How Misinformation Density Affects Health Information Search

Anonymous Author(s)
Submission Id: 2013

ABSTRACT

Search engine results can include misinformation that is inaccurate, misleading, or even harmful. But people may not recognize or realize false information results when searching online. We suspect that the percentage of search results with misinformation (misinformation density) may influence people's search activities, learning outcomes, and user experience. We conducted a zoom-mediated "lab" user study to examine this matter. The experiment used a between-subjects design. We asked 60 participants to finish two health information search tasks using search engines with *High*, *Medium*, or *Low* misinformation density levels. To create these experimental settings, we trained task-dependent text classifiers to manipulate the number of misinformation and correct information results displayed on SERPs. We collected participants' search activities, responses to pre-task and post-task surveys, and answers to task-related factual questions before and after searching.

Our results indicate that search result misinformation density strongly affects users' search behavior. High misinformation density makes people search more frequently, use longer queries, and click on more results. However, such increased search activities did not lead to better search outcomes. Participants using the *High* misinformation density search engine answered factual questions less accurately and learned very limitedly from a search session than the two other systems. Moreover, participants in systems with a balanced amount of misinformation and correct information (*Medium*) could learn factual knowledge as effectively as others in a system with little misinformation (*Low*). Surprisingly, participants using different misinformation density systems did not rate their search experience or perceived goodness of search systems with significant differences. Our experiment and findings have disclosed the effects of misinformation density on health information search and offered insights to improve online health information search.

CCS CONCEPTS

• **Computer systems organization → Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

health information search, misinformation, false information, interactive information retrieval

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 ACM 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.
Woodstock '18, June 03–05, 2018, Woodstock, NY
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

ACM Reference Format:
Anonymous Author(s). 2018. How Misinformation Density Affects Health Information Search. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Web search engines have become the ubiquitous gateway of information for almost everyone on almost everything. People search for various information online, ranging from everyday topics to those with potentially huge (negative) impacts on individuals (e.g., how to commit suicide [26]) and society (e.g., presidential election [13]). However, the web is full of misinformation and even disinformation. But current web search engines cannot provide consistent support for filtering out such misinformation or informing users of the risks across all topics. Online misinformation significantly affects people's life and society.

Here we focus on online health information seeking through web search engines. With the development of the Internet, more and more people choose to acquire health information online [28]. For example, online health information helps people be more engaged in health decision-making and improves patient-physician relationships. Also, online health information can complement expert opinion and support people to make health decision [27]. However, many public health researchers are also very concerned about the potential of the Internet as a tool for disseminating health-related information due to health misinformation [1, 9]. The credibility of online health information varies greatly. People may also be overwhelmed with extraneous and often conflicting online health information [8, 12, 18]. Online health misinformation can delay or prevent effective care, harming people's health and even their lives. In some cases, the accumulation of these unfounded stories, pseudoscientific beliefs, and conspiracy theories can spark social movements, such as the anti-vaccine movement, with far-reaching consequences on public health [23, 47]. However, we know little about how online health misinformation influences people's online health information search through web search engines.

We conducted a zoom-mediated "lab" experiment to study the influence of misinformation on health information search. We asked participants to search for health information in search systems with different misinformation density levels. Here we define search results' misinformation density as the percentage of misinformation results among topically relevant ones. We built task-dependent text classifiers to categorize search results into misinformation, correct information, and irrelevant (non-relevant) ones. We filtered out misinformation or correct information results to create search systems with *High*, *Medium*, or *Low* misinformation density levels. Also, we collected participants' search behavior logs and measured their correctness for answering task-related factual questions both before and after search tasks. The rest of this article introduces our experiments and findings.

2 RELATED WORK

2.1 Health Information Seeking

The rapid development of the Internet has significantly changed the way people access information. In the health care domain, in terms of finding and receiving health information, people are beginning to actively acquire health information through internet searches rather than passively through books and magazines [28, 48]. Also, people search the web for health information not only for the sake of the information itself but may support their health decisions [41, 44].

Previous studies proposed models to understand the effect of health information seeking, health information sources, and health information overload on information anxiety (psychological organism), and consequent behavioral response, information avoidance during the global health crisis (COVID-19) [41]. Researchers also revealed the barriers, facilitators and demographics that influence patients' disclosure of online health information during consultations, as well as the various mechanisms patients use to reveal these findings. They also demonstrated the possible impact of online information on patient healthcare, such as patient-physician communication and relationships [44].

Previous studies have shown that online health information search can can support patients' health decisions [44]. However, at the same time, there is a lot of irrelevant, incomplete and even incorrect health information on the Internet, which can be confusing for people [4, 35, 45]. Previous study showed a range of search and appraisal skills among participants, with many reporting a limited awareness of how they found and evaluated Internet-based information on medicines [35].

Also, people can have misconceptions about health information due to the lack of professional knowledge [5]. Berland et al. found that coverage of key information on English- and Spanish-language Web sites is poor and inconsistent, although the accuracy of the information provided is generally good. A high reading level is required to understand web-based health information. [5]. All these factors interfere with the use of Internet health information and cause adverse effects on people's health [35].

Previous studies have also examined how people find and appraise health information online [14], and its impact on health-related behaviors [3, 39]. Sillence et al. conducted a longitudinal study in which fifteen women faced decisions concerning Menopause and hormone replacement therapy (HRT). They found that they used online materials to generate and test hypotheses and theories about HRT, despite their limited ability to process some of the information. The women then reported combining online advice with offline advice from friends, family and physicians and felt confident in their final decision. The women felt that the Internet influenced their decision making and improved communication with their physicians [39]. Ayers et al. found out that the more frequently a person uses the Internet as a source of health information, the more likely they are to change their health behavior [3]. Researchers also explored the cognitive strategies and attitudes of health information seekers, as well as the influence of factors such as cognitive strategies and technical skills on users' ability to filter and analyze information and perceive health problems [15].

2.2 Online Misinformation

Misinformation is false, inaccurate, or misleading information that is communicated regardless of an intention to deceive [2, 31]. Misinformation includes false rumors, pranks, and misleading use of facts, etc. In this study, we define misinformation in the health field as information that is contrary to the epistemic consensus of the scientific community regarding a phenomenon [43]. There is also much disinformation in the health area, which is a part of misinformation and is deliberately created and circulated to gain profits, power, or reputation. We did not tease apart disinformation from misinformation and mainly focused on whether the information is true or false based on current research.

There is a lot of misinformation in online health information, which has many negative effects on people who look to the Internet for health support [12]. Misinformation about healthcare online can delay or prevent effective care, harming people's health, and those misbeliefs can even lead to social movements, such as the anti-vaccine movement, with profound effects on public health [23, 47].

Previous research has found that health misinformation in web search affect people's search behavior and their health decisions. Researchers noticed that even when careful and focused on the task, people can still be affected by biased search engine results and tend to make decisions consistent with the bias [16]. Although most people say that online searches make them better at making decisions about health issues, there are plenty of challenges to finding correct information online [20]. Participants also reported skepticism about the validity of health information, indicating that they were unable to identify reliable strategies for evaluating health information on the Internet [20, 29, 37]. A previous study investigated health information search by asking laypeople to search the Internet for information under a hypothetical medical emergence scenario. The authors found that regardless of their search skills and experience, laypeople may have difficulties in searching for health information. And these challenges might be linked to their prior beliefs based on their life experience and domain knowledge [16]. These difficulties may be related to constructing and evaluating hypotheses based on their domain knowledge [25, 34]. A large amount of information that may be retrieved in a health information search makes it difficult to distinguish between true and false information and to interpret conflicting findings and facts [43].

2.3 Search as Learning

Information scientists have observed that people of all ages increasingly turn to search engines for health information to support their healthcare decisions [43]. And various studies have indicated that searching is not only a tool designed to help people find information needed for learning but also a learning process itself, in which people will search to critically analyze, comprehend, integrate, evaluate and use information [17, 32, 33, 46]. Previous studies have explored search as learning in different contexts, such as academia and business [49, 51]. Several studies have investigated potential indicators of search as learning and the relationship between search behavior and learning outcomes by designing search tasks [11, 21]. Some learning-related tasks were conducted in a health context, exploring the effects of personal stress levels and time constraints on searcher behavior, but did not delve into the users' learning process of in health information search [30].

Online health information search can be considered a learning process [7, 21, 52]. Complex problems in the health-related field often involve conflicting evidence and have various solutions. Previous studies have explored people's evaluation and selection of search results, comparison and integration of information from various websites during health information search and learning [10, 36]. Therefore, during the online health information searching process, health misinformation in search results can make information processing very difficult for users [22]. Users need to evaluate and compare different or even contradictory information to gain a correct understanding of health issues [38, 42]. When search results have increased misinformation, users are exposed to more conflicting information and need to compare and discriminate information more often. Previous research has found that as the level of cognitive complexity of a task increases, users spend more time, issue more queries, click on more search results, and visit more URLs [24, 40, 50].

Our research extends prior work in several ways. We are the first to examine the effects of misinformation density on people's health information search behaviors using an experimental approach. Second, we explored how misinformation density levels affect users' learning process and correctness of answering factual questions. Our study enriches the current understanding of misinformation in health information search.

3 EXPERIMENTS

We designed a lab-based user study to examine the influence of the amount of misinformation (*misinformation density*) in search results on people's search behavior and the effectiveness of learning factual information from search. We trained task-dependent task classifiers to categorize results into misinformation, correct information, and irrelevant ones, such that we can manipulate the ratio of misinformation results in search engine result pages (SERPs). We measured participants' correctness of answering task-related factual questions, search experience, and perceived goodness of the systems through Likert-style questions. We recruited 60 college students and moderated the experiments through zoom-mediated "lab" experimental sessions.

3.1 Misinformation Density

We are interested in the influence of misinformation on health information seeking through general web search engines. Specifically, our study focuses on misinformation density in web search results and how it affects people's search behaviors and learning outcomes from search. Here we define search results' *misinformation density* as the proportion of displayed topically relevant search results containing substantial misinformation (excluding the irrelevant ones). We have three hypotheses:

- H1—Misinformation density affects people's search activities (e.g., how they formulate queries and click results) in health information search.
- H2—Misinformation density affects people's correctness of answering factual questions after health information search.
- H3—Misinformation density affects search user experience and perceived goodness of the search systems.

Table 1: Two search tasks used in our experiment.

	Task Description
Task 1 (Aspirin)	Your friend shared an article about **whether Aspirin helps with dandruff and white hair**. After reading this article, you hope to search for more information to know this topic better.
Task 2 (Vitamin B12)	Your friend shared an article about **whether Vitamin B12 in functional drinks increase liver and kidney burden**. After reading this article, you hope to search for more information to know this topic better.

We design experiments accordingly to examine the independent and dependent variables and hypotheses.

3.2 Experimental Design and Ethics

Our experiment uses a between-subjects design. The independent variable is search systems with different misinformation density levels (*Low*, *Medium*, or *High*). The dependent variables correspond to the three hypotheses and have included search behavior measures, questions measuring participants' knowledge to answer factual questions, and responses to search experience questions. We assign each participant to the search system of a specific misinformation density level (*Low*, *Medium*, or *High*) to finish two search tasks. We record their search activities and answers to the questions before and after each search task.

Our experimental design expects a lab-based setting. However, we conducted experiments through zoom-mediated sessions to reduce the health risks of participants during the COVID-19 pandemic. Moreover, participants may expose themselves to health misinformation during the experiment, leading to potentially harmful decisions if they experience similar real-life issues. To reduce possible risks, we specifically instructed participants after the experiments that they may have read misinformation and should consult health professionals if they experience similar real-file issues. Our experiment has received institutional IRB approval for human-subjects research ethics.

3.3 Health Information Search Tasks

We designed two health information search tasks. Table 1 shows the two tasks about Aspirin and Vitamin B12. The tasks simulate the scenario where people read a misinformation article shared by their friends on social media and then search for information to verify the correctness of the content. We have also provided the URL of the article along with the task description in experiments. We required the users to read the background article before they started searching for information. All participants finished the same two tasks, but we have rotated the sequence of the two tasks for different participants.

When we designed the tasks, we excluded some candidate ones through pilot studies. For example, some tasks did not seem credible for regular college students (our participants) even without an information search. We suspect a search engine's misinformation density levels are less likely to impact these tasks. Also, we excluded tasks if they had too little misinformation or correct information results from the search API we used for building our experimental systems. This is to ensure that we can successfully manipulate the three misinformation density levels in search systems. Moreover, we excluded tasks we were not able to verify the correctness of its relevant information.

3.4 Search Systems

Our experimental search systems resemble a commercial web search engine in outlook and functionality, except we can manipulate search results' misinformation density at three levels (*Low*, *Medium*, and *High*). Users may submit and reformulate any text queries and click on any search results. The search systems provide filtered search results from the Bing search API, but we only retain regular results (the 10-blue links). We do not show other SERP elements such as knowledge cards and vertical results to exclude their possible effects on the experiments. Also, it is unclear how to manipulate the misinformation density levels of multiple different SERP elements consistently. Our systems display search results in the same outlook as they would appear on Bing, including font size, style, and color. The systems with different misinformation density levels have the same outlook and search functionality. The participants could not distinguish the systems of different misinformation density levels only from the search tasks. We also did not inform them of the treatment of our experiments.

We built task-dependent text classifiers to manipulate search results' misinformation density. For each task, we trained a classifier to divide its search results into three categories: misinformation, correct information, or irrelevant. After we designed the two search tasks, two authors independently formulated ten search queries for each task. We retrieved the top 10 results for each query and manually annotated the results by the three categories. The size of the annotated dataset is about 300 results for each task. We trained a Naïve Bayes classifier using bag-of-words features for each task to automatically determine the categories of new retrieval results. The model only used results' SERP abstracts for classification as we found it difficult to ensure timely response if we classify based on the result link's full text. The trained classifiers had achieved accuracy scores of 0.8 and 0.64 on the annotated datasets for tasks 1 and 2 in a cross-validation setting.

We apply the trained classifiers to filter search results from the Bing search API in real-time to manipulate misinformation density. The *High* misinformation density system removes search results that are judged or classified as correct information while retraining the relative sequence of the other results. In contrast, the *Low* misinformation density system removes results that are judged or classified as misinformation. The *Medium* misinformation density system aims to provide a balanced number of misinformation and correct results. We remove misinformation or correct information results (judged or classified ones) such that the number of displayed misinformation and correct results on any SERPs do not differ more than one.

As our results in Section 5.1 show, our manipulation of search results' misinformation density was very successful. If we exclude irrelevant results, the percentage of misinformation search results among topically relevant ones displayed on the SERPs is 86% (*High*), 53% (*Medium*), and 9% (*Low*), respectively.

3.5 Survey and Factual Questions

We use pre-task and post-task surveys to measure participants' expectations, search experience, and perceived goodness of the systems. Table 2 lists the questions and items for the surveys. The pre-task survey asked participants' prior experience with the topic and their expectations of the task. In contrast, the post-task survey asked participants' actual search experience and their perceived goodness of the search engines. All the questions and items are Likert style, and we map participants' responses to 1–5 numeric values.

We have also developed ten factual questions for each task to measure participants' knowledge and ability to answer factual questions correctly. Table 3 shows one example question for each task (we do not list all questions due to limited space). We asked the same list of questions twice before and after a search task to examine the effects of the search process on their responses. All the factual questions are statements we selected manually from the result webpages (some of them are correct, and some are incorrect). Participants need to answer to what extent they agree with these statements using a 5-point Likert scale from *Strongly Disagree* (1) to *Strongly Agree* (5).

3.6 Experiment Procedure

We conducted the experiments through zoom due to the COVID-19 pandemic. We invited participants to remotely control the experiment coordinator's computer to finish the tasks. We used the same computer setting (e.g., browser, resolution) for all participants. During the experiment, we set the monitor's resolution to 1280×800 such that all participants' monitors are compatible with our setting. The environment is similar to a lab-based experiment where all participants come to the same place and use the same computer to finish tasks, except that we moved the procedure online.

For each task, participants needed to finish the following steps:

- *Introduction*—Participants needed to follow the task description and read the provided misinformation article. Participants did not know whether the article contained misinformation, but we instructed them that the article did not necessarily provide credible information.
- *Pre-task Questions*—Participants needed first to finish the pre-task survey questions and then the factual questions. We instructed participants to answer the factual questions according to their common sense and prior knowledge or information from the background article they read in the previous step.
- *Search* (10 minutes)—Participants needed to search for 10 minutes using the assigned experimental search engine. Participants could submit as many queries as they wanted and click on any search results during this period. However, we specifically instructed them not to use other search engines during the task.

Table 2: Pre-task and post-task survey questions.

Type	Acronym	Questions and Options
Pre-task	`preFAM`	Are you familiar with the topic of the task before our experiment? *Very Unfamiliar* (1)—*Very Familiar* (5)
	`preEXPL`	Have you explored the topic of the task before our experiment? *No Exploration At All* (1)—*Lots of Exploration* (5)
	`preINTR`	Are you interested in the topic of the task? *Not Interested At All* (1)—*Very Interested* (5)
	`preCRED`	Do you think the article you just read is credible? *Not Credible At All* (1)—*Very Credible* (5)
	`preDIFF`	Do you expect it is difficult to finish the task through information from a search engine? *Very Easy* (1)—*Very Difficult* (5)
	`preCAPB`	Do you expect you have enough knowledge and skills to finish the task? *Not Capable At All* (1)—*Very Capable* (5)
Post-task	`postSUFF`	Have you found sufficient information in the task you have just finished? *Very Insufficient* (1)—*Very Sufficient* (5)
	`postEXPL`	Have you fully explored the topic in the task you have just finished? *No Exploration At All* (1)—*Lots of Exploration* (5)
	`postEFFT`	Have you tried hard to collect information on the task you have just finished? *Not At All* (1)—*Lots of Effort* (5)
	`postUSEF`	Do you believe the search engine provided useful information for finishing the task? *Not Useful At All* (1)—*Very Useful* (5)
	`postCRED`	Do you believe the search engine provided credible information for finishing the task? *Not Credible At All* (1)—*Very Credible* (5)
	`postCONF`	Are you confident about your answers to the post-task tests you have just finished? *Not Confident At All* (1)—*Very Confident* (5)

Table 3: Example factual questions for each task.

	Example Factual Questions and Items	
Task 1	False	Aspirin can soften blood vessels and promote the circulation of blood, which provides sufficient nutrition to hair follicles, accelerate the production of Melanocytes, and eventually make white hair black again. *Strongly Disagree* (1)—*Strongly Agree* (5)
Task 2	True	Like other water-soluble vitamins, Vitamin B12 will be excreted through urine when our human body has acquired more than needed. *Strongly Disagree* (1)—*Strongly Agree* (5)

- *Post-task Questions*—Participants need to finish the factual questions again. Then, they needed to answer the post-task survey questions.

We recruited 60 participants (45 female and 15 male) through online ads posted to a University's social media groups and forums. All the participants are students of that University, and their age ranges from 19 to 24. Their major areas of study included sciences, social sciences, computer and information sciences, engineering, and humanities. We have intentionally excluded students studying medical or health sciences as we expected their background knowledge to be much different from others. More specifically, we expect some of them are less likely to be affected by misinformation during health information search due to their strong domain expertise. We provided the same base monetary compensation to all participants. To motivate participants to search and learn, we informed them before the experiment that we would examine their answers and provide the top 10% performed participants with an incentive (twice as much as the regular compensation).

4 COLLECTED DATA

In total, we have collected data from 60 participants and 120 search sessions. We have collected data from 20 participants and 40 search sessions for each misinformation density level because we used a between-subjects design. On average, the participants issued 9.8 queries (counting multiple SERP pages for the same query) and clicked on 8.1 results in a session. To verify the effectiveness of our experiment manipulation, we have judged all the displayed SERP results for all users in all tasks. We manually categorized these 3,284 SERP results into misinformation, correct information, and irrelevant ones.

5 RESULTS

We report and analyze our experimental results in this section. We compare results from systems of three misinformation density levels using ANOVA with the Bonferroni correction post-hoc tests. All the dependent variables' distributions are within the acceptable range for normal distributions, with skewness between $(-2, 2)$ and kurtosis between $(-7, 7)$ [6, 19]. We use *, **, and *** to indicate significant differences at $p < 0.05$, 0.01, and 0.001 in our figures and tables, respectively.

5.1 Displayed and Clicked Results

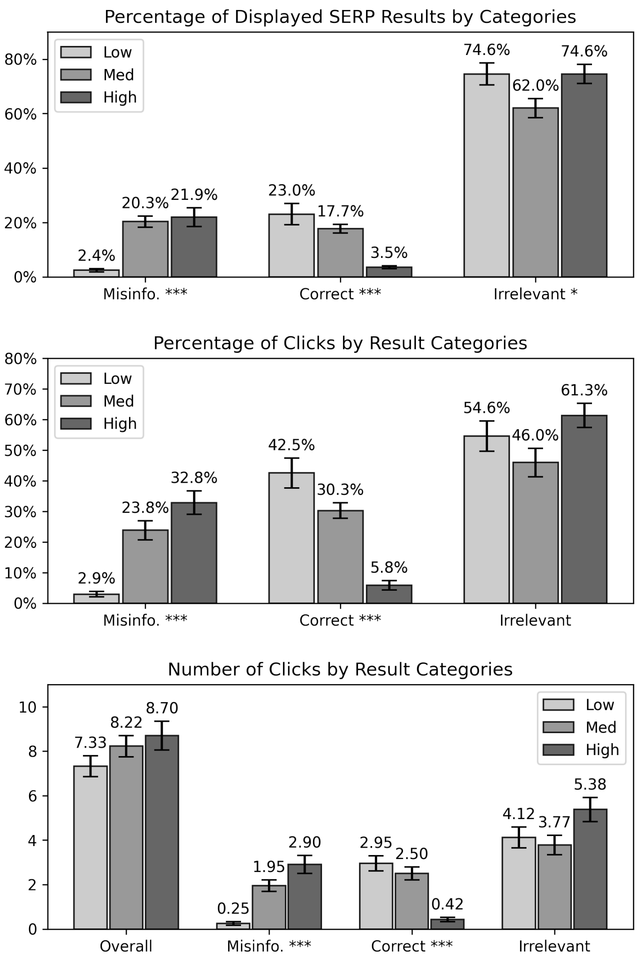
Results show that our experimental manipulation of misinformation density in search results was successful. The actual percentage of misinformation search results among topically relevant ones in the *High*, *Medium*, and *Low* misinformation density systems is 86%, 53%, and 9%, respectively. The vast differences in actual misinformation density levels also caused significant differences in users' actual access to misinformation and correct information results.

Before examining our experimental results, we first validate whether our manipulation of the search system's misinformation density level is successful. We have manually annotated all the results displayed on the SERPs by all participants in all tasks. Figure 1 shows the percentage of SERP results that are misinformation, correct information, or irrelevant.

The results verified that our manipulation of misinformation density levels in this experiment was successful. The *High* misinformation density system displayed 21.9% false information results but only 3.5% correct information ones. In contrast, the *Low* misinformation density system showed 2.4% false information results but 23.0% correct information ones. The *Medium* misinformation density system retrieved a roughly balanced number of false and correct information results (20.3% and 17.7%, respectively). If we exclude irrelevant results, the percentage of misinformation search results among topically relevant ones is 86%, 53%, and 9%, respectively.

The ANOVA tests have found significant differences ($p < 0.001$) among the three levels regarding the percentage of displayed misinformation and correct results. Post-hoc tests found *High*>*Low* and

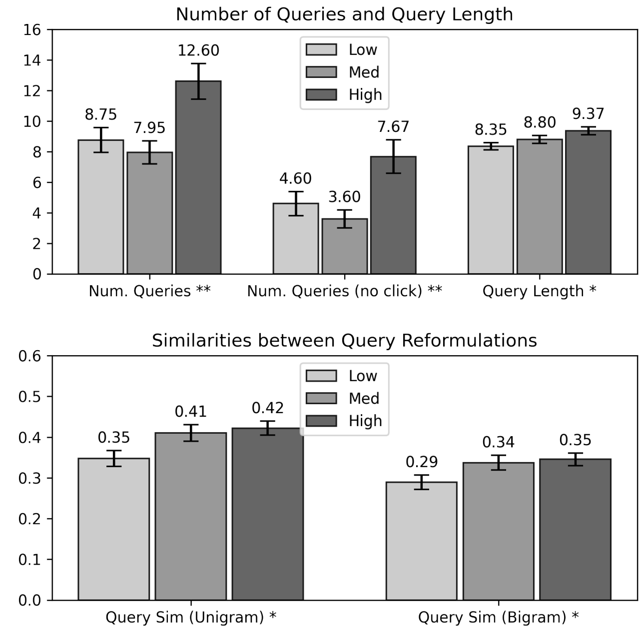
Figure 1: Displayed and clicked results that are misinformation, correct information, or irrelevant among systems with *High*, *Med*, and *Low* misinformation density levels.



Medium>*Low* for misinformation results ($p < 0.001$), but no significant differences between *High* and *Medium*. Similarly, post-hoc tests found *Low*>*High* and *Medium*>*High* for correct information results ($p < 0.001$), but no significant differences between *Low* and *Medium*.

Unsurprisingly, the three systems with vastly different misinformation density levels also caused significant differences in click behaviors. As Figure 1 also shows, the percentage of clicked results resembles those of the displayed SERP results. Systems with three different misinformation levels have significant differences in the ratio of clicked misinformation and correct results ($p < 0.001$; the post-hoc tests between each pair are also significant at least at 0.01 level). In contrast, we found no significant differences in the ratio of clicked irrelevant results among the three levels ($p = 0.06$). These results further indicate that our successful manipulation of misinformation density levels had also caused significant impacts on searchers' actual access to the results.

Figure 2: Query behavioral measures for systems with *High*, *Med*, and *Low* misinformation density levels.



To conclude, the results in this section have verified the success of our experimental manipulation and effects. Particularly, we may reasonably interpret participants' responses to the factual questions as the effects of misinformation, considering that the three systems had significantly affected searchers' access to misinformation and correct information results.

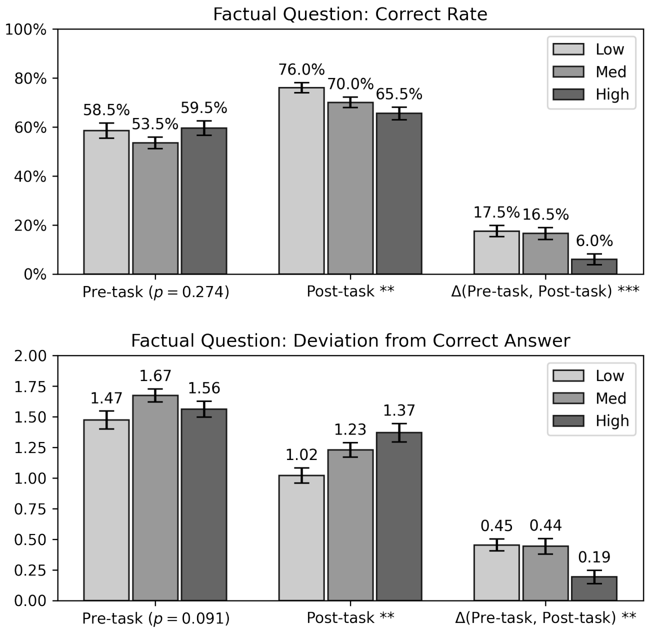
5.2 Search Queries

Search queries are the primary inputs of users when they interact with search engines. Our results show that search engines' misinformation density level significantly affects users' search queries from various aspects. Here we look into three measures: the number of queries, query length, and similarity of query reformulation pairs.

We found that *High* misinformation density level leads to significantly more search queries than the *Medium* and *Low* levels. This means that participants in the *High* misinformation density system had searched more frequently than those in *Low* and *Medium* systems (since we let them search for 10 minutes in all settings). The differences between *High* and *Medium* and *High* and *Low* are significant at 0.01 level in post-hoc tests, while we found no significant difference between *Medium* and *Low*. The finding is consistent regardless of whether we count the number of raw search queries or unique ones. Note that when searchers turned to the second page of results, our systems would count it as another search query.

In addition, we observed a significant effect of misinformation density level on query length ($p < 0.05$). The post-hoc tests have found significant differences for *High*>*Low* ($p < 0.05$). This shows that searchers tend to use longer queries (a possible indicator of more complex queries) in the *High* misinformation density condition.

Figure 3: Correctness of answering factual questions before and after search tasks in systems with *High*, *Med*, and *Low* misinformation density levels.



Moreover, we examine the similarity of users' query reformulations as previous studies found that a high similarity of query reformulations may indicate struggling search sessions. Here we define a query reformulation pair as a user switching from one search query to a next one different from the previous query in content. We compare the two queries in each query reformulation pair by their unigram and bigram similarities. Figure 2 compares the average query reformulation similarities of the three levels. We observed significant effects of misinformation level on both unigram and bigram similarities ($p < 0.05$). More specifically, the post-hoc tests found significant differences of *Low* versus *Medium* and *High* settings in both unigram and bigram similarity ($p < 0.05$), but no significant differences between *Medium* and *High*. The significantly lower query reformulation similarity in the *Low* misinformation density system suggests that reducing misinformation in search results may help users avoid struggling search sessions.

5.3 Factual Questions

We further examine participants' correctness of answering the factual questions before and after each search task. We found that higher misinformation density will significantly reduce the chances of answering factual questions correctly and make participants' responses deviate from the correct answers to a greater extent. In addition, *High* misinformation density in search results prevents people from learning correct information effectively from the search session.

Here we look into two measures. The first measure is the correct rate of answering factual questions. We divide the participants' 5-point Likert-style responses into correct or false responses. For a true statement, we count responses 4 and 5 as correct, and for a false statement, we count responses 1 and 2 as correct ones. Then, we measure participants' correct rates on the ten factual questions in each task. The second measure is the average deviation of participants' responses from the correct answers. For true and false statements, we count their "correct" responses as 5 and 1, respectively. Then, we measure the absolute difference between participants' actual responses and the "correct" ones. Figure 3 reports the results for pre-task and post-task factual questions and their differences Δ(Pre-task, Post-task).

Table 4: Example post-task factual queries that have significant differences among systems of *High*, *Med*, and *Low* misinformation density levels.

Question	Low	Med	High	ANOVA *p*-value
Post-task Factual Questions (Examples)				
Task2-FQ1	2.40 (0.29)	2.20 (0.35)	3.35 (0.37)	0.044 *
Task2-FQ2	1.30 (0.16)	1.25 (0.16)	1.95 (0.30)	0.049 *
Task2-FQ8	1.50 (0.17)	1.85 (0.26)	2.60 (0.26)	0.005 **
Δ(Pre-task, Post-task) (Examples)				
Task2-FQ1	−1.75 (0.34)	−1.95 (0.34)	−0.80 (0.30)	0.036 *
Task2-FQ9	1.50 (0.28)	1.20 (0.28)	0.45 (0.34)	0.047 *

Experimental results show that participants' correctness of answering factual questions does not differ significantly in the three systems before the search tasks. However, their correctness varies significantly after the search tasks regardless of the correct rate ($p < 0.01$) or the average deviation from the correct answers ($p < 0.01$). Post-hoc tests also found significant differences between each adjacent misinformation level (*High* vs. *Medium* and *Medium* vs. *Low*) at least at 0.05 level. These results suggest that higher misinformation density in search results prevents people from understanding factual information correctly.

Moreover, we observed that misinformation density levels also significantly affect the differences of participants' pre-task and post-task correctness on the factual questions ($p < 0.001$ for correct rate and $p < 0.01$ for deviation from correct answers). However, we note that post-hoc tests only found significant differences between the *High* misinformation density level and other settings, but not between *Medium* and *Low* levels. These findings are consistent for both correct rate and deviation from the correct answers. These results indicate that a mixture of misinformation and correct information did not significantly affect people's learning of factual information from search results. But a vast majority of misinformation results with little correct information is harmful, preventing people from learning correct information.

We further examined each factual question for each task separately. In total, this means 20 questions, as we have included 10 for each task. For the pre-task responses, none of the questions showed significant differences among systems of different misinformation density levels. In contrast, we observed three questions for task 2 but none for task 1 with significant differences in post-task responses. We note that most questions' ratings did not show significant differences among systems with different misinformation density levels. We suspect this is most likely because of the limited sample size since the aggregated correct rate and deviation

Table 5: Pre-task and post-task survey responses (mean and standard error of the mean).

Question	Low	Med	High	ANOVA *p*-value
preFAM	1.40 (0.11)	1.70 (0.16)	1.90 (0.13)	0.037 *
preEXPL	1.25 (0.10)	1.38 (0.13)	1.57 (0.11)	0.122
preINTR	3.60 (0.16)	3.92 (0.13)	3.85 (0.12)	0.224
preCRED	2.58 (0.13)	2.70 (0.13)	2.77 (0.15)	0.579
preDIFF	2.50 (0.09)	2.83 (0.14)	2.62 (0.15)	0.200
preCAPB	3.15 (0.15)	3.25 (0.18)	3.17 (0.13)	0.892
postSUFF	3.75 (0.13)	3.52 (0.14)	3.38 (0.14)	0.164
postEXPL	3.48 (0.16)	3.50 (0.15)	3.30 (0.17)	0.631
postEFFT	4.25 (0.12)	4.17 (0.10)	4.40 (0.09)	0.317
postUSEF	4.15 (0.13)	4.05 (0.12)	3.77 (0.12)	0.096
postCRED	3.77 (0.13)	3.40 (0.11)	3.45 (0.13)	0.063
postCONF	3.70 (0.14)	3.42 (0.15)	3.70 (0.13)	0.276

from correct answers have shown clear and consistent significant differences. However, the several observed significant differences for post-task factual questions are highly consistent with the aggregated measures we compared before.

To sum up, results in this section showed that misinformation density in our experiments has significantly affected participants' correctness of answering factual questions after search and their effectiveness of learning this factual information during the search process.

5.4 Survey Questions

Despite significantly affecting search behavior and learning effectiveness, misinformation density did not seem to influence participants' search experience or perceived goodness of the systems.

Table 5 reports the mean responses of each question in pre-task and post-task surveys in the three systems. Participants' responses for most pre-task questions did not show significant differences between the three misinformation density levels except for preFAM (prior familiarity about the task topic). Note that we assigned participants to different settings randomly. Thus the observed significant differences for preFAM are most likely a random outcome.

In contrast, however, it is surprising that participants' responses to the post-task survey questions did not show any significant differences among the three misinformation density levels, although their search behavior and factual question correctness varied greatly. According to the mean ratings, we suspect that High misinformation density may prevent people from acquiring sufficient information (postSUFF) and exploring tasks fully (postEXPL) and cost them more effort (postEFFT). But none of the differences are not statistically significant at 0.05 levels. Regarding participants' perceived goodness of the system, we found that they did rate the system in the High misinformation density level as less useful (postUSEF), less credible (postCRED), and felt less confident about their answers (postCONF). Nonetheless, the differences were not significant at the 0.05 level.

Table 6: Summary of major findings.

High (86%) vs. *Medium* (53%)	• Clicked more misinformation results but fewer correct information results. • Submitted over 50% more search queries during the same 10-minute search session. • Reduced the correct rate for answering factual questions after a search task by 4.5%. • Prevented learning factual information throughout a session by 10.5% correct rate.
Low (9%) vs. *Medium* (53%)	• Clicked fewer misinformation results. • Lower similarity of query reformulations. • Increased the correct rate for answering factual questions after a search task by 6%. • No differences in the learning outcome throughout a search session.

6 CONCLUSION

To conclude, we used a lab-based user study to examine the effects of misinformation density on users' search activities, learning outcomes, and search experience. Table 6 summarizes our findings. Our study has made the following contribution:

First, we have examined the effects of misinformation density on people's health information search behaviors. Through novel and successful manipulation of misinformation during a web search, we have observed that a high misinformation density of search results may increase search effort (e.g., searching more frequently, using longer queries, and clicking more results). These observations support our hypothesis H1 and help understand user interaction with search engines of different misinformation density.

Second, we have studied how misinformation density levels affect people's correctness of answering factual questions and learning task-related knowledge from search. We have observed consistent and clear effects of misinformation density, making people answer factual questions less accurately and preventing them from learning online information effectively. Also, we found that participants could learn factual information comparably well in systems with Low and Medium misinformation density levels. This suggests that misinformation may not be too concerning if the search engine SERPs have provided a decent amount of correct information. These findings support our hypothesis H2 and help understand how search engine misinformation influences people and society's knowledge and learning outcome.

Third, our experimental results did not show any solid evidence supporting that misinformation density could influence people's search experience or the perceived goodness of the system (hypothesis H3). Even if such an effect exists, we suspect it is a relatively mild effect compared with those for search behaviors and the correctness of answering factual questions (H1 and H2).

We also acknowledge certain limitations of our study. For example, we have used a very limited number of tasks and included only 60 participants (though a regular number for a lab-based user study). We believe it also requires further studies to confirm some of our findings further.

REFERENCES

[1] Wasim Ahmed. 2018. *Using Twitter data to provide qualitative insights into pandemics and epidemics.* Ph. D. Dissertation. University of Sheffield.

[2] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31, 2 (May 2017), 211–36. https://doi.org/10.1257/jep.31.2.211

[3] Stephanie L. Ayers and Jennie Jacobs Kronenfeld. 2007. Chronic illness and health-seeking information on the Internet. *Health* 11, 3 (2007), 327–347. https://doi.org/10.1177/1363459307077547

[4] Mike Benigeri and Pierre Pluye. 2003. Shortcomings of health information on the Internet. *Health Promotion International* 18, 4 (2003), 381–386. https://doi.org/10.1093/heapro/dag409

[5] Gretchen K. Berland, Marc N. Elliott, Leo S. Morales, Jeffrey I. Algazy, Richard L. Kravitz, Michael S. Broder, David E. Kanouse, Jorge A. Muñoz, Juan-Antonio Puyol, Marielena Lara, Katherine E. Watkins, Hannah Yang, and Elizabeth A. McGlynn. 2001. Health Information on the InternetAccessibility, Quality, and Readability in English and Spanish. *JAMA* 285, 20 (05 2001), 2612–2621. https://doi.org/10.1001/jama.285.20.2612

[6] Barbara M. Byrne. 2016. *Structural Equation Modeling With AMOS: Basic Concepts, Applications, and Programming* (3 ed.). Routledge.

[7] Yijin Chen, Yiming Zhao, and ZiYun Wang. 2020. Understanding online health information consumers' search as a learning process: Search platform sequences, query reformulation and conceptual changes. *Libr. Hi Tech* 38 (2020), 859–881.

[8] Yen-Yuan Chen, Chia-Ming Li, Jyh-Chong Liang, and Chin-Chung Tsai. 2018. Health Information Obtained From the Internet and Changes in Medical Decision Making: Questionnaire Development and Cross-Sectional Survey. *Journal of Medical Internet Research* 20 (2018).

[9] Cynthia Chew and Gunther Eysenbach. 2010. Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. *PLoS ONE* 5 (2010).

[10] Kevyn Collins-Thompson, Preben Hansen, and Claudia Hauff. 2017. Search as Learning (Dagstuhl Seminar 17092). *Dagstuhl Reports* 7, 2 (2017), 135–162. https://doi.org/10.4230/DagRep.7.2.135

[11] Kevyn Collins-Thompson, Soo Young Rieh, Carl C. Haynes, and Rohail Syed. 2016. Assessing Learning Outcomes in Web Search: A Comparison of Tasks and Query Strategies. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval* (Carrboro, North Carolina, USA) *(CHIIR '16)*. Association for Computing Machinery, New York, NY, USA, 163–172. https://doi.org/10.1145/2854946.2854972

[12] Jose Yunam Cuan-Baltazar, Maria José Muñoz-Perez, Carolina Robledo-Vega, Maria Fernanda Pérez-Zepeda, and Elena Soto-Vega. 2020. Misinformation of COVID-19 on the Internet: Infodemiology Study. *JMIR Public Health Surveill* 6, 2 (9 Apr 2020), e18444. https://doi.org/10.2196/18444

[13] Robert Epstein and Ronald E. Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences* 112, 33 (2015), E4512–E4521. https://doi.org/10.1073/pnas.1419828112 arXiv:https://www.pnas.org/content/112/33/E4512.full.pdf

[14] Gunther Eysenbach and Christian Köhler. 2002. How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ* 324, 7337 (2002), 573–577. https://doi.org/10.1136/bmj.324.7337.573 arXiv:https://www.bmj.com/content/324/7337/573.full.pdf

[15] Markus A Feufel and S Frederica Stahl. 2012. What do Web-Use Skill Differences Imply for Online Health Information Searches? *J Med Internet Res* 14, 3 (13 Jun 2012), e87. https://doi.org/10.2196/jmir.2051

[16] Amira Ghenai, Mark D. Smucker, and Charles L.A. Clarke. 2020. A Think-Aloud Study to Understand Factors Affecting Online Health Search. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (Vancouver BC, Canada) *(CHIIR '20)*. Association for Computing Machinery, New York, NY, USA, 273–282. https://doi.org/10.1145/3343413.3377961

[17] Souvick Ghosh, Manasa Rath, and Chirag Shah. 2018. Searching as Learning: Exploring Search Behavior and Learning Outcomes in Learning-Related Tasks. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 22–31. https://doi.org/10.1145/3176349.3176386

[18] Lisa Neal Gualtieri. 2009. The Doctor as the Second Opinion and the Internet as the First. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI EA '09)*. Association for Computing Machinery, New York, NY, USA, 2489–2498. https://doi.org/10.1145/1520340.1520352

[19] Joseph F. Hair, Bill Black, Barry J. Babin, and Rolph E. Anderson. 2010. *Multivariate Data Analysis: Global Edition* (7 ed.). Pearson Education.

[20] Anat Hashavit, Hongning Wang, Raz Lin, Tamar Stern, and Sarit Kraus. 2021. Understanding and Mitigating Bias in Online Health Search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 265–274. https://doi.org/10.1145/3404835.3462930

[21] Bernard J. Jansen, Danielle Booth, and Brian Smith. 2009. Using the Taxonomy of Cognitive Learning to Model Online Searching. *Inf. Process. Manage.* 45, 6 (Nov. 2009), 643–663. https://doi.org/10.1016/j.ipm.2009.05.004

[22] Wonkwang Jo, Jaeho Lee, Junli Park, and Yeol Kim. 2020. Online Information Exchange and Anxiety Spread in the Early Stage of the Novel Coronavirus (COVID-19) Outbreak in South Korea: Structural Topic Model and Network Analysis. *J Med Internet Res* 22, 6 (2 Jun 2020), e19455. https://doi.org/10.2196/19455

[23] Allison B Kaufman and James C Kaufman. 2018. *Pseudoscience: The conspiracy against science.* Mit Press.

[24] 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 *Proceedings of the 2015 International Conference on The Theory of Information Retrieval* (Northampton, Massachusetts, USA) *(ICTIR '15)*. Association for Computing Machinery, New York, NY, USA, 101–110. https://doi.org/10.1145/2808194.2809465

[25] 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. https://doi.org/10.1197/jamia.M2449 arXiv:https://academic.oup.com/jamia/article-pdf/15/4/484/2183649/15-4-484.pdf

[26] Ashiqur R. KhudaBukhsh, Paul N. Bennett, and Ryen W. White. 2015. Building Effective Query Classifiers: A Case Study in Self-Harm Intent Detection. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (Melbourne, Australia) *(CIKM '15)*. Association for Computing Machinery, New York, NY, USA, 1735–1738. https://doi.org/10.1145/2806416.2806594

[27] Briege M. Lagan, Marlene Sinclair, and W. George Kernohan. 2010. Internet Use in Pregnancy Informs Women's Decision Making: A Web-Based Survey. *Birth* 37, 2 (2010), 106–115. https://doi.org/10.1111/j.1523-536X.2010.00390.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1523-536X.2010.00390.x

[28] Kenneth Lee, Kreshnik Hoti, Jeffery David Hughes, and Lynne M Emmerton. 2015. Consumer Use of "Dr Google": A Survey on Health Information-Seeking Behaviors and Navigational Needs. *J Med Internet Res* 17, 12 (29 Dec 2015), e288. https://doi.org/10.2196/jmir.4345

[29] Diane Levin-Zamir and Isabella Bertschi. 2018. Media Health Literacy, eHealth Literacy, and the Role of the Social Environment in Context. *International Journal of Environmental Research and Public Health* 15, 8 (2018). https://doi.org/10.3390/ijerph15081643

[30] Chang Liu, Ying-Hsang Liu, Tom Gedeon, Yu Zhao, Yiming Wei, and Fan Yang. 2019. The effects of perceived chronic pressure and time constraint on information search behaviors and experience. *Information Processing & Management* 56, 5 (2019), 1667–1679. https://doi.org/10.1016/j.ipm.2019.04.004

[31] Caio Machado, Beatriz Kira, Vidya Narayanan, Bence Kollanyi, and Philip Howard. 2019. A Study of Misinformation in WhatsApp Groups with a Focus on the Brazilian Presidential Elections.. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1013–1019. https://doi.org/10.1145/3308560.3316738

[32] Gary Marchionini. 2006. Exploratory Search: From Finding to Understanding. *Commun. ACM* 49, 4 (April 2006), 41–46. https://doi.org/10.1145/1121949.1121979

[33] Felipe Moraes, Sindunuraga Rikarno Putra, and Claudia Hauff. 2018. Contrasting Search as a Learning Activity with Instructor-Designed Learning. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Torino, Italy) *(CIKM '18)*. Association for Computing Machinery, New York, NY, USA, 167–176. https://doi.org/10.1145/3269206.3271676

[34] Raymond S. Nickerson. 1998. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* 2, 2 (1998), 175–220. https://doi.org/10.1037/1089-2680.2.2.175 arXiv:https://doi.org/10.1037/1089-2680.2.2.175

[35] Geraldine Peterson, Parisa Aslani, and Kylie A Williams. 2003. How do Consumers Search for and Appraise Information on Medicines on the Internet? A Qualitative Study Using Focus Groups. *J Med Internet Res* 5, 4 (19 Dec 2003), e33. https://doi.org/10.2196/jmir.5.4.e33

[36] Peter L. T. Pirolli. 2007. *Information Foraging Theory: Adaptive Interaction with Information* (1 ed.). Oxford University Press, Inc., USA.

[37] Lesley Rennis, Gloria McNamara, Erica J. Seidel, and Yuliya Shneyderman. 2013. Google It!: Urban Community College Students' Use of the Internet to Obtain Self-Care and Personal Health Information. *College student journal* 49 (2013), 414–426.

[38] Dietram A. Scheufele and Nicole M. Krause. 2019. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences* 116, 16 (2019), 7662–7669. https://doi.org/10.1073/pnas.1805871115 arXiv:https://www.pnas.org/content/116/16/7662.full.pdf

[39] 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. https://doi.org/10.1016/j.socscimed.2007.01.012

[40] Xiaoxuan Song, Chang Liu, and Yan Zhang. 2021. Chinese College Students' Source Selection and Use in Searching for Health-related Information Online. *Information Processing & Management* 58, 3 (2021), 102489. https://doi.org/10.

1016/j.ipm.2021.102489

[41] Saira Hanif Soroya, Ali Farooq, Khalid Mahmood, Jouni Isoaho, and Shan e Zara. 2021. From information seeking to information avoidance: Understanding the health information behavior during a global health crisis. *Information Processing & Management* 58, 2 (2021), 102440. https://doi.org/10.1016/j.ipm.2020.102440

[42] Brian G. Southwell, Jeff Niederdeppe, Joseph N. Cappella, Anna Gaysynsky, Dannielle E Kelley, April Oh, Emily B. Peterson, and Wen ying Sylvia Chou. 2019. Misinformation as a Misunderstood Challenge to Public Health. *American journal of preventive medicine* (2019).

[43] Briony Swire-Thompson and David Lazer. 2020. Public Health and Online Misinformation: Challenges and Recommendations. *Annual Review of Public Health* 41, 1 (2020), 433–451. https://doi.org/10.1146/annurev-publhealth-040119-094127 arXiv:https://doi.org/10.1146/annurev-publhealth-040119-094127 PMID: 31874069.

[44] Sharon Swee-Lin Tan and Nadee Goonawardene. 2017. Internet Health Information Seeking and the Patient-Physician Relationship: A Systematic Review. *J Med Internet Res* 19, 1 (19 Jan 2017), e9. https://doi.org/10.2196/jmir.5729

[45] Tabitha Tonsaker, Gillian Bartlett, and Cvetan Trpkov. 2014. Health information on the Internet: Gold mine or minefield? *Canadian family physician Médecin de famille canadien* 60 (05 2014), 407–8.

[46] Pertti Vakkari, Mikko Pennanen, and Sami Serola. 2003. Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information Processing & Management* 39, 3 (2003), 445–463. https://doi.org/10.1016/S0306-4573(02)00031-6

[47] Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. 2019. Systematic Literature Review on the Spread of Health-related Misinformation on Social Media. *Social Science & Medicine* 240 (2019), 112552. https://doi.org/10.1016/j.socscimed.2019.112552

[48] David Ka-Ki Wong and Man-Kuen Cheung. 2019. Online Health Information Seeking and eHealth Literacy Among Patients Attending a Primary Care Clinic in Hong Kong: A Cross-Sectional Survey. *J Med Internet Res* 21, 3 (27 Mar 2019), e10831. https://doi.org/10.2196/10831

[49] Dan Wu, Shaobo Liang, and Wenting Yu. 2018. Collaborative information searching as learning in academic group work. *Aslib J. Inf. Manag.* 70 (2018), 2–27.

[50] Wan-Ching Wu, Diane Kelly, Ashlee Edwards, and Jaime Arguello. 2012. Grannies, Tanning Beds, Tattoos and NASCAR: Evaluation of Search Tasks with Varying Levels of Cognitive Complexity. In *Proceedings of the 4th Information Interaction in Context Symposium* (Nijmegen, The Netherlands) *(IIIX '12)*. Association for Computing Machinery, New York, NY, USA, 254–257. https://doi.org/10.1145/2362724.2362768

[51] Pengyi Zhang and Dagobert Soergel. 2016. Process patterns and conceptual changes in knowledge representations during information seeking and sensemaking: A qualitative user study. *Journal of Information Science* 42, 1 (2016), 59–78. https://doi.org/10.1177/0165551515615834 arXiv:https://doi.org/10.1177/0165551515615834

[52] Yan Zhang. 2019. *Consumers' Selection and Use of Sources for Health Information.* 83–101. https://doi.org/10.1007/978-3-030-14714-3_5