. However, we conjecture that the user preference for the search engine is related to the freedom of exploration along with the explanation it provided about the symptoms given.

While the generalized interpretation of findings across search engines and symptom checkers is hard to make given the limitations of this study, our results provide insight into how symptom checkers could be developed to improve user satisfaction. For example, they could provide broader information about the symptoms, like search engines. These results might also motivate the investigation of alternative solutions to both search engines or symptom checkers, e.g., a system that structures interactions like symptom checkers but provides the freedom of exploration and hypothesis verification of search engines [2].

This study certainly offers a cautionary tale for the use of online systems for self-diagnosis. Poor answer correctness, underestimation of severity (i.e., triage), and over confidence were all observed in this study. The impact of these observations on real people making real health decisions can be severe, even fatal. Given this, it may be understandable why many medical professionals vehemently oppose people performing any self-diagnosis online. Yet, surveys over decades have shown people do not heed these warnings and continue to self-diagnose online. If we cannot stop them, how might we help them? There is certainly a fertile, and very active, area of research on building better search engines and symptom checkers [12]. There is also a lot to be learnt to better understand users — how they go about self-diagnosing online, when and why they succeed or fail, and how best to support them. It is to this last area that this study hopes to help the most.

ACKNOWLEDGMENTS

The authors thank Dr Jimmy (University of Surabaya) for implementing the code to generate the result analysis and figures, and Harry Scells (University of Queensland) for his feedback on this manuscript. The authors also acknowledge the invaluable feedback of the reviewers. Associate Professor Guido Zuccon is the recipient of an Australian Research Council DE- CRA Research Fellowship (DE180101579) and a Google Faculty Award.

REFERENCES

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

[2] Marc-Allen Cartright, Ryen W White, and Eric Horvitz. 2011. Intentions and Attention in Exploratory Health Search. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. ACM, Beijing, China, 65–74. https://doi.org/10.1145/2009916.2009929

[3] Zhongbo Chen and Martin R Turner. 2010. The internet for self-diagnosis and prognostication in ALS. *Amyotrophic Lateral Sclerosis* 11, 6 (2010), 565–567.

[4] Donna M D'Alessandro, Peggy Kingsley, and Jill Johnson-West. 2001. The readability of pediatric patient education materials on the World Wide Web. *Archives of pediatrics & adolescent medicine* 155, 7 (2001), 807–812.

[5] Hamish Fraser, Enrico Coiera, and David Wong. 2018. Safety of patient-facing digital symptom checkers. *The Lancet* 392, 10161 (2018), 2263–2264.

[6] Michella Gaye Hill. 2020. Appraisal of free online symptom checkers and applications for self-diagnosis and triage: An Australian evaluation. (2020).

[7] Semigran HL, Linder JA, Gidengil C, and Mehrotra A. 2015. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ* 351 (2015). https://doi.org/10.1136/bmj.h3480

[8] Yifeng Hu and Jessica Haake. 2010. Search your way to an accurate diagnosis: Predictors of Internet-based diagnosis accuracy. *Atlantic Journal of Communication* 18, 2 (2010), 79–88.

[9] Jimmy, Guido Zuccon, Gianluca Demartini, and Bevan Koopman. 2019. Health Cards to Assist Decision Making in Consumer Health Search. In *AMIA'19*.

[10] 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.

[11] 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.

[12] Liadh Kelly, Hanna Suominen, Lorraine Goeuriot, Mariana Neves, Evangelos Kanoulas, Dan Li, Leif Azzopardi, Rene Spijker, Guido Zuccon, Harrisen Scells, et al. 2019. Overview of the CLEF eHealth evaluation lab 2019. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 322–339.

[13] Alla Keselman, Allen C. Browne, and David R. Kaufman. 2008. Consumer Health Information Seeking as Hypothesis Testing. *Journal of the American Medical Informatics Association* 15, 4 (07 2008), 484–495.

[14] Annie YS Lau and Enrico W Coiera. 2007. Do people experience cognitive biases while searching for information? *Journal of American Medical Informatics Association* 14, 5 (2007), 599–608.

[15] 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.

[16] Carla Teixeira Lopes and Cristina Ribeiro. 2013. Query Behavior: The Impact of Health Literacy, Topic Familiarity and Terminology. In *Proceedings of the International Conference on Human Factors in Computing and Informatics*.

[17] Tana M Luger, Thomas K Houston, and Jerry Suls. 2014. Older adult experience of online diagnosis: results from a scenario-based think-aloud protocol. *Journal of medical Internet research* 16, 1 (2014), e16.

[18] Michael L Millenson, Jessica L Baldwin, Lorri Zipperer, and Hardeep Singh. 2018. Beyond Dr. Google: the evidence on consumer-facing digital tools for diagnosis. *Diagnosis* 5, 3 (2018), 95–105.

[19] Frederick North, William J Ward, Prathibha Varkey, and Sidna M Tulledge-Scheitel. 2012. Should you search the internet for information about your acute symptom? *Telemedicine and e-Health* 18, 3 (2012), 213–218.

[20] Joao Palotti, Guido Zuccon, Jimmy, Pavel Pecina, Mihai Lupu, Lorraine Goeuriot, Liadh Kelly, and Allan Hanbury. 2017. Clef 2017 task overview: The ir task at the ehealth evaluation lab. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*.

[21] 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).

[22] Ira Puspitasari. 2017. The impacts of consumer's health topic familiarity in seeking health information online. In *Proceedings of the 15th IEEE International Conference on Software Engineering Research, Management and Applications (SERA)*.

[23] 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.

[24] Fox S and Duggan M. 2013. *Health online 2013*. Technical Report. Pew Research Center.

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

[26] Ryen W White and Eric Horvitz. 2009. Cyberchondria: studies of the escalation of medical concerns in Web search. *ACM Transactions on Information Systems (TOIS)* 27, 4 (2009), 1–37.

[27] Ryen W White and Eric Horvitz. 2009. Experiences with web search on medical concerns and self diagnosis. In *AMIA'09*, Vol. 2009. American Medical Informatics Association.

[28] Joel A Wolf, Jacqueline F Moreau, Oleg Akilov, Timothy Patton, Joseph C English, Jonhan Ho, and Laura K Ferris. 2013. Diagnostic inaccuracy of smartphone applications for melanoma detection. *JAMA dermatology* 149, 4 (2013), 422–426.

[29] Qing T Zeng, Sandra Kogan, Robert M Plovnick, Jonathan Crowell, Eve-Marie Lacroix, and Robert A Greenes. 2004. Positive attitudes and failed queries: an exploration of the conundrums of consumer health information retrieval. *International Journal of Medical Informatics* 73, 1 (2004), 45–55.

[30] Guido Zuccon, Bevan Koopman, and Joao Palotti. 2015. Diagnose this if you can. In *European Conference on Information Retrieval 2015*. Vienna, Australia, 562–567.