A Think-Aloud Study to Understand Factors Affecting Online Health Search

Amira Ghenai[1], Mark D. Smucker[2], and Charles L. A. Clarke[1]
[1] David R. Cheriton School of Computer Science, University of Waterloo, Canada
[2] Department of Management Sciences, University of Waterloo, Canada
{aghenai,mark.smucker,charles.clarke}@uwaterloo.ca

ABSTRACT

The majority of US Internet users have searched the internet for health-related information. When people conduct these health searches, searching for information about medical treatments is among the more common reasons. While being a convenient and fast method to collect information, when used by people for health search, search engines can be biased toward results saying treatments are helpful, regardless of the truth. The presence of incorrect information in search results may potentially cause harm, especially if people believe what they read without further research or professional medical advice. In this paper, we aim to better understand the decision making process of determining the efficacy of medical treatments using search result pages. We conducted a think-aloud study in order to gain insights on strategies people use during online search for health related topics. We found that, even when participants are careful and focused on the task, biased search engine results can significantly influence people to make decisions consistent with the bias. The chief reason biased search engines results were able to influence participants is that participants often considered what the majority of the search results stated as part of their decision-making. We also found that participants looked for indications of authoritativeness and quality when evaluating online content. While rank bias and a bias towards wanting treatments to be helpful has been found in prior studies, our participants did not reveal these biases as part of their spoken thoughts. Our results imply that more attention should be paid to search engines' biases given people's bias towards accepting the most common answer in the results as the correct answer. When search results are biased toward incorrect results for health-related searches, dire consequences may be the result.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**; **Retrieval effectiveness**.

KEYWORDS

Health Search; User Study; Misinformation; Decision-Making

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHIIR '20, March 14–18, 2020, Vancouver, BC, Canada
© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6892-6/20/03…$15.00
https://doi.org/10.1145/3343413.3377961

ACM Reference Format:
Amira Ghenai[1], Mark D. Smucker[2], and Charles L. A. Clarke[1]. 2020. A Think-Aloud Study to Understand Factors Affecting Online Health Search. In *2020 Conference on Human Information Interaction and Retrieval (CHIIR '20), March 14–18, 2020, Vancouver, BC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3343413.3377961

1 INTRODUCTION

The majority of US Internet users have used web search to look for information about a health issue or a medical treatment [7]. However, there is an increased concern over the lack of accountability and dubious quality of this online content. Prior research [38] has shown that search engines can be biased towards stating that medical treatments are helpful, regardless of the truth.

In our prior work [25], we measured the effect of search results on people's ability to correctly determine the effectiveness of health treatments. We discovered that search results in the health domain can have a substantial, statistically significant, effect on people's decisions. When search results were biased towards incorrect information, the study participants' accuracy in making correct decisions was reduced from 43% to 23%; when biased towards correct information, participants' accuracy increased from 43% to 65%. While we had expected the biased results to influence people, we were surprised by the strength of the effect. A potential concern was that participants were in some way simply echoing back to us the most common answer as a means to please us researchers who had created the search results. Another concern was that participants might have failed to approach the task seriously and were not carefully considering the search results, and thus failed to find the correct information among incorrect results.

To investigate these concerns and better understand the decision making process while people use search engines for health related purposes, we designed an experiment that combined participant think-aloud and one-on-one interviews with the original experiment. Collecting and analyzing think-aloud data has been used to build models of cognitive processes during a problem solving task[35].

For the experiment reported in this paper, we asked participants to determine the effectiveness of four medical treatments. We provided participants with search engine results pages that they had to use to answer the questions about the treatments' efficacy. While doing the task, we asked participants to say out loud what goes through their mind by stating directly what they think. Later, we asked participants about their decisions during the task, as well as generally about their use of search engines for health related purposes. We found that:

- Even with a careful focus on the task, participants are being heavily influenced by a search result bias. When biased towards correct information, participants' accuracy reached 67%, and the accuracy was reduced to 32% when search results were biased towards incorrect information. It seems that even when working diligently at this task, it is easy to be influenced by the search results.
- The majority view in the search results, the apparent authoritativeness of sources, and the apparent quality of sources are among the most important aspects people pay attention to when using search engine results to answer health related questions.
- While prior work shows that rank [1, 11] and optimism bias (believing treatments help) [37, 39] are factors that effect people's online search, participants did not think-aloud these biases. Rank and optimism bias are examples of subconscious biases during the complex process of online health search.

We next discuss related work. We then cover the details of the study and present the study's results, along with our conclusions.

2 RELATED WORK

The work proposed in this paper builds directly on that of Pogacar et al. [25]'s work. The key finding of that study is that search results can have a substantial and statistically significant effect on people's decision about the efficacy of medical treatments. The study showed that when search results are biased towards incorrect information, people's accuracy in decision making was reduced from 43% to 23% while, when biased towards correct information, people's accuracy increased from 43% to 65%. Furthermore, the work showed that the rank of the topmost correct result has an effect on people's accuracy. The study found that prior knowledge of the medical treatment helped protect participants from the presence of incorrect information in search results, and that participants were generally biased towards stating that a medical treatment is helpful, regardless of the truth. Results showed that, even when there was always a correct answer in either rank 1 or 3, participants were not always able to successfully find the correct answer. More importantly, search engines can potentially harm people with a mix of correct and incorrect information. This research showed that search engines may have a substantial impact on people's decisions about the efficacy of medical treatments. In order to overcome current search engine limitations and build ones that better support people's decision making, it is paramount to further explore the reasons leading people to be heavily influenced by misleading information in search results.

In the Pogacar et al. [25] study, limited information about the participants' interaction with search results, e.g., click behavior and search logs, was collected. In the work presented here, we extend the Pogacar et al. [25] study by investigating, in more detail, the interaction between people and search engines. In doing this, we aim to shed light on possible explanations of the impact of search results on people's decision making about the efficacy of medical treatments.

A large amount of prior work has looked at the quality of online health-related content [14, 28, 33] and how people evaluate the credibility of online medical information to make their decisions [4, 8, 13, 34]. Tang et al. [33] compared the search results of Google to those of a domain-specific health and depression search engine. Results showed that, while Google returns more relevant documents, the domain specific search engine returns more correct search result pages. While the quality of online medical content is questionable, prior work looked at the amount of trust people have in the online health information they find online [30–32]. White et al. [36–39] demonstrated that search engines have a strong content bias towards stating that medical treatments are helpful even when they actually are not.

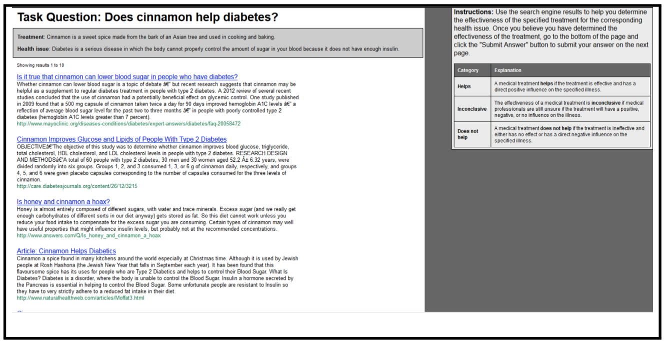
Lau et al.[18] conducted a controlled laboratory study to understand whether providing high quality search results improves people's accuracy when searching online for health related information. Results showed that when participants were provided with high-quality search results from reliable sources (such as PubMed, MedlinePlus, and HealthInsite), the participants' accuracy in answering health questions increased compared to when they were not provided with search results.

Kammerer et al. [15] conducted a controlled laboratory study to understand the behavior and decision making of people when they evaluated web search sources about specific medical issues. The authors selected two different treatments for a certain health issue. The controlled search results had a mixture of different source credibility levels (medical institutions, journals, forums). Later, the authors asked participants to evaluate which treatment was better. Using eye tracking, participants' logs, and verbal protocols, they found that people spend less time and effort evaluating search results when they believe the Web to be a reliable and accurate source of information. Furthermore, people tend to be more certain and require less justifications of information when they trust the Web as a source.

There is a substantial amount of research on using the think aloud method to study the criteria people employ in assessing information sources. Prior studies [21, 26] used verbal protocols to look at the features of search result pages used to gauge search result quality. Lucassen et al.[21], for example, investigated how users evaluate the trustworthiness of Wikipedia content by analysing the think-aloud data. Authors manipulated the quality and the topics to generate varied Wikipedia content with topics familiar to the users. Results showed that participants consider the textual features, references and images when evaluating Wikipedia content.

Further, think-aloud protocols have been used in prior studies [5, 20] in order to shed light on the process of evaluating the credibility of online sources. Elsweiler et al.[5] conducted a think-aloud user study in order to look at how people access the credibility of search result pages. Results showed that people are not certain when accessing the credibility of online sources. People use ten different cues in order to access the credibility of sources and the use of these cues differ for each participant and each topics.

Additionally, verbal protocols in prior studies showed the importance of trust in source selection. Sillence et al. conducted a significant amount of work looking at trust in online health information ([30–32]). They [31, 32] designed a three-stage model of trust when searching for information online. They also [32] conducted a think-aloud user experiment with menopausal patients and a larger-scale experiment [30, 31]. These studies showed that



Task Question: Does cinnamon help diabetes?

Treatment: Cinnamon is a sweet spice made from the bark of an Asian tree and used in cooking and baking.

Health issue: Diabetes is a serious disease in which the body cannot properly control the amount of sugar in your blood because it does not have enough insulin.

Showing results 1 to 10

Is it true that cinnamon can lower blood sugar in people who have diabetes?
Whether cinnamon can lower blood sugar is a topic of debate â€" but recent research suggests that cinnamon may be helpful as a supplement to regular diabetes treatment in people with type 2 diabetes. A 2012 review of several recent studies concluded that the use of cinnamon had a potentially beneficial effect on glycemic control. One study published in 2009 found that a 500 mg capsule of cinnamon taken twice a day for 90 days improved hemoglobin A1C levels â€" a reflection of average blood sugar level for the past two to three months â€" in people with poorly controlled type 2 diabetes (hemoglobin A1C levels greater than 7 percent).
http://www.mayoclinic.org/diseases-conditions/diabetes/expert-answers/diabetes/faq-20058472

Cinnamon Improves Glucose and Lipids of People With Type 2 Diabetes
OBJECTIVEâ€"The objective of this study was to determine whether cinnamon improves blood glucose, triglyceride, total cholesterol, HDL cholesterol, and LDL cholesterol levels in people with type 2 diabetes. RESEARCH DESIGN AND METHODSâ€"A total of 60 people with type 2 diabetes, 30 men and 30 women aged 52.2 Â± 6.32 years, were divided randomly into six groups. Groups 1, 2, and 3 consumed 1, 3, or 6 g of cinnamon daily, respectively, and groups 4, 5, and 6 were given placebo capsules corresponding to the number of capsules consumed for the three levels of cinnamon.
http://care.diabetesjournals.org/content/26/12/3215

Is honey and cinnamon a hoax?
Honey is almost entirely composed of different sugars, with water and trace minerals. Excess sugar (and we really get enough carbohydrates of different sorts in our diet anyway) gets stored as fat. So this diet cannot work unless you reduce your food intake to compensate for the excess sugar you are consuming. Certain types of cinnamon may well have useful properties that might influence insulin levels, but probably not at the recommended concentrations.
http://www.answers.com/Q/Is_honey_and_cinnamon_a_hoax

Article: Cinnamon Helps Diabetics
Cinnamon a spice found in many kitchens around the world especially at Christmas time. Although it is used by Jewish people at Rosh Hashona (the Jewish New Year that falls in September each year). It has been found that this flavoursome spice has its uses for people who are Type 2 Diabetics and helps to control their Blood Sugar. What Is Diabetes? Diabetes is a disorder, where the body is unable to control the Blood Sugar. Insulin a hormone secreted by the Pancreas is essential in helping to control the Blood Sugar. Some unfortunate people are resistant to Insulin so they have to very strictly adhere to a reduced fat intake in their diet.
http://www.naturalhealthweb.com/articles/Moffat3.html

Instructions: Use the search engine results to help you determine the effectiveness of the specified treatment for the corresponding health issue. Once you believe you have determined the effectiveness of the treatment, go to the bottom of the page and click the "Submit Answer" button to submit your answer on the next page.

Category	Explanation
Helps	A medical treatment **helps** if the treatment is effective and has a direct positive influence on the specified illness.
Inconclusive	The effectiveness of a medical treatment is **inconclusive** if medical professionals are still unsure if the treatment will have a positive, negative, or no influence on the illness.
Does not help	A medical treatment **does not help** if the treatment is ineffective and either has no effect or has a direct negative influence on the specified illness.

Figure 1: User interface showing search results. Clicking on a result's title took the user to that result's web page. On the right side, the page shows instructions and the different efficacy definitions [25].

participants rejected sales sites as well as low-quality design content even though they were legitimate sources. Second, when looking at high-quality designed sites, participants trusted content coming from medical institutions or health experts but also personalized content from people similar to the searchers. Third, participantsâĂŹ decision-making processes were influenced by online information: they used online content to reinforce a decision they had already made to find supporting facts and build confidence about their decisions [32]. Other think-aloud studies [22, 23] focused on the importance of people as information sources especially when dealing with the clinical decisions.

3 MATERIALS AND METHODS

3.1 Study Design

At the start of the study, each participant read and signed a consent form and then completed a demographics' questionnaire. We also calibrated our eye tracking device to measure the participant's eye movement. In addition, participants read detailed instructions about their participation before proceeding with the study.

After these preliminary steps, participants had the chance to practice determining the effectiveness of a medical treatment using the search results. During the practice task, participants were asked to articulate and say their thoughts out loud. After this practice task, we started video and audio recording of the participants. Then, participants began the main study where they had to determine the effectiveness of four medical treatments while thinking out loud (*concurrent think-aloud*).

While participants were doing this search task, a researcher took notes about their verbal and non-verbal interactions. After finishing the search task, we showed participants the video recording of their participation, with their eye movements, to help them remember their thoughts, and asked them questions about their decisions (*retrospective think-aloud*). Finally, we asked participants general questions about their usage of search engines for health related purposes (*questionnaire*). The study was designed as a web application and the search results were modelled as a traditional style of web search engine. We recreated the interface of Pogacar et al. [25] for our study. Figure 1 shows the interface that participants used during the study.

Table 1: This table shows the medical treatments with their corresponding efficacy

T	Medical Treatment	Efficacy
T1	Do antioxidants help female subfertility?	Unhelpful
T2	Do benzodiazepines help alcohol withdrawal?	Helpful
T3	Do probiotics help treat eczema?	Unhelpful
T4	Does caffeine help asthma?	Helpful
T5	Does cinnamon help diabetes?	Unhelpful
T6	Does melatonin help treat and prevent jet lag?	Helpful
T7	Does surgery help obesity?	Helpful
T8	Does traction help low back pain?	Unhelpful

3.2 Study Material

Our study material is publicly available[1]. In this section, we briefly explain the study material. Refer to Pogacar et al. [25]'s work for a detailed explanation regarding the study material.

We controlled search result content in two respects. First, *search result bias*, which was either *correct* or *incorrect*. Second, *topmost correct search result*, where we placed the first correct result at either rank 1 or 3. Furthermore, we measured participants' performance by tracking the fraction of correct decisions and the fraction of harmful decisions. Participants had to determine the efficacy of medical treatments as either *helpful* (the medical treatment has a direct positive influence on a specific illness), *unhelpful* (the medical treatment has either a direct negative influence or no influence on a specific illness) or *inconclusive* (medical professionals are not sure about the effectiveness of the medical treatment).

3.2.1 Medical Treatments. We used a list of eight medical treatments from our prior study [25]. Each medical treatment can either be: *helpful* or *unhelpful*. These treatments were originally taken from those developed by White and Hassan [38] who determined the effectiveness of each of these treatments by reading the corresponding Cochrane Review [3, 12] and then reaching a consensus to determine the treatmentâĂŹs efficacy.

Out of the eight medical treatments, four were *helpful* and four were *unhelpful*. Table 1 shows the list of the medical treatments with their corresponding effectiveness. To re-create the conditions of our prior study [25], participantâĂŹs prior beliefs concerning the eight medical treatments were purposely not collected in this study. Specifically, we aim to better understand the results of the prior study and we are not studying how beliefs change. Furthermore, by not asking for prior beliefs, we aim to have the participants be as natural as possible in the experiment.

3.2.2 Search Results. During the study, we asked participants to pretend they had a question about the effectiveness of a medical treatment and had decided to use a search engine to help them answer this question. We showed participants a web page that had ten search results, with the general appearance of a standard search engine results page (SERP). The search results were either biased

[1] https://cs.uwaterloo.ca/~aghenai/user_study_pages.html

Table 2: This table shows the list of post-task questions along with the counts of responses for each question.

No	Question	Yes	No	Maybe
1	Do you believe that exposure (i.e. most results say the treatment helps/does not help) is important in determining the effectiveness of the medical treatment? And why?	13	2	1
2	Do you believe that rank (i.e. highly ranked results say the treatment helps/does not help) is important in determining the effectiveness of the medical treatment? And why?	9	6	1
3	Do you believe that quality is important in determining the effectiveness of the medical treatment? And please elaborate on what quality means to you?	15	0	1
4	Do you believe that the web page layout is important in determining the effectiveness of the medical treatment? And why?	12	2	2
5	Do you believe that social factors (i.e. experience of other people you know such as friends, family etc.) is important in determining the effectiveness of the medical treatment? And why?	9	5	2
6	Did you notice any manipulation of the search results? If yes, then can you guess what was it?	9	7	0
7	How do you describe your experience with the think-aloud process?	-		

towards correct or incorrect information. When biased towards correct information, we showed eight correct search results and two incorrect ones. When biased towards incorrect information, we showed participants eight incorrect search results and two correct ones. We further controlled the rank of the topmost correct result page to either be at rank 1 or 3. We randomly assigned the search results to the corresponding ranks from a pool of 8-10 correct and 8-10 incorrect documents. For every search result, we first showed the title followed by a snippet and a link to the actual page where the participant can click to check the page content. We did not force the participants to check all the ten presented search results because we wanted them to behave as naturally as possible.

3.2.3 Documents and Snippets. To create the SERP pages used in our study, we collected documents about the efficacy of the medical treatment. We used the same 158 documents used in our previous study [25] for this purpose. Every document is either correct (contains information about the treatment efficacy that agrees with medical consensus, i.e., agrees with the Cochrane review) or incorrect (contains information about the treatment efficacy that contradicts with medical consensus, i.e., disagrees with the Cochrane review). For every search result, we showed the document's title, URL, and snippet. We use the same snippets generated for the previous study [25].

3.3 Performance and Statistical Significance

We measured the participants' performance by computing two different measures: 1) the fraction of correct decisions and, 2) the fraction of harmful decisions. A participant's decision is correct if it agrees with medical consensus. For these treatments an inconclusive decision is considered an incorrect decision as all medical treatments in the study were either helpful or unhelpful. Further, a participant's decision is harmful if it is opposite to medical consensus where inconclusive is not considered a harmful decision.

The fractions of correct and harmful decisions are the dependent variables. The search result bias and the topmost correct are the independent variables. In order to measure the statistical significance of the independent variables on the fractions of correct and harmful decisions, we used generalized linear mixed effect model in R. More details about the modeling method can be found in the previous study [25].

3.4 Think-aloud Protocols

We used a think-aloud protocol during the study in order to reveal potential factors influencing the decision making process of people using search engines to answer health-related questions. We combine two types of the think-aloud protocol in the study: concurrent and retrospective think-aloud.

3.4.1 Concurrent Think-aloud. During the think-aloud task, we asked subjects to articulate their thinking and decision making process while doing the search task (*concurrent think-aloud* - CTA). We chose to apply CTA as it is helpful in extracting immediate thoughts while doing the task [17]. This is helpful in order to reveal potential factors influencing the decision making process of people using online search to answer health-related questions.

We captured the think-aloud data through audio recording of the participants with the aid of a computer microphone. In addition, we recorded the screen of the computer as the participant completes the search task using Tobii Pro Studio software[2]. While participants articulated their thoughts, we noted the non-verbal responses during the think-aloud in addition to the words said by the participant (such as pauses, smiles, misreading, periods of silence, pace of speech, body movements, tone variations and volume changes).

One known challenge of the concurrent think-aloud method is that participants might find it difficult to simultaneously articulate their thoughts while doing the search task [16]. In order to address this limitation, we implemented a number of mitigating strategies. First, we restricted our recruitment process to people with English as their first language. With English as their native language, we believe that it is easier for participants to express thoughts while performing study tasks. Second, we designed a practice task where participants complete a short training task before starting the actual study. We presented ten search results and asked the participant to answer a question about the effectiveness of a medical treatment. During this training task, participants get a chance to think out

[2] https://www.tobiipro.com/product-listing/tobii-pro-studio/

loud while determining the effectiveness of a medical treatment. Third, we believe that the study is suitable to apply the think-aloud protocol as the tasks are of intermediate level of difficulty [2]. Finally, during the think-aloud task, there is no interaction between the participant and the searcher in order to not annoy or distract the participant. Instead, a "KEEP TALKING" sign is used to remind participants to talk and encourage the thinking-aloud.

3.4.2 Stimulated Retrospective Think-Aloud. After the concurrent think-aloud, a *stimulated retrospective think-aloud* - RTA was used in which we asked participants about their thoughts after completing the search task [16, 27]. The RTA, where participants were asked questions after finishing the study tasks, is a more natural activity than the concurrent think-aloud process [16]. We implemented the RTA method as it is helpful in the case where participants do not verbalize enough of their ideas. It is also a chance to obtain deeper thoughts and better interpret and validate the CTA (such as asking about pauses and facial expressions, etc.) [17].

In the stimulated retrospective think-aloud study, there was a delay between the study task and the discussion afterwards. In order to help participants remember their thoughts, we used eye tracking during the search task. During the RTA part, while playing back the video recording of the concurrent think-aloud data, we showed participants the captured eye movements to help them recall their thoughts and ideas. We performed eye tracking using a Tobii Pro X3-120 [3] device mounted on the monitor.

Later, participants were asked further **ad-hoc** questions about their decisions and interactions with the search results as the CTA video recording is being played. When formulating the questions, we paid special attention to not introduce any bias, and we avoided leading questions. For this reason, we always made sure to ask questions that start with "What", "When", "Where" and "How". Examples of the questions we ask participants in this part are:

- What was it that made you decide to click on this specific page?
- How did you make up your mind and decided that the treatment is *unhelpful*?
- What did you think of the content in this web page?

3.5 Post-task Questionnaire

After the CTA and RTA (the think-aloud parts), we provided a post-task questionnaire where we orally asked participants general questions about using online search for health purposes. Furthermore, this questionnaire provided a chance to gather feedback about the think-aloud experience. In this part, we asked **open** questions where subjects have the ability to provide responses in the way they prefer (no restricted choices). This type of question is helpful to gain additional varied insights about the decision making process of participants while doing the search talk [16]. The full list of questions asked in the post-task questionnaire are shown in Table 2. While designing the questions, we paid special attention to the wording and made sure that the questions were not biased or double-barreled [16].

3.6 Transcription

We video recorded the concurrent and retrospective think-aloud process while participants interacted with the search task and audio recorded the questionnaire part. An outside vendor transcribed all the parts of the collected data. The reported results in this paper were based on the transcribed data. The transcription service included timestamps for the transcribed scripts, and filler words were removed from the transcripts.

3.7 Coding Scheme

After transcribing the think-aloud recordings, we undertook a coding process in which we generated tags in order to quantify the observations during the think-aloud. We used QSR International's NVivo 12 qualitative data analysis software [19] for the coding process.

We performed a qualitative analysis for the think-aloud data using a mixed methods research approach for both the bottom-up and the top-down approach [10]. Some of the codes were inspired by existing research about possible cognitive biases of using web search for health related purposes such as prior belief [37] and rank [1, 11] (top-down). While other codes had been added and modified as we explore the think-aloud transcribed data such as advertisements, statistics and studies (bottom-up). Applying the mixed methods approach, we aimed to discover the possible strategies participants apply when using search engine to answer a health related question. Further, we performed non-mutually exclusive codes to allow more than one code per item.

The initial coding process was performed by one of the authors (A). Then, the coding was repeated once again by the same author at a later date (B). This process was applied to increase the reliability of the final generated codes. We computed the intra-coder reliability to verify the consistency of the coding between (A) and (B) [9]. To test the intra-rater reliability, we used Cohen's kappa [24]. Cohen's kappa is the ratio of difference between observed agreement and probability of chance agreement over probability of chance disagreement. Cohen's kappa is known to be more robust than a simple agreement percentage as it takes into account the possibility of the agreement occurring by chance [24].

When coding, we kept track of each coding occurrence to compute the frequency counts. Table 6 shows the list of codes with corresponding frequency counts (references). We used this quantitative method in order to identify which of the codes are more and less important for participants during the decision making process.

3.8 Participants

We obtained ethics approval from the Office of Research Ethics at our university. Next, we recruited participants using posters and email announcements to different graduate student email lists. As the user study involved an English language think-aloud process, and in order for participants to be able share their thoughts easier, one of the recruiting requirements was to have only native English speakers. All participants gave their informed consent. Following their participation, we debriefed all participants and provided them with the correct answers regarding the efficacy of the medical treatments. We paid participants $15. Participants were 16 students (7 male, 9 female) from different majors (7 from engineering and

[3] https://www.tobiipro.com/product-listing/tobii-pro-x3-120/

Table 3: Main results. Based on the decisions the 16 participants made, we compute the fraction of correct and harmful decisions. Fractions are shown along with their standard errors.

Results Bias	Fraction of Decisions	
	Correct	Harmful
Correct	0.67 ± 0.08	0.06 ± 0.03
Incorrect	0.32 ± 0.06	0.28 ± 0.06

Table 4: Statistical significance of independent variables.

Independent Variable	Dependent Variable	Pr(>Chisq)
Search Results Bias	Correct Decision	$\ll 0.001$
Search Results Bias	Harmful Decisions	$\ll 0.01$
Topmost Correct Rank	Correct Decision	0.8
Topmost Correct Rank	Harmful Decisions	0.05

Table 5: Confusion matrices. This table shows the decisions made by the study participants regarding the efficacy of the 2 helpful and 2 unhelpful medical treatments.

Truth	Participants			Total
	Unhelpful	Helpful	Inconclusive	
Unhelpful	13	6	13	32
Helpful	5	18	9	32
Total	18	24	22	64

mathematics, 8 from arts and sciences and 1 from environmental studies) with an age between 18 and 28 years old (37.5% less than 20, 56.25% between 20 and 25 and 6.25% greater than 25, with an average age of 21).

4 RESULTS AND DISCUSSION

4.1 Main Results

Table 3 reports the fraction of correct and harmful decisions of the 16 participants corresponding to the search results bias. We see that, similar to the prior study of Pogacar et al. [25], results with bias towards correct information leads to an increased accuracy up to 67% while lowering harmful decisions to 6%. Conversely, results biased towards incorrect information reduces accuracy to 32% while increasing harmful decisions to 28%.

Table 4 reports the statistical significance of the search results bias and topmost correct rank on the correct and harmful decisions. Similar to the prior study of Pogacar et al. [25], we find that the search result bias has a statistically significant effect on the fraction of correct decisions and harmful decisions. Due to the smaller sample, we find that the topmost correct rank has less of an effect on the correct and harmful decisions.

While doing the verbalization process, participants took longer to finish the search tasks (4 minutes average participation time per question in the prior study [25] compared to 10 minutes average participation time per question in the current study). As the think aloud study involved closer observation compared to the prior large computer lab study with very little supervision [25], we expected participants to take the task more seriously and be more conscious about their decisions. As a result, we thought search results bias will have less or no effect on their decisions in this study. However, results demonstrated once again that search results have a potentially strong effect on participants' decisions.

Pogacar et al. [25] as well as White and Hassan [38] demonstrated that their studies' participants have a strong bias towards believing that treatments are helpful. In literature, this human behavior of expecting and believing positive events when there is no evidence to support such expectations is defined as *optimism* bias [29]. Looking at the current think-aloud data, we split the participants' answers about the medical treatment types into "helpful", "unhelpful" and "inconclusive" treatments to further investigate the optimism bias trend. Similar to prior work [25, 38], the results in Table 5 show that helpful is the most frequent option people tend to answer during the study. Furthermore, participants are more likely to answer inconclusive more frequently than what Pogacar et al. [25] observed i.e., when thinking out loud, people tend to respond inconclusive more frequently than when not thinking out loud.

4.2 Think-aloud Method

The coding process shows some insights regarding the potential reasons why people are influenced by the search results even when a correct answer is always placed at either rank 1 or 3.

The average participation time of the concurrent think-aloud part is 39 minutes with a maximum participation of 1 hour and 39 minutes and a minimum of 14 minutes. Table 6 shows the number of participants mentioning each code and the total number of references for that corresponding code. The codes are arranged in a descending ordered by the number of participants, and then by the number of references. In this table, we only report the coding performed during the first time period for two main reasons. First, the Cohen Kappa inter-rater ratio computed was computed between the coding of the two different time periods and the overall value was 0.7 which is a substantial inter-rater ratio [24]. Second, we reached the same main results with both codings. The main coding results are described as follows:

First, from the transcribed data, 14 out of 16 participants mentioned the notion of *Majority* with a total number of 36 mentions. Majority means that participants try to find out what most websites state about the treatment effectiveness or try to look for an agreement between them. If participants are exposed to results geared towards a specific direction, they end up being influenced by what the majority of the search results state. This finding explains why search result bias (both in this study as well as in [25]) has a significant effect on people's decisions. Here, we provide examples of the majority effect from the think-aloud transcript with the participant number in parentheses:

> **(Participant 5)** *I'm going to say helps because a lot of people, like it was just, the vast number were in agreement.*
> **(Participant 6)** *So I'm seeing a lot of doctors recommending the melatonin pill. Yeah, I think this helps.*
> **(Participant 9)** *I think that's the common trend that we're seeing. So I'm going to submit and say that it does help.*

Table 6: The list of codes with their corresponding description. Each code is assigned a label C1-C16 that we use throughout the paper to refer to specific codes. The table also shows the number of participants mentioning a particular code, and the total number of references assigned to the code.

No	Name	Description	Participants	References
C1	*Majority*	The majority of the search results stating that the treatment helps or that the treatment does_not_help or looking for a consensus of different search results.	14	36
C2	*Authoritativeness*	The trustworthiness and reliability in the content of the search results page.	13	153
C3	*Statistics & Studies*	The presence of statistics, numbers and detailed research studies in the search results page.	12	20
C4	*Advertisements*	The presence of messages to promote or sell a product, service or idea in a search results page.	7	16
C5	*Date*	The date and time the search results page was first published to the public or the dates mentioned in the page content reflecting how old the information is.	7	15
C6	*References*	Having a list of sources that have been cited to support the information in the search results page.	7	12
C7	*Negative information*	Mentioning negative information about the treatment in the search results page such as listing the side effects or explaining the dangers of using the treatment etc.	6	15
C8	*Information representation*	The information related to the style of the content presented in the search results page such as list versus grid representation, colors, the page layout, capital letters and special characters etc.	5	18
C9	*Prior belief*	Trusting the information that agrees with our prior knowledge (or belief) and disregarding facts that contradict with it, regardless of the actual truth [39].	5	8
C10	*Readability*	The style of writing and the quality of content being easy to read [6].	4	8
C11	*Relevance*	The relevance to the topic about the effectiveness of the medical treatment.	4	7
C12	*Past experience*	Having a prior experience with the topic (either the medical condition or the treatment) that may effect how much we trust the information in the search results page regardless of the factual correctness.	3	3
C13	*Text length*	The amount of text content in the search results page which might impact the reliability. For example, longer explanations might lead to higher levels of trust.	3	3
C14	*Images*	The presence of visuals in the search results page. The intuition behind this is that images might help better remember the information which may interfere with the decision making process.	2	6
C15	*Rank*	The order of search results in the SERP page that might effect the trustworthiness and reliability of the sources.	2	4
C16	*Social factor*	Relate the information about the topic to people we know. For example, whether a friend or a family member's opinion effects our preferences and decision making.	1	2
		Overall	16	326

It is important to note that some people view search results as individuals having opinions (Participants 5 & 6 in the above examples). They lean towards a specific direction because they believe that the majority of search results reflects the majority of opinions in real life, which is a potentially dangerous misconception. Further, we find that 45% of the total codes are about *authoritativeness* with 13 participants talking about it, with a total of 153 references. Authoritativeness refers to the degree of reliability and trustworthiness with respect to specific content. We observed that participants talk about authoritativeness in three different ways: 40% of the time, people state that the content is not authoritative (negative authoritativeness), 34% of the mentions state that the content is trustworthy (positive authoritativeness) and the remaining 26% are about not being sure whether or not to trust the content (neutral authoritativeness). Below, we show some examples of each case from the think-aloud transcript:

> **(Participant 17)** *Health.com, I've seen it before, not really ... I don't really rely on it for information the first time I see it.*
>
> **(Participant 10)** *WebMD. It's a more trustworthy source, I think.*
>
> **(Participant 14)** *Okay. I don't really know what this website is. Medications for management of alcohol withdrawal.*

The high percentage of mentions regarding authoritativeness shows the importance of this factor to participants when evaluating the effectiveness of the treatments. When Pogacar et al. [25] designed the user study, they did not control for the authoritativeness of search results, as authoritativeness was not one of the independent variables. As a result, correct answers are not guaranteed to be in authoritative search results (a correct answer might appear in a non-trustworthy source such as personal blogs or forums). This might potentially negatively affect participants' performance especially with an incorrect search results bias. Not controlling for authoritativeness might be another possible reason why people have been heavily influenced during the study.

Participants talk about many factors that define the quality of search results during the think-aloud. Concepts C3-6, C8, C10 and C13-14 in Table 6 are all about quality. For example, 12 participants mention 20 times the statistical analysis and detailed research studies during the think-aloud process (C3) in order to evaluate the quality of information in the search results. Examples of such beliefs can be found in the transcriptions:

> **(Participant 12)** *...so this is explaining a study. Who had been given cinnamon reduced their blood sugar by 18 to 29 percent. Well that seems like some good numbers. So that's interesting. I think, based on that, I'd probably say that it helps because it had really evidence from a study.*
>
> **(Participant 15)** *So this looks like a research study, so I think it's pretty reliable.*

We, further note some notion about prior beliefs during the think-aloud where prior belief (C9 in Table 6) refers to the idea of believing information that agrees with our prior beliefs or knowledge and ignoring content that contradicts with it. A total of five participants mentioned the prior belief concept during the think-aloud process 8 times. Below, we show some examples:

> **(Participant 16)** *And I was also taught from school that benezenes are harmful to health so though I might be bias I have this thought that benzene would not exactly help with certain health concerns.*
>
> **(Participant 3)** *So this Kurt Donsbach, PhD ... He will claim that it has no positive function at all, but I've heard different, so right away I'm not convinced by this page.*

We also coded the concept of rank where Table 6 shows that only 2 participants out of 16 mentioned rank a total of 6 times. We show an example from the think-aloud transcript below:

> **(Participant 19)** *I'll just go to the first link, even though it's wikiHow, it is the first link. I don't really know much about search engines, but I feel like the first link ... they're trying to give you the most helpful link. So I'll just open it, but still.*

Looking at the results, people rarely talk about rank when, in prior research [1, 11, 25], it has been shown that rank has a potential effect on people's decisions. A possible explanation is that people are unconsciously influenced with the higher ranked search results, however, they are not enough aware of this effect to vocalize it during a think-aloud.

We know from White et al. [37, 39] that people have a strong *confirmation bias* when using search engines. Further, we noticed from results in Table 5 and from Pogacar et al. [25]'s work that people have a bias towards believing that treatments are helpful (optimism bias). However, in the think-aloud transcription data and coding process, we fail to find any mention of such biases (confirmation or optimism bias). Perhaps participants are not aware of these influences, but are still being biased with such factors during the search. Further, confirmation as well as optimism biases are other examples of unconscious biases that the think-aloud study fails to reveal.

4.3 Retrospective Think-aloud

We audio recorded, transcribed and summarized the data gathered during the retrospective think-aloud. The average participation time of the retrospective think-aloud part was 25 minutes with a maximum participation of 37 minutes and a minimum of 17 minutes. Looking at the retrospective think-aloud transcription helps in giving insights of new strategies participants used during the study that might not be captured during the concurrent think-aloud part. Here is a list of strategies caught from the retrospective think-aloud summaries (between parentheses we specify the participant mentioning the strategy):

- Reading pages that state the medical treatment *does not help* in order to understand the opposite arguments. (Participants 3, 7 and 18)
- Finding reliable sources first, then quickly checking relevant less reliable websites (such as answers.com and Yahoo answers) in order to look for consistency. (Participant 4)
- Reading the search result page first to understand the causes of the health issue, before reading about the effectiveness of the medical treatment. (Participant 7)
- Using the first clicked on link as a base reference for all future websites the participant decides to look at. (Participant 8)
- In case no consistency exists between search results, the participant tries initiating a new search query with different keywords. (Participant 9)
- In case of no consistency between search results, the participant looks at the dates the information was published in order to check whether the non-agreement happens because of time difference. (Participants 9 and 17)
- When the same hostname/website appears more than once in the SERP page, the participant believes that this is a reliable source. (Participant 9)

- Deciding which websites to click on by looking at URL titles to check whether they contain the exact words as the search keywords (Participant 13).
- The participant trusts a non-credible website when there are other websites that state the same information as the non-credible website. (Participant 16)
- The participant only opens websites based on prior experience i.e. the participant opens websites that have been reliable and helpful in the past and does not trust or does not open websites that are not familiar. (Participant 17)

4.4 Post-task Questionnaire

Table 2 shows the list of questions as well as the participants answers. From the table, we can notice that 13 out of 16 participants believe that exposure is important when evaluating the search result pages, where we define exposure as what the majority of the search results state. This observation aligns with what we observed in the think-aloud transcriptions as majority was mentioned by 14 participants in total. Participants are consciously aware of the influence of majority while evaluating the treatment's effectiveness because they possibly believe that majority reflects real life opinions. When asked about majority, participants explained that this means what most of the content they were exposed to stated. It is important to note that the majority opinion is constructed differently for every participant. While some participants look at the whole SERP page without clicking on all ten search results (they only click when they want to know more about the content) to understand the majority, others define the majority opinion by only looking at the items clicked on (for example, they only check the highly ranked results and ignore the lower ranked ones).

Further, when explicitly asked about the rank, only 57% of the participants believe that rank is important in evaluating the search result pages. Similarly, looking back at the think-aloud data, we observed that only two participants mention rank during the think-aloud task. Again, this shows that rank is a subconscious factor that effect people's decision making while doing online search. Next, 15 out of 16 participants strongly believe that quality is important in doing online search. Eleven participants explained quality as a notion of authoritativeness, while two participants believe that quality has to do with readability and three participants stated that layout is the major factor to determine the quality of websites. When specifically asked about the layout, 12 out of 16 participants believe that the page design is important in evaluating the search results page.

When asked about the social factor (i.e. the experience of other people we know such as friends and family etc.), only 9 out of 16 participants believe that it is important in evaluating search results. Social factor is a type of subconscious bias where people tend to believe that family and friends do not effect the decisions when they, subconsciously, do.

We, further, asked participants whether they noticed any manipulation of the search results during the study and, if they did, we asked whether they could guess the factor of manipulation. Seven out of sixteen participants could feel that there is a manipulation while seven participants could not notice any manipulation. Four participants guessed that rank was the manipulation factor while another four suggested authoritativeness as the manipulation factor. Two participants thought that the URL was changed during the study design. One participant suggested that we introduced duplicate results and one participant felt the manipulation was about correctness (which was the only right guess among all participants). The participants' responses suggest that we successfully designed the deceptive aspects of the study (rank and correctness) so that participants behaved normally (without any behavioral influences that would make the observations invalid).

Finally, when asked about the experience of participants with the think-aloud process, five participants found the study interesting and insightful, five participants found the study was a good experience, two participants found the think-aloud part to be hard because of the thinking while talking process, and one participant found herself being more conscious about the decisions while stating them out loud.

5 CONCLUSION

When people perform online search regarding the effectiveness of medical treatments, search engine result pages often contain incorrect and misleading results. Due to known content biases, people can be influenced to believe that ineffective treatments are actually effective, potentially causing harm. In order to create search engines that provide better support for decision making about medical treatments, we need to gain insights into the strategies people use during this decision making process. Understanding cognitive biases associated with the use of search engines to answer health related questions is a complex problem, partly because there is a potentially large number of such biases and other unconscious factors effecting the decision making process. In this paper, we conducted a think-aloud study where we asked participants to verbalize their thoughts while using search results to decide about the effectiveness of a medical treatment. Results revealed some strategies people use during online searches for health related topics.

We thought that the think-aloud process might lessen the effect of the search result bias, since participants carefully performed the task in front of a researcher. However, participants were still significantly influenced by the misinformation, demonstrating the degree to which search biases can impact the decision making process.

Additionally, biased content led participants to believe that search results reflects real life opinions. In particular, when the majority of search results reflect a certain view, this can be interpreted as a majority consensus in the world at large. The implications are profound when, for example, searching for cancer treatments on today's popular web search engines might return a mix of correct and incorrect results.

Finally, when people use search engines to answer questions, there are many factors that may unconsciously effect their decision making, which the think-aloud method used here failed to catch.

6 ACKNOWLEDGMENTS

Special thanks to Joan Bartlett for her advice and help. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (Grants CRDPJ 468812-14 and RGPIN-2014-03642).

REFERENCES

[1] Ahmed Allam, Peter Johannes Schulz, and Kent Nakamoto. 2014. The impact of search engine selection and sorting criteria on vaccination beliefs and attitudes: two experiments manipulating Google output. *Journal of medical Internet research* 16, 4 (2014), e100.

[2] Elizabeth Charters. 2003. The use of think-aloud methods in qualitative research an introduction to think-aloud methods. *Brock Education Journal OLD* 12, 2 (2003).

[3] A Cipriani, TA Furukawa, and C Barbui. 2011. What is a Cochrane review? *Epidemiology and psychiatric sciences* 20, 3 (2011), 231–233.

[4] Sameer Dhoju, Md Main Uddin Rony, Muhammad Ashad Kabir, and Naeemul Hassan. 2019. Differences in Health News from Reliable and Unreliable Media. In *Companion Proceedings of The 2019 World Wide Web Conference*. ACM, 981–987.

[5] David Elsweiler and Markus Kattenbeck. 2019. Understanding credibility judgements for web search snippets. *Aslib Journal of Information Management* (2019).

[6] Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd international conference on computational linguistics: Posters*. Association for Computational Linguistics, 276–284.

[7] Susannah Fox and Maeve Duggan. 2013. Health online 2013. *Health* 2013 (2013), 1–55.

[8] Krisandra S Freeman and Jan H Spyridakis. 2004. An examination of factors that affect the credibility of online health information. *Technical Communication* 51, 2 (2004), 239–263.

[9] Lisa M Given. 2008. *The Sage encyclopedia of qualitative research methods*. Sage publications.

[10] Yongqi Gu. 2014. To code or not to code: Dilemmas in analysing think-aloud protocols in learning strategies research. *System* 43 (2014), 74 – 81. https://doi.org/10.1016/j.system.2013.12.011 Language Learning Strategy Research in the Twenty-First Century: Insights and Innovations.

[11] Alexander Haas and Julian Unkel. 2017. Ranking versus reputation: perception and effects of search result credibility. *Behaviour & Information Technology* 36, 12 (2017), 1285–1298.

[12] Julian PT Higgins. 2008. Cochrane handbook for systematic reviews of interventions version 5.0. 1. The Cochrane Collaboration. *http://www. cochrane-handbook. org* (2008).

[13] Eun Hwa Jung, Kim Walsh-Childers, and Hyang-Sook Kim. 2016. Factors influencing the perceived credibility of diet-nutrition information web sites. *Computers in Human Behavior* 58 (2016), 37–47.

[14] Jatin Kaicker, Wilfred Dang, and Tapas Mondal. 2013. Assessing the Quality and Reliability of Health Information on ERCP Using the DISCERN Instrument. *Health Care: Current Reviews* (2013), 1–4.

[15] Yvonne Kammerer, Ivar Bråten, Peter Gerjets, and Helge I Strømsø. 2013. The role of Internet-specific epistemic beliefs in laypersonsâĂŹ source evaluations and decisions during Web search on a medical issue. *Computers in Human Behavior* 29, 3 (2013), 1193–1203.

[16] 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.

[17] Hannu Kuusela and Paul Pallab. 2000. A comparison of concurrent and retrospective verbal protocol analysis. *The American journal of psychology* 113, 3 (2000), 387.

[18] Annie Lau and Enrico Coiera. 2008. Impact of web searching and social feedback on consumer decision making: a prospective online experiment. *Journal of medical Internet research* 10, 1 (2008), e2.

[19] QSR International Pty Ltd. Version 12, 2018. *NVivo qualitative data analysis software*.

[20] Teun Lucassen, Rienco Muilwijk, Matthijs L Noordzij, and Jan Maarten Schraagen. 2013. Topic familiarity and information skills in online credibility evaluation. *Journal of the American Society for Information Science and Technology* 64, 2 (2013), 254–264.

[21] Teun Lucassen and Jan Maarten Schraagen. 2010. Trust in wikipedia: how users trust information from an unknown source. In *Proceedings of the 4th workshop on Information credibility*. 19–26.

[22] Andrea P Marshall, Sandra H West, and Leanne M Aitken. 2011. Preferred information sources for clinical decision making: critical care nursesâĂŹ perceptions of information accessibility and usefulness. *Worldviews on Evidence-Based Nursing* 8, 4 (2011), 224–235.

[23] Andrea P Marshall, Sandra H West, and Leanne M Aitken. 2013. Clinical credibility and trustworthiness are key characteristics used to identify colleagues from whom to seek information. *Journal of Clinical Nursing* 22, 9-10 (2013), 1424–1433.

[24] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica* 22, 3 (2012), 276–282.

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

[26] Soo Young Rieh. 2002. Judgment of information quality and cognitive authority in the Web. *Journal of the American society for information science and technology* 53, 2 (2002), 145–161.

[27] Neil J Salkind. 2010. *Encyclopedia of Research Design*. Vol. 1. SAGE.

[28] Katja Schmidt and Edzard Ernst. 2004. Assessing websites on complementary and alternative medicine for cancer. *Annals of Oncology* 15, 5 (2004), 733–742.

[29] Tali Sharot, Alison M Riccardi, Candace M Raio, and Elizabeth A Phelps. 2007. Neural mechanisms mediating optimism bias. *Nature* 450, 7166 (2007), 102.

[30] Elizabeth Sillence, Pam Briggs, Lesley Fishwick, and Peter Harris. 2004. Trust and mistrust of online health sites. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 663–670.

[31] Elizabeth Sillence, Pam Briggs, Peter Harris, and Lesley Fishwick. 2006. A framework for understanding trust factors in web-based health advice. *International Journal of Human-Computer Studies* 64, 8 (2006), 697–713.

[32] Elizabeth Sillence, Pam Briggs, Peter Richard Harris, and Lesley Fishwick. 2007. How do patients evaluate and make use of online health information? *Social science & medicine* 64, 9 (2007), 1853–1862.

[33] Thanh Tin Tang, Nick Craswell, David Hawking, Kathy Griffiths, and Helen Christensen. 2006. Quality and relevance of domain-specific search: A case study in mental health. *Information Retrieval* 9, 2 (2006), 207–225.

[34] Teresa L Thompson. 2014. *Encyclopedia of health communication*. Sage Publications.

[35] Maarten W. van Someren, Yvonne F. Barnard, and Jacobijn A.C. Sandberg. 1994. *The Think Aloud Method: A Practical Guide to Modelling Cognitive Processes*. Academic Press, London.

[36] Ryen White. 2013. Beliefs and biases in web search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 3–12.

[37] Ryen W White. 2014. Belief dynamics in Web search. *Journal of the Association for Information Science and Technology* 65, 11 (2014), 2165–2178.

[38] Ryen W White and Ahmed Hassan. 2014. Content bias in online health search. *ACM Transactions on the Web (TWEB)* 8, 4 (2014), 25.

[39] Ryen W White and Eric Horvitz. 2015. Belief dynamics and biases in web search. *ACM Transactions on Information Systems (TOIS)* 33, 4 (2015), 18.