The Positive and Negative Influence of Search Results on People's Decisions about the Efficacy of Medical Treatments

Frances A. Pogacar[1], 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
{fapogacar,aghenai,mark.smucker,charles.clarke}@uwaterloo.ca

ABSTRACT

People regularly use web search engines to investigate the efficacy of medical treatments. Search results can contain documents that present incorrect information that contradicts current established medical understanding on whether a treatment is helpful or not for a health issue. If people are influenced by the incorrect information found in search results, they can make harmful decisions about the appropriate treatment. To determine the extent to which people can be influenced by search engine results, we conducted a controlled laboratory study that biased search results towards correct or incorrect information for 10 different medical treatments. We found that search engine results can significantly influence people both positively and negatively. Importantly, study participants made more incorrect decisions when they interacted with search results biased towards incorrect information than when they had no interaction with search results at all. For search domains such as health information, search engine designers and researchers must recognize that not all non-relevant information is the same. Some non-relevant information is incorrect and potentially harmful when people use it to make decisions that may negatively impact their lives.

KEYWORDS

Health Search; User Study; Misinformation; Harmful Effects

1 INTRODUCTION

Wei Zexi, a 21 year old Chinese student, died on April 12, 2016, of synovial sarcoma, a form of cancer [10]. In the early stages of his illness, doctors treated him with conventional treatments. But when these treatments were not successful, his family reportedly spent 200,000 yuan (US$30,650) on an experimental treatment not approved for use in China. Wei Zexi's story is notable because he found the hospital offering the treatment via the Baidu search engine. The treatment did not help, and he later learned, via a friend using the Google search engine outside of China, that there was no scientific evidence that this treatment would help him. Shortly before his death, he wrote a web posting denouncing Baidu for violating his trust. Following the web post and his death, public outrage resulted in the Chinese government passing new regulations regarding search engines [1]. Apparently, Wei Zexi had found the treatment via an advertisement on Baidu's search results page, and among these new regulations was the requirement that search engines clearly identify advertisements as different from natural or organic search results [1].

When people search for health information online, as 72% of U.S. internet users do, the majority are seeking information about a health issue or medical treatment [7]. While the majority of U.S. internet users are confident searchers, and believe they are finding accurate information [12], it is likely that there are many like Wei Zexi who have used a search engine for health information and have ended up making incorrect decisions that either wasted their money or negatively impacted their health. Indeed, White and Hassan [16] have shown that search engines can be biased towards incorrectly indicating that medical treatments help when they do not, and that these errors may be amplified by people's bias towards positive information [14]. If people find and believe incorrect information regarding medical treatments, there is the potential for these people to be harmed.

To measure the actual effect of search bias on people's ability to correctly determine the efficacy of medical treatments, we conducted a controlled laboratory study with 60 participants. In our study, we biased search results towards being correct or towards being incorrect. We also controlled the topmost rank of a correct result to investigate the effect of rank.

Our study's participants had to determine the efficacy of ten different medical treatments. We asked participants to pretend that they had a question about the effectiveness of a medical treatment and that they had decided to use a search engine to help them answer the question. For each of the ten treatments, we either presented the participants with a search results page or a control condition where they had to directly answer the question without any search results at all.

We found that:

- Search results have a statistically significant, strong effect on people's ability to make correct decisions. Results biased towards incorrect information reduced people's accuracy from 43% to 23%. Results biased towards correct information increased accuracy from 43% to 65%.
- The topmost rank of a correct result appears to have some effect on people's accuracy. While not statistically significant, when shown results biased towards correct information, participants' accuracy was only 59% if the top two results were incorrect compared to 70% accuracy when the rank 1 item was correct.

ICTIR '17, October 1–4, 2017, Amsterdam, Netherlands.

ISBN 978-1-4503-4490-6/17/10…$15.00
DOI: https://doi.org/10.1145/3121050.3121074

- Knowledge of the medical treatment can perhaps inoculate people against incorrect information. We found more self-reported knowledge to reduce the effect of incorrect information on accuracy (p = 0.04).
- Like White and Hassan [16], we found that participants were biased towards saying treatments were helpful.

In addition, we collected information about search behaviour via a questionnaire and report on participant's confidence in their answers and their click behaviour.

Our results demonstrate that search engines have a great potential to both help and harm people. Indeed, when searchers decide that ineffective treatments will help them, they open themselves up to at best being swindled out of money and at worst being harmed by these ineffective treatments either directly or from lack of proper treatment.

We next review related work, and then cover the details of the study and present the study's results. Following the results, we discuss implications and conclude the paper.

2 RELATED WORK

Our work builds directly on the results of White and his co-authors [14–17]. White has established that web search engines have a bias towards search results that report that medical treatments help health issues even when the evidence is either inconclusive or actually says that the treatment is unhelpful. White's work has looked at both medical queries with yes and no answers [14, 15] and queries about the efficacy of medical treatments [16, 17]. An example of a yes/no question from [15] is "Does mono in children cause bruising?" An example of an efficacy query from [17] is "Does melatonin work for jet lag".

A key finding of this body of work is that people are both biased towards answers of "yes" and "helps" and that people's beliefs are difficult to change if they are already decided on an answer. White [15] did find that when search results are biased towards one answer (yes or no) and these answers are all ranked above the contradictory answers (all yes above all no, and vice versa), that people could be influenced to select the dominant answer in search results. When the correct answer to a question was *yes*, White was able to get participants to correctly answer 74.9% of the time. When the correct answer was *no*, users' accuracy could reach 63.1% when the results were biased to *no* and all *no* results were ranked above *yes* answers.

Our work specifically looks at searcher accuracy for determining the efficacy of medical treatments rather than yes/no questions. While White and Horvitz [17] looked at search accuracy for efficacy queries, they did not measure the impact of bias and rank on accuracy. White and Horvitz [17] examined organic search results, which have an uncontrolled bias, as well as controlled search results with a 50/50 mix of answers, i.e. unbiased search results. In this paper, we look at biasing results both towards correct and towards incorrect results. We also look specifically at the rank of the topmost correct document, which is a more subtle notion of rank than White [15] examined where he ranked all yes/no answers above all no/yes answers.

Both White [15] and White and Horvitz [17] focus their study on the process by which search engine results can change searcher beliefs. To study the dynamics of search beliefs, White and his co-authors first asked study participants about their beliefs before searching. In this paper, we purposely avoided asking study participants about their beliefs prior to searching for fear of biasing participants towards their pre-existing beliefs. If we were to ask participants for their prior beliefs, this would mean that to change their belief, a participant would need to admit to the experimenter that their prior belief was wrong. Instead, our control condition is to measure participant's accuracy without any exposure to search results. While we cannot measure how a single participant changes their belief, we can measure how a population's accuracy can be influenced. White [15] and White and Horvitz [17] both found that it is difficult to change beliefs, while we show that with a significant bias in results, large shifts in a population's accuracy can be achieved. As an additional aspect of our work, we examine the impact of participants' self-reported knowledge of the medical treatments and health issues.

Other than the work of White and co-authors, the work most relevant to our paper is that of Epstein and Robertson [6], who studied the impact of search results in the political domain. Epstein and Robertson designed large scale, controlled experiments to understand the influence of search engine results on political elections. Results showed that preferences of undecided voters can be significantly influenced and that the extent of the influence was associated with certain demographic characteristics. Epstein and Robertson's work is similar to ours as we both study the influence of search results on people's decisions, but while Epstein and Robertson focus heavily on the effect of rank on preferences in the political domain, we explore how search results, biased with correct and incorrect information, as well as rank, can lead participants towards or away from correct decisions in the health domain.

In other related work, Kammerer et al. [8] designed a controlled user study to understand the behaviour of people when evaluating web search sources regarding medical issues. Kammerer et al. [8] chose two medical treatments for a health issue, crafted search results using different types of sources (medical institutions, journals, forums, etc.), and then asked participants to evaluate which treatment was better. Using eye tracking, participant logs and verbal protocols, results showed that people spend less time and effort evaluating search results when information sources seem accurate and reliable. Even though their study design is similar to ours, their main focus was to evaluate the validation of sources. In this paper, we evaluate the influence of search bias and rank when searching the efficacy of medical treatments.

Kulshrestha et al. [9] studied search bias in Twitter, a social media website. More specifically, Kulshrestha et al. [9] introduced three different aspects of bias for search systems: query bias, output bias, and ranking bias. Results showed that query bias (such as query topic or how the query is phrased) and ranking bias play an important role in producing bias in search results.

3 METHODS AND MATERIALS

To measure the effect of search results bias on people's ability to correctly determine the efficacy of medical treatments, we created a controlled, within-subjects, laboratory study. We first provide an overview of the experiment and then detail each of the parts.

3.1 Overview

Our experiment had two independent variables each with two levels. The first independent variable was the search results bias with the levels: *correct* and *incorrect*. The second independent variable was the rank of the topmost correct search result with levels of 1 and 3, indicating the position of the first correct result. The experiment also had a control condition, in which no search results are presented to the user. The two independent variables with two levels each, plus a control, produces five experimental conditions. Participants had to determine the efficacy of medical treatments and a treatment could be either *helpful* or *unhelpful*. So that each of the five experimental conditions would be measured on both *helpful* and *unhelpful* treatments, we selected five of each for a total of ten treatments. The experiment had two dependent variables: 1) the fraction of *correct* decisions and 2) the fraction of *harmful* decisions made by the participant. In addition, we collected data from a questionnaire and feedback on each decision made. We also logged computer interactions for the entire study.

3.2 Medical Treatments

To select our medical treatments, we first received from White and Hassan [16] a list of 249 treatments that they had judged for their study. White and Hassan together determined the effectiveness of each treatment by reading the corresponding Cochrane Review [4, 5] and then reaching a consensus to determine the treatment's efficacy. A Cochrane Review is a systematic review that synthesizes the clinical evidence and informs clinical decision making. White and Hassan settled on three categories of efficacy: *helps*, *inconclusive*, and *does not help*.

For each medical treatment in our study, our participants needed to decide on its efficacy by selecting one of these three categories. We described the categories to our study participants as follows:

- **Helps**: The 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 specified illness.
- **Does not help**: The 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.

To help our study participants better understand each category, these definitions modify and expand upon the definitions that White and Hassan [16] used in their paper. To save space in this paper, we report results using the labels: *helpful*, *inconclusive*, and *unhelpful*.

Table 3 shows the ten treatments we selected for our study. Each medical treatment is associated with a stated health issue. We selected five *helpful* and five *unhelpful* treatments, and we tried to select treatments and health issues that might be of interest to university students, who would form the majority of our study participants.

3.3 Control Condition

The control condition required participants to decide on the efficacy of a medical treatment without any assistance, i.e., they were not shown a search engine results page (SERP). This control condition allows us to determine the fraction of correct and harmful decisions that participants would make if they did not interact with a search engine. Participants experienced the control condition for two of the ten medical treatments that they judged.

3.4 Search Results - Independent Variables

For eight of the ten medical treatments, we instructed participants to pretend that they had a question about the effectiveness of a medical treatment and had decided to use a search engine to help them answer the question. In these cases, we showed participants a web page that looked like a search engine results page (SERP) with ten search results displayed with snippets.

All ten of the search results were about the medical treatment, but they were biased towards either *correct* or *incorrect* information regarding the efficacy of the medical treatment.

To bias our search results towards correct information, we selected eight of the results to be correct and two to be incorrect. A correct result is a document that contains information about the efficacy of the medical treatment that supports the truth, and an incorrect result contains information that contradicts the true efficacy of the medical treatment. To bias the search results towards incorrect information, we selected eight to be incorrect and two to be correct.

Our amount of bias is similar to that which can be found in actual search engines. White and Hassan [16] found that, for a major web search engine, at rank 10, on average, 80.69% of the results for a query about a medical treatment reported that the treatment was *helpful*. The remainder of the top 10 results consisted of 12.29% being *inconclusive* and 7.01% being *unhelpful*.

In addition to controlling the result bias to be correct or incorrect, we also controlled the rank of the topmost correct document to be either at rank 1 or at rank 3. We selected these ranks because eye tracking studies show that the first two results are viewed at very high rates, but that attention from rank 1 to rank 3 drops by about 50% [11].

For each participant and each display of search results, we used randomization to generate the search results. For each medical treatment, we had pools of 8-10 correct and 8-10 incorrect documents. To generate search results biased towards *correct* information, we randomly selected two incorrect documents and eight correct documents from their respective pools. Conversely, for results biased towards *incorrect* information, we randomly selected two correct and eight incorrect documents from their respective pools. The topmost correct document was randomly assigned into rank 1 or rank 3, corresponding to the experimental condition. If the experimental condition had the topmost correct document in rank 3, then rank 1 and rank 2 were assigned two random incorrect documents. The rest of the incorrect and correct documents were then randomly distributed across the remaining ranks. After generating search results pages, we verified that the correct and incorrect documents were randomly distributed among the ranks and across the participants.

3.5 Documents and Snippets

In order to build search engine result pages for every medical treatment, we collected documents containing information about the treatment's efficacy. We used Bing, Yahoo, and Google to collect a total of 158 documents relevant to determining the efficacy of the ten medical treatments. As described in the previous section, for each medical treatment we created pools of 8-10 correct and 8-10 incorrect documents. A correct document contains information about the efficacy of the medical treatment that supports the truth (see Table 3). An incorrect document contains information that contradicts the true efficacy of the medical treatment.

We divided the task of collecting and labeling documents as either correct or incorrect between two of the paper's authors. For some of the medical treatments, it was difficult to find eight documents stating that the medical treatment was unhelpful. In these cases, we selected documents that did not directly support or oppose the truth, but rather listed negative side effects or possible harm of the treatment.

For the search results pages, we showed the document's title, its url, and a snippet. We manually constructed the snippets. For topics T1-T8 (Table 3), one of the authors selected the first two sentences of the document as the snippet. For topics T9 and T10, a different author selected what appeared to be the most important and descriptive sentences. We did not realize that different techniques were employed until after the experiment was concluded. Given that we did not see significantly different click behavior across the different medical treatments, we do not believe that the different selection of snippets affects the results. As part of publication, we intend to release copies of these documents and the snippets for others to be able to replicate the experiment.

3.6 Dependent Variables

We study two dependent variables. The first is the fraction of decisions that are correct. A participant's decision about the efficacy of a medical treatment is correct if their decision matches the truth (Table 3). Note that if a participant decides that the efficacy of a medical treatment is inconclusive, that decision will always be wrong because our ten medical treatments are either helpful or unhelpful.

Our second dependent variable is the fraction of decisions that are harmful. We consider a harmful decision to be one where the participant decides that the efficacy is the opposite of the truth, i.e., the participant decides a medical treatment is helpful when in fact it is unhelpful, or unhelpful when it is helpful. If a participant decides that a medical treatment's efficacy is inconclusive, we do not count that as a harmful decision because our reasoning is that the participant will still need to find more information before making a final decision.

3.7 Study Design

After consenting to participate, participants filled out a questionnaire to capture demographic information as well as information about their usage of search engines for health related purposes. Following the questionnaire, the participants read instructions and had to answer correctly a set of questions regarding the study before they could proceed with the study. Next the participants had a chance to practice with the system by determining the efficacy of two medical treatments not used in the main study. For one medical treatment, they could use search results and for the other they experienced the control condition, with no search results. The participants then began the main study where they had to decide on the efficacy of the ten medical treatments while experiencing the experimental conditions. For each medical treatment decision task, we asked pre-task and post-task questions. Before the task, we asked participants about their knowledge of the health issue and treatment. After the task we asked the participants about their confidence in their answer. At the end of the study, participants were debriefed and provided with the truth about each of the medical treatments.

3.7.1 User Interface. We built the study as a web application. For 8 of the 10 medical treatments, participants interacted with a search engine results page. For the other two medical treatments, the participants received the control condition, with no search results. We modelled the search results page after the traditional style of web search engines. At the top of the page, we displayed the medical treatment question that the user is asked to answer followed by a short boxed paragraph showing definitions of the health issue and treatment. We obtained the definitions from either Merriam-Webster's[1] or the Mayo Clinic's[2] medical dictionaries. We showed the definitions to avoid confusion and to make sure participants had a basic understanding of what was meant by the health issue and medical treatment. The medical treatment question and definitions remained visible throughout the entire task.

The search results page allowed participants to click on the search results, but they could not issue additional queries or obtain additional results. On the right side of the search results page, we displayed a reminder of the definitions of the different categories of medical treatment efficacies: *helps*, *does not help* and *inconclusive*.

For every document summary, we first showed the document title followed by a snippet and a link to the actual page. When a participant clicked on a search result, we took them to a screenshot of the web page rather than to the actual web page. We did this because we wanted to make sure that the participant was not able to click on any links and view any pages outside the scope of the study. In addition, this approach allowed us to be certain that each participant was exposed to the same version of the web page, and we did not have to fear the loss of pages during the study. We placed a button at the bottom of the search results page that, when pressed, took the user to a page to submit their decision regarding the efficacy of the medical treatment.

3.7.2 Balanced Design. We used a 10x10 Graeco-Latin square to create a fully balanced design and randomize medical treatments and experimental conditions. We create each 10x10 block by first creating four smaller 5x5 squares as follows. To balance the *helpful* medical treatments with the *unhelpful* ones, we generated three Latin Squares: one for the five experimental conditions, one for the five helpful medical treatments and one for the five unhelpful medical treatments. Overlaying the Latin square of experimental conditions over the helpful treatments and over the unhelpful treatments individually, we create two separate Graeco-Latin squares ensuring that both the *helpful* treatments and *unhelpful* treatments

[1] https://www.merriam-webster.com/

[2] http://www.mayoclinic.org/

have an equal and systematic balance of the experimental conditions. Finally by randomizing the columns and rows, this process creates two separate 5x5 Graeco-Latin Squares - one for the *helpful* treatments and one for the *unhelpful* treatments. Repeating the above process generates two new Graeco-Latin squares for *helpful* and *unhelpful* treatments and gives us four separate Graeco-Latin Squares (two *helpful* and two *unhelpful*). Combining the four squares we generate a 10x10 Graeco Latin square, and randomize the columns and rows and then randomly assign participants to each row.

3.8 Participants

We obtained ethics approval from our university and then recruited participants via posters and email announcements to different graduate student email lists at the university. 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 60 students (27 male, 33 female) from different majors (36 from engineering and mathematics, 20 from arts and sciences and 4 from other majors) with an age between 18 and 36 years old (22% less than 20, 50% between 20 and 25 and 28% greater than 25, with an average age of 23).

3.9 Data Cleaning

During the course of the study, four participants had to be replaced because of failure to successfully complete the study due to technical or other issues. After a careful examination of the study data from the 60 participants, we did not find any irregularities and thus did not clean or modify the data before analysis.

3.10 Statistical Significance and Modelling

To determine the statistical significance of our results, we used generalized linear (logistic) mixed effect models as implemented in R [13] and the lme4[2] package. We used logistic regression because our dependent variables of correct and harmful decisions are binary outcomes. We modeled participants and medical treatments as random effects. Our independent and explanatory variables were fixed effects. We test the effect of each independent and explanatory variable on our dependent variables individually. To analyze the significance of these variables, for each variable we build and compare two models using a likelihood ratio test that reports a Chi-Square test statistic and p-value. The first model is the complete model. The complete model includes the dependent variable, the applicable independent variables, and the random effects. The second model is the null model, which includes everything in first model minus the variable of interest. With the two models, the null model, without the variable of interest, and the complete model, with the variable of interest, we perform the likelihood ratio test. The p-values are then determined by chi-square tests on the log-likelihood values.

When analyzing our entire dataset, which includes all five experimental conditions (a 2x2 factorial design plus one control), we do not include the Topmost Correct Rank as a fixed effect in the model. This is because the control condition has no search results, and therefore rank is not applicable. The majority of our other analyses are done on the four search results experimental conditions without the control. For these analyses, we include both independent variables of Search Results Bias and Topmost Correct Rank in our models.

4 RESULTS AND DISCUSSION

The main results of our study focus on the effect of our independent variables on the participants' ability to correctly determine the efficacy of the ten medical treatments. The participants either interact with controlled search results to help them answer the question, or they are asked to directly answer the question without any search results (control condition). The search results are either biased towards correct or incorrect information regarding the medical treatment, with the topmost correct document at rank 1 or rank 3.

Table 1 reports the fraction of correct and harmful decisions of the 60 participants corresponding to the independent variables of Search Results Bias and Topmost Correct Rank. Refer to Section 3.6 for the definitions of correct and harmful decisions. We see that results with the rank 1 document correct and biased towards correct information can lead to increased accuracy up to 70%, while lowering harmful decisions from 20% to 6%. Conversely, results biased towards incorrect information significantly reduces accuracy from 43% to 23%, while doubling the incidence of harmful decisions.

Table 2 reports the statistical significance of the independent variables on the dependent variables from Table 1. Measuring significance using the techniques described in Section 3.10, we found the effect of the search result bias is statistically significant on the fraction of correct decisions and harmful decisions. We found that the topmost correct rank had less of an effect on the dependent variables, yet it did demonstrate some explanatory significance for our model with a nearly statistically significant effect (p = 0.06) on the fraction of harmful decisions made by the participant.

These results demonstrate the strong effect that search results can have on people's ability to use search results to correctly determine the efficacy of medical treatments. We have shown that with exposure to correct information, searchers perform better. On the other hand, we see that there is harm that can be done by incorrect information. For our experimental conditions where the search bias is towards incorrect information, the results show that participants actually perform worse than if they had no search results at all. Although the bias is towards incorrect information, the search results still contain two correct documents, with one always located in the top three ranks of the result list. The possibility to find the correct information is there, yet participants perform worse than if they were given no extra information.

Table 3 shows the fraction of correct decisions made by the participants for each of the ten medical treatments. For nine of the ten treatments, we see that search results biased towards incorrect information, decreases the accuracy with respect to the control. The treatment that does not behave as expected is T6 *Does caffeine help asthma?* (truth = helpful). For this specific treatment, the search results biased towards incorrect information improves performance over the control. The control shows that most participants generally begin with an incorrect belief. We may speculate that when exposed to the search results, participants may find the correct documents

Independent Variables		Dependent Variables	
Results Bias	Topmost Correct Rank	Fraction of Decisions	
		Correct	Harmful
Incorrect	3	0.23 ± 0.04	0.41 ± 0.05
Incorrect	1	0.23 ± 0.04	0.35 ± 0.04
Control (No search results)		0.43 ± 0.05	0.20 ± 0.04
Correct	3	0.59 ± 0.05	0.13 ± 0.03
Correct	1	0.70 ± 0.04	0.06 ± 0.02

Table 1: Main results. Users either interact with a page of search results or had to determine the efficacy of the medical treatment with no search results (control condition). We biased the search results towards incorrect or correct answers. We also controlled the topmost correct result to be at either rank 1 or 3. Based on the decisions the 60 participants made, we compute the fraction of correct and harmful decisions. Fractions are shown along with their standard errors. Table 2 reports the statistical significance of the independent variables.

Independent Variable	Dependent Variable	Pr(>Chisq)
Search Results Bias	Correct Decision	$\ll 0.001$
Search Results Bias	Harmful Decision	$\ll 0.001$
Topmost Correct Rank	Correct Decision	0.16
Topmost Correct Rank	Harmful Decision	0.06

Table 2: Statistical significance of independent variables. When the dependent variable is either the participant making a correct or a harmful decision, the search bias is statistically significant for the outcomes in Table 1. The rank of the topmost correct result shows significance near the 0.05 level with a p-value of 0.06 when the dependent variable is whether or not the participant makes a harmful decision.

and this slightly improves their performance. For eight of the ten treatments, we see that search results biased towards correct information, increases the accuracy with respect to the control. The two cases which do not behave as expected are T7 *Does cinnamon help diabetes?* (truth = unhelpful) and T9 *Does surgery help obesity?* (truth = helpful). For T9, the accuracy decreases slightly under the *correct* search results and may be due to random noise. On the other hand, T7 creates some speculation for what is actually going on for that specific treatment. A follow up study, including observations and debriefing participants would help better analyze these trends.

As White and Hassan [16] have demonstrated, participants and search engines have strong biases towards positive information. We split our data by the medical treatment type of *helpful, inconclusive, unhelpful* to investigate the trends and behaviours of our participants. Table 4 shows this data separately for the control condition and the other experimental conditions. Both tables show that there is an overall bias towards deciding that a health treatment is *helpful*. For the control condition, where the medical treatment is truly unhelpful, results show that participants correctly answer *unhelpful* about as often as they answer *inconclusive*. For the controlled experimental conditions, where the medical treatment is truly unhelpful, results show that participants are actually more likely to answer *inconclusive* than to decide that the treatment is *unhelpful*. This suggests that users are looking for information that is positive, and would rather respond *inconclusive* than believe that a treatment is *unhelpful*. This is a dangerous bias. When a treatment is truly unhelpful, searchers want to find positive information, and therefore can be heavily influenced by search results with incorrect information, claiming that the treatment is *helpful*.

4.1 Knowledge and Confidence

Before participants saw any search results for a given medical treatment, we asked them separately about their knowledge of the health issue and the medical treatment. Participants answered the questions on a rating scale, which we coded from 1 to 5 to mean "nothing", "heard of it", "know generally about it", "quite familiar", and "know extensive details". Knowledge of the health issue and medical treatment were positively correlated as determined by the Pearson correlation coefficient of $r = 0.40$ ($p \ll 0.001$). After submitting their decision about the medical treatment, we asked participants to report their confidence in their answer on a 5 point scale from 1="very uncertain" to 5="very certain".

We did not find that knowledge had a statistically significant effect on our dependent variables, but we did see a general trend for more knowledge to result in a greater fraction of correct decisions when the search results were biased towards incorrect information. Looking closer, we decided to group decisions made with the two highest levels of knowledge into one group, *high* and the lowest three levels of knowledge into another group, *low*.

Considering only the experimental conditions when the search results are biased towards incorrect information, we can examine the fraction of correct decisions made plus and minus its standard error for both low and high knowledge levels of both health issue and medical treatment. The fraction of correct decisions and its standard error for *low health issue* knowledge was 0.19±0.03. When the knowledge of the health issue is *high*, the fraction increases to 0.28 ± 0.04. Applying a Chi-squared test, the difference between these two rates is not statistically significant (p=0.14).

When the knowledge of the *medical treatment* is low, the fraction correct is 0.20 ± 0.03 and it increases to 0.34 ± 0.06 when knowledge is *high*. The difference between these rates is statistically significant (p=0.04). Knowledge of medical treatment can result in a significantly higher fraction of decisions made correctly when exposed to search results biased towards incorrect information. Table 3 shows that under the control condition, the fraction of decisions made correctly was 0.43 ± 0.05. Applying a two-sided t-test to compare the control condition to the decision made with high knowledge of the medical treatment, we fail to reject the null that they are the same rates (p=0.21). Even so, having knowledge of the health issue and medical treatment are not enough to raise performance above no exposure to search results. In other words, it is not as though the knowledgeable participants could fully ignore the incorrect information and only focus on the correct information and exceed the control condition's performance.

If we perform these same analyses for the fraction of decisions that are harmful, we find that more knowledge is associated with

			Fraction of Decisions Correct		
			Control (no	Search Results Bias	
T	Medical Treatment (Cochrane ID Suffix)	Efficacy	search results)	Incorrect	Correct
T1	Do antioxidants help female subfertility? (7807.pub2)	Unhelpful	0.58 ± 0.15	0.08 ± 0.06	0.71 ± 0.09
T2	Do benzodiazepines help alcohol withdrawal? (5063.pub3)	Helpful	0.33 ± 0.14	0.29 ± 0.09	0.63 ± 0.10
T3	Do insoles help back pain? (5275.pub2)	Unhelpful	0.33 ± 0.14	0.17 ± 0.08	0.50 ± 0.10
T4	Do probiotics help treat eczema? (6135.pub2)	Unhelpful	0.33 ± 0.14	0.17 ± 0.08	0.75 ± 0.09
T5	Do sealants prevent dental decay in the permanent teeth? (1830.pub4)	Helpful	0.67 ± 0.14	0.46 ± 0.10	0.83 ± 0.08
T6	Does caffeine help asthma? (1112.pub2)	Helpful	0.08 ± 0.08	0.25 ± 0.09	0.79 ± 0.08
T7	Does cinnamon help diabetes? (7170.pub2)	Unhelpful	0.50 ± 0.15	0.00 ± 0.00	0.38 ± 0.10
T8	Does melatonin help treat and prevent jet lag? (1520)	Helpful	0.67 ± 0.14	0.38 ± 0.10	0.79 ± 0.08
T9	Does surgery help obesity? (3641.pub3)	Helpful	0.67 ± 0.14	0.46 ± 0.10	0.63 ± 0.10
T10	Does traction help low back pain? (3010.pub5)	Unhelpful	0.17 ± 0.11	0.08 ± 0.06	0.46 ± 0.10
	Overall		0.43 ± 0.05	0.23 ± 0.03	0.65 ± 0.03

Table 3: This table shows the medical treatments with their corresponding efficacy and suffix to their Cochrane [5] source ID. The Cochrane ID has been condensed from the full ID. The prefix for each suffix listed in the table is 14651858.CD00*. Each treatment is also assigned a label T1 - T10 that we use throughout the paper to refer to specific medical treatments. The table also shows the fraction of decisions correctly made by participants for each of the 10 medical treatments under the control condition, and the experimental conditions of search results biased toward incorrect and correct information. Fractions are shown along with their standard errors.

Control Condition (No Search Results)

	Participant Decision			
Truth	Unhelpful	Helpful	Inconclusive	Total
Unhelpful	23	16	21	60
Helpful	8	29	23	60
Total	31	45	44	120

Experimental Conditions (Interact with Search Results)

	Participant Decision			
Truth	Unhelpful	Helpful	Inconclusive	Total
Unhelpful	79	64	97	240
Helpful	50	132	58	240
Total	129	196	155	480

Table 4: Confusion matrices. These tables show the decisions made by the study participants regarding the efficacy of the 5 helpful and 5 unhelpful medical treatments. The upper table shows the decisions under the control condition when participants decide without any assistance at all. The lower table shows the decisions made under the experimental conditions that allow the participants to interact with controlled search results.

more harmful decisions. On investigation, we found that this is because people who are less knowledgeable are more likely to decide a medical treatment is *inconclusive*, which highlights a limitation of analyzing results in terms of the fraction of harmful decisions.

For the confidence of the decision, we found that participants who decide that a medical treatment's efficacy is *inconclusive*, are less confident in their answer than those deciding a treatment is *unhelpful* or *helpful*.

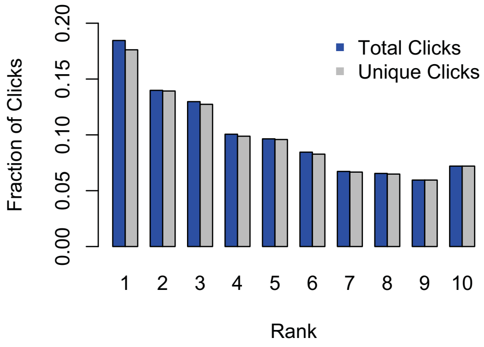

Figure 1: This graph shows the fraction of total clicks and unique clicks for each of the 10 search result ranks.

4.2 Clicks

Figure 1 shows the distribution of clicks over the search result ranks. We see that the total number of clicks and unique number of clicks per question and overall are very similar. The biggest difference between the total clicks and unique clicks occurs at rank 1, which shows that rank 1 is so important that some participants click on it multiple times in the same session. The overall distribution of clicks over search rank shows a similar result to what is seen with real search engines and provides some evidence that our participants interacted with our search results in a realistic fashion.

Over the four SERP experimental conditions, the average number of total clicks per question was 3.50 ± 0.1. For all correct decisions of SERP experimental conditions, the average number of total clicks was 3.73 ± 0.2. For all incorrect decisions of SERP experimental conditions, the average number of total clicks was 3.32 ± 0.2. For all harmful decisions of SERP experimental conditions (response was opposite to the correct answer), the average number of clicks was

Dependent Variables	Average Number of Clicks
Harmed Decisions	3.02 ± 0.3
Unharmed Decisions	3.65 ± 0.3
Correct Decisions	3.73 ± 0.2
Incorrect Decisions	3.32 ± 0.2

Table 5: Average Number of Clicks for each dependent variable: Correct Decisions, Harmful Decisions. This analysis only applies to the 4 SERP experimental conditions. Control data (No SERP) is not relevant.

3.02±0.3. Conversely, all unharmful decisions of SERP experimental conditions (their response was correct), the average number of clicks was 3.65 ± 0.3. The difference between the mean number of clicks for correct and harmful decisions is statistically significant. Participants that interact more with the search results are more likely to make a correct decision and may be working harder to determine the correct answer.

5 CONCLUSION

When people use search engines to answer health questions, their interaction with the system has the potential for both positive and negative outcomes. When people find medical treatments or information that will prolong or improve their life, or that of a loved-one, search engines demonstrate an ability to make people's lives better. When search engines intermix correct and incorrect information, we have shown that there is the potential for harm.

In this paper, we showed that search results can significantly affect people's decisions about the efficacy of medical treatments. Compared to not using a search engine, when people interacted with search results biased toward incorrect information, their accuracy dropped from 43% to 23%. Thankfully, when people interact with search results biased towards correct information, their accuracy climbed to 65% (Table 3).

There has long been people who prey on the hopes of others for cures to terrible diseases, and now their webpages can become intermingled with those of reputable medical organizations. For example, a search for Hoxsey Therapy, an ineffective cancer treatment [3], on today's popular web search engines, returns a mix of results that either explain it is ineffective or explain how it can help a patient with cancer. We found that people are biased towards wanting treatments to be helpful, and this bias combined with incorrect information has the potential to cause people harm.

The implications of these results extend beyond health search. Information retrieval researchers typically use curated collections. These curated collections contain high quality and trustworthy documents. On the open web, we already know that there is spam, and we actively filter it out of web results. We now can see that web search needs more than spam filtering. Web search also needs a form of automated curation to be available to searchers so that they can have confidence in the quality of the information being provided to them. It is not enough to rely on searchers' own media literacy to protect them from incorrect information.

Likewise, information retrieval evaluation needs to expand its understanding of the effects of documents beyond graded relevance. Non-relevant does not always mean innocuous. A document that leads a searcher to form a harmful belief about a medical treatment is damaging. A non-relevant document in today's effectiveness measures only causes a loss of time or effort and is represented as having zero gain. An incorrect document can increase the likelihood of a searcher forming a harmful belief and undoing the value of relevant documents, i.e. an incorrect document could be perceived to have some notion of a *negative gain*, which to our knowledge, is a new concept in information retrieval.

ACKNOWLEDGMENTS

We thank Ryen White for his assistance, feedback, and sharing of resources. Nimesh Ghelani and Gaurav Baruah both assisted with the running of the study. We thank the University of Waterloo Survey Research Centre for their advice. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (Grants CRDPJ 468812-14 and RGPIN-2014-03642), in part by Google, in part by the Qatar Foundation for Education, Science and Community Development, and in part by the University of Waterloo.

REFERENCES

[1] Alyssa Abkowitz. 2016. China Issues New Internet Search Rules Following Baidu Probe; Regulator mandates 'objective, fair and authoritative results'. *Wall Street Journal (Online)* (Jun 26 2016).

[2] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. DOI: http://dx.doi.org/10.18637/jss.v067.i01

[3] Barrie R. Cassileth and Helene Brown. 1988. Unorthodox cancer medicine. *CA: A Cancer Journal for Clinicians* 38, 3 (1988), 176–186.

[4] A Cipriani, TA Furukawa, and C Barbui. 2011. What is a Cochrane review? *Epidemiology and psychiatric sciences* 20, 03 (2011), 231–233.

[5] J. P. T. Higgins (Ed.). 2008. *Cochrane Handbook for Systematic Reviews of Interventions*. Vol. 5. The Cochrane Collaboration. www.cochrane-handbook.org.

[6] 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.

[7] Susannah Fox and Maeve Duggan. 2013. Health Online 2013. Pew Research Center. (2013).

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

[9] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, IIEST Shibpur, India Krishna P Gummadi, and Karrie Karahalios. 2017. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. In *Proc. of CSCW*.

[10] Yadan Ouyang. 2016. Student's death highlights gaps in China's health regulations. *Lancet Oncology* 17, 6 (2016), 709.

[11] Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. 2007. In Google We Trust: Users' Decisions on Rank, Position, and Relevance. *Journal of Computer-Mediated Communication* 12, 3 (2007), 801–823.

[12] Kristen Purcell, Joanna Brenner, and Lee Rainie. 2012. Search Engine Use 2012. Pew Research Center. (2012).

[13] R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

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

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

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

[17] Ryen W White and Eric Horvitz. 2015. Belief dynamics and biases in web search. *ACM Transactions on Information Systems (TOIS)* 33, 4 (2015), 18:1–18:46.