Content Bias in Online Health Search

RYEN W. WHITE and AHMED HASSAN, Microsoft Research

Search engines help people answer consequential questions. Biases in retrieved and indexed content (e.g., skew toward erroneous outcomes that represent deviations from reality), coupled with searchers' biases in how they examine and interpret search results, can lead people to incorrect answers. In this article, we seek to better understand biases in search and retrieval, and in particular those affecting the accuracy of content in search results, including the search engine index, features used for ranking, and the formulation of search queries. Focusing on the important domain of online health search, this research broadens previous work on biases in search to examine the role of search systems in contributing to biases. To assess bias, we focus on questions about medical interventions and employ reliable ground truth data from authoritative medical sources. In the course of our study, we utilize large-scale log analysis using data from a popular Web search engine, deep probes of result lists on that search engine, and crowdsourced human judgments of search-result captions and landing pages. Our findings reveal bias in results, amplifying searchers' existing biases that appear evident in their search activity. We also highlight significant bias in indexed content and show that specific ranking signals and specific query terms support bias. Both of these can degrade result accuracy and increase skewness in search results. Our analysis has implications for bias mitigation strategies in online search systems, and we offer recommendations for search providers based on our findings.

Categories and Subject Descriptors: **H.3.3 [Information Storage and Retrieval]**: Information Search and Retrieval – *Search process, Selection process*.

General Terms: Experimentation, Human Factors, Measurement.

Additional Key Words and Phrases: Content biases; Health search.

1. INTRODUCTION

Health search is prevalent on the Web. Over 80% of U.S. adult Internet users have performed online health searches [Fox and Duggan 2013]. By nature, health queries may have significant consequences, but one scenario of particular importance to health information seekers is determining the efficacy of medical interventions, e.g., in queries such as [can green tea help with weight loss]. Some medical interventions may be effective, but some may not and may even have harmful effects (e.g., Politi et al. [2007] reported that 7% of all medical treatments involve uncertain tradeoffs between benefit and harm). In making such determinations, searchers gather evidence from search results, meaning that they are susceptible to their own cognitive biases [White and Horvitz 2009a; White 2013] but also biases in the information surfaced by search engine ranking algorithms. These algorithms can exhibit skew in the information that they cover [Gerhart 2004; Vaughan and Thelwall 2004; Goldman 2006] and in how they decide to order their search results [Mowshowitz and Kawaguchi 2002a; 2002b]. In addition, search systems have limited access to information about a searcher's current situation and background probabilities on health conditions both for them and more generally, patient populations. The trust that people are known to place in result

Author's addresses: R. W. White, Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA; A. Hassan, Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA.

DOI:http://dx.doi.org/10.1145/0000000.0000000

rankings can lead to heightened concerns, erroneous beliefs, and negative emotional outcomes [White and Horvitz 2009a; Lauckner and Hsieh 2013].

The availability heuristic is a cognitive bias leading people to overestimate the likelihood of events based on the ease of retrieval and recall [Tversky and Kahneman 1973; Simon 1991]. Searchers may be susceptible to related biases in indexed and retrieved content that favors particular outcomes, often misaligned with reality [White and Horvitz 2009a]. Such biases in information availability can affect people's beliefs and decision making processes [Simon 1991].

Recent work on confirmation biases in information retrieval has focused on their impact on search-result examination behavior [White 2013]. This study is different in that we target biases in search engine functionality, in terms of levels of skew and accuracy in indexed and retrieved content, and the extent to which this deviates from known distributions. We refer to these as *content biases* and define them as follows:

DEFINITION: *Given a query, content bias in its search results describes a deviation from a known or accepted truth that negatively affects result accuracy*.

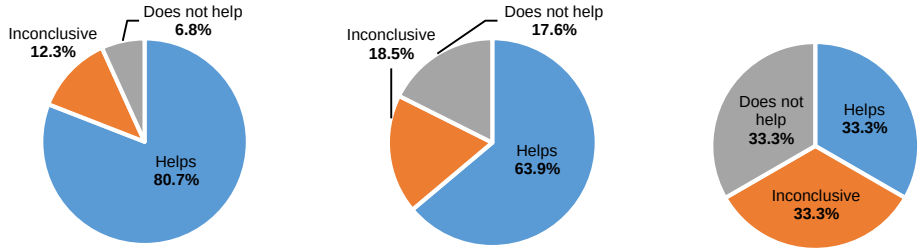
Content biases can stem from three sources: (i) creators of the online content (i.e., more could be written about particular outcomes irrespective of the truth); (ii) search engines, which could index and retrieve a skewed set of the available content, and; (iii) searchers, who are affected by their own biases in terms of how they seek, perceive, and react to the information presented to them. In the context of health seeking, addressing content bias is important for the sizable subset of the general public who rely on search engines for healthcare decision making [Fox and Duggan 2013].

The authoritativeness or reliability of results has been considered during crawling [Tang et al. 2005] and ranking [Brin and Page 1998], as well as to inform better result selection decisions [Schwarz and Morris 2011]. Biases in search engine result lists, toward particular brands [Mowshowitz and Kawaguchi 2002a], websites [Fortunato et al. 2006], or perspectives [Pariser 2011; Kacimi and Gamper 2011; 2013], has also been studied. There are also legal implications of bias in areas such as e-Government, where recent legislation requires that information be available to an informed citizenry.[1] Measures of bias such as *retrievability* [Azzopardi and Vinay 2008] and *discoverability* [Dasgupta et al. 2007] have been proposed to detect biases in ranked and indexed content respectively. However, to our knowledge none of these studies or methods focus directly on deviations from known or accepted distributions as we do in this research.

In this article, we seek to better understand biases in search engine operation, specifically the indexed and returned content, associated ranking signals (including implicit feedback derived from aggregated search behavior), and the impact of query formulation. To meet these objectives, we employ search engine log analysis, result scraping, and crowdsourced judging. Before diving deeply into the nature of any content bias within search engines, we sought an initial sense for the extent of any content bias in the search results for medical intervention queries. These are an important class of query where people seek to assess the efficacy of medical treatment options by examining search results.

To quantify bias in this setting, we obtained a set of intervention queries derived via matching log entries from the Microsoft Bing Web search engine against evidence-based expert reviews from the Cochrane Collaboration (cochrane.org). We retrieved the results for these queries from the Bing search engine, obtained answer labels for the retrieved results from third-party judges (i.e., the Web page suggests that the

[1] U.S. Legislation: E-Government Act 2002, and the e-Government Reauthorization Act 2007.



(a) Top 10 search results (b) Search engine index (c) Expected distribution

Figure 1. Distribution of answers about interventions in (a) top results, (b) matching index content, and (c) the expected (true) distribution given our sampling criteria (33% per answer).

intervention *helps*, *does not help*, or is *inconclusive*), and computed the distribution of answers. We compared the answer distribution in the top 10 search results against the distribution of matching content in the search engine index (see Figure 1, overleaf). While the expected (true) distribution per our uniform sampling strategy (described later in the article) is 33% per answer, on average 64% of available results (i.e., in the index and match the query in some way) are positive (*helps*), and on average 81% of the top 10 results returned by the search engine are positive. If the search results and indexed content were unbiased we would expect their distributions to match (or at least be similar) to the equal distribution of answers in our ground truth. From Figure 1, it is clear that this is not the case; the engine is surfacing content that is skewed positively (toward *helps*). The magnitude of this bias is quite startling and we explore its characteristics and its ramifications in more detail in the remainder of the article.

This study make the following contributions:

— Introduces the notion of *content bias* in search, where bias is defined in terms of a deviation from the truth rather than simply a skew toward a viewpoint, site, etc. as has been considered in previous work, e.g., [Azzopardi and Vinay 2008].
— Studies content biases in online health search for queries aimed at determining the efficacy of medical interventions, using ground truth on intervention effectiveness via authoritative medical reviews sourced from the Cochrane Collaboration.
— Quantifies the extent of content biases in health search results for queries of this type, and studies its source, including content and aggregated user behavior.
— Estimates the effect of content bias on SERP examination behavior, and shows that by controlling for this bias, a more accurate behavioral signal—potentially useful as implicit feedback (e.g., [Joachims 2002; Agichtein et al. 2006])—is attainable.
— Demonstrates significant variation in both accuracy and result skewness across different Cochrane reviews, and shows that reviews with no variation in accuracy or skew can likely be explained by missing content in the search engine index.
— Establishes the query terms that have most impact on content biases in our setting. Given an understanding of the role of each query term, we demonstrate some early promise in automatic query alteration methods that use term substitution (i.e., replacing troublesome terms with suitable synonyms) to improve result accuracy.
— Offers design implications for Web search providers to mitigate some of the effects of content bias, and ultimately improve the accuracy of the results that they serve.

The remainder of this article is structured as follows. In Section 2 we present relevant related work in areas such as biased result presentation and examination, topical skew in result sets, question answering, and health search. Our study is described in Section

3. Section 4 describes the findings, grouped by research question. We discuss our findings and their implications in Section 5, and conclude in Section 6.

2. RELATED WORK

The research in a number of areas is relevant to the work described in this article. This includes research on: (i) bias and diversity in result sets; (ii) beliefs and cognitive biases in search; (iii) search result examination and result ranking from aggregated behavior; (iv) support for question answering; (v) the quality of online health information, and; (vi) health-related information seeking. We now describe related work in each of these research areas in turn. We conclude the section by summarizing the contributions of this specific article over prior work.

2.1 Bias and Diversity in Search Results

Search engine ranking algorithms can exhibit biases in the information that they cover [Vaughn and Thelwall 2004] and how they choose to order search results [Mowshowitz and Kawaguchi 2002a; 2002b]. Research in this area has considered national biases in search results (favoring Websites from particular countries), and has shown that it is website visibility, in terms of inlinks, not the language that influenced whether a site appears in a search engine index [Vaughn and Thelwall 2004]. Others have considered whether search engines are presenting a fair ordering of results by comparing the search results from one engine against a pooled set of search results from many engines [Mowshowitz and Kawaguchi 2002a; 2002b]. Mowshowitz and Kawaguchi [2002a] claimed that the choice of search terms is unimportant in determining bias per their bias definition. In our setting, we show that the specific query formulation can be critical in determining the veracity of the search results returned by the search engine.

Bias has been studied in terms of inclusion or exclusion of particular results in search result rankings. Bharat and Broder [1998] studied the effect of various biases in the context of overlap between search engines, including ranking biases (influenced by search engine ranking strategy), and query bias (associated with query processing factors, e.g., whether query terms are handled conjunctively or disjunctively), as well as a number of biases related specifically to their measurement methods (e.g., checking bias, experimental bias, malicious bias). Azzopardi and Owens [2009] showed that search engines had a predilection toward particular news Websites, with implications on the perspectives to which searchers are exposed. Lawrence and Giles [1999] studied the impact of different amounts of website coverage in influencing the degree of bias in results. Others have focused on addressing such biases by increasing the visibility of pages to search crawlers [Upstill et al. 2002; Dasgupta et al. 2007].

Focusing more specifically on the role of the ranking function rather than the indexing, researchers have studied the role of document attributes such as bias toward document length [Singhal et al. 1996] or a bias toward popular Web pages that receive many incoming hyperlinks from other Websites [Pandey et al. 2004]. These factors can lead to such pages being favored by the search engine, irrespective of their relevance to the search query. Azzopardi and Vinay [2008] proposed a measure of *retrievability* to quantify the ease with which particular documents can be retrieved by search systems. This can be used to estimate degrees of search engine bias and is related to the research reported in this article. Wilkie and Azzopardi [2014] consider bias in term weighting schemes used in retrieval operations, and the extent to which these schemes favor certain documents given document characteristics. Our research differs in that we focus on characterizing the accuracy and the nature of the information surfaced by the search engine and its impact on subsequent search behavior, and search outcomes. Our goal is not to propose another method for measuring the degree of bias in search

results. Instead we focus on demonstrating the significant inaccuracies that exist in health search results, on attempting to explain the content bias, and on proposing bias mitigation strategies to help search systems direct searchers to accurate answer pages.

Biases in the presentation of results on personal and controversial topics have also been examined [Pariser 2011; Kacimi and Gamper 2011; 2013; Yom-Tov et al. 2014]. Via functionality such as search personalization, search engines can filter available information to only that supportive of the searcher's position [Pariser 2011]. Methods have been proposed to expose searchers to opposing perspectives [Yom-Tov et al. 2014], to promote a balanced view on important issues and encourage civil discourse. Kacimi and Gamper [2011; 2013] seek to mine opinions on underrepresented perspectives in search results for subjective queries. Their method enhances results by considering semantic and sentiment diversity in addition to relevance during ranking operations. Kacimi and Gamper's also present a summary of different arguments and sentiments related to the query topic. Other research has focused on the detection of controversy or disputed claims (including those with a medical focus) in Web pages [Ennals et al. 2010; Dori-Hacohen and Allan 2013], the generation of summaries of opinions on political controversies [Awadallah et al. 2012], and teaching users about controversial topics by highlighting contrasting viewpoints [Vydiswaran et al., 2012].

Our focus is different: we target scenarios where people seek answers to questions where there is an accepted truth, and the search engine (due to a number of factors) may return pages containing incorrect answers. Bias and controversy has also been modeled in the context of reviewing, considering the source and target of the review in addition to the review itself [Lauw et al. 2006]. This is relevant in our setting given the use of human-generated Cochrane reviews—meta-analyses of other studies—as the ground truth in our analysis. Research on result diversity (e.g., [Zhai et al. 2005]) is also relevant, but typically focuses on aspects such as topical variance, rather than results skewed toward erroneous outcomes. Other research has considered differences in retrieval performance for variations of the same search intent [Belkin et al. 1993; Ingwersen 1994], but targeted relevance and not biased perspectives or accuracy.

2.2 Biases in Search Behavior

Biases in the results returned toward one perspective can reinforce existing beliefs and leave searchers susceptible to the effects of information availability, which have been well studied in the psychology community [Tversky and Kahneman 1973; Simon 1991] where their ability to make rational decisions is limited by the information that they have access to [Simon 1991] and the ease with which information is recalled maps to likelihood estimates [Tversky and Kahneman 1973]. White [2013] focused on biased beliefs in health search and the role of search engines in reinforcing those beliefs. However, that work did not focus on content bias or use an authoritative ground truth as we do in this research. Instead, it used the independent opinions of two physicians, and disagreement among physicians is well known [Inlander 1993]. This creates some concern regarding the quality of ground truth data in that particular study.

A number of authors have proposed methods for using behavioral data of various forms—including queries, result clicks, and post-click navigation behavior—to improve search-result relevance [Joachims 2002; Agichtein et al. 2006; Bilenko and White 2008]. Despite its utility for ranking, search behavior can still be affected by biases related to the presentation order of search results on the SERP. Joachims et al. [2005] analyzed searchers' decision processes via search-result clickthrough and gaze tracking, and compared implicit feedback from result clicks against manual judgments. They found that clicks are informative but biased (favoring results located at higher rank positions (near the top of the list) irrespective of relevance), yet *relative* result preferences

derived from clicks mirror searchers' preferences. Searcher models can capitalize on such consistent behavior to infer result attractiveness and document relevance (e.g., [Craswell et al. 2008]).

Other factors beyond rank position can introduce bias into search. Clarke et al. [2007] introduced click inversions as a way to study features of result captions that increase caption attractiveness. Yue et al. [2010] studied connections between caption attractiveness and search behavior. Searchers in their study showed a click preference for results with more attractive titles. Beyond captions, Ieong et al. [2010] studied the effect of domain biases, whereby a result is deemed more relevant because of its source domain. Recent research by White and Horvitz [2013] found that potentially-alarming content appearing in result captions (e.g., mentions of serious illnesses in captions presented for symptom queries) lead to increased user engagement. Over time all of these biases can lead to focus of attention on top-ranked content and can create a cycle of preferential attachment whereby clicks reinforce popular results [Cho and Roy 2004] and as mentioned earlier, popularity can be important in ranking [Pandey et al. 2004]. That said, for some queries where there are differences in the search target, this bias can be offset by the heterogeneity of searchers' topical interests [Fortunato et al. 2006].

2.3 Support for Question Answering

Support for question answering (QA) in search engines has also been proposed, and question queries are increasingly common [Pang and Kumar 2011]. To help people answer questions, question-answering methods have been integrated into search systems [Dumais et al. 2002; Collins-Thompson et al. 2004]. Search engines can offer direct answers on search engine result pages (SERPs) for some simple questions (e.g., basic algebra, conversions, weather) [Chilton and Teevan 2011], although these methods can have low query coverage. Questions such as those associated with determining the efficacy of medical interventions that we focus on in this study are unlikely to receive dedicated answers, and searchers must rely on the answer pages returned by the search engine in the results to address their information needs. Beyond answering questions directly, measures of Web page quality have been used to improve the reliability of the search results processed during crawling and indexing [Tang et al. 2005], and in ranking using the hyperlink structure connecting documents on the Web [Brin and Page 1998; Kleinberg 1999], or even the online behavior of people during Web browsing activity [Richardson et al. 2006]. Collins-Thompson et al. [2004] demonstrated a relationship between retrieval quality and accuracy in QA systems. Kelly et al. [2002] showed differences in results retrieved depending on the nature of the question, specifically contrasting the effect of task- vs. fact-oriented questions.

2.4 Quality of Online Health Information

Online health information is often of low quality [Bengeri and Pluye 2003]. Surveys have shown that around 75% of online health seekers ignore key quality indicators such as source validity or source creation date when examining health content [Fox 2006]. Cline and Haynes [2001] suggest that public health professionals should be concerned about the prevalence of online health seeking; in their study they consider the potential benefits of this activity, synthesize quality concerns, and identify criteria that could be used to evaluate online health information. Eysenbach and Kohler [2002] systematically reviewed health Website evaluations and found that the most frequently-applied quality criteria included accuracy, completeness, and design (e.g., visual appeal, readability). They also note that 70% of the studies that they examined concluded that the quality of health-related Web content is low.

To address quality concerns, services have emerged that offer external verification on the reliability of health-related web content (e.g., Health on the Net (hon.ch) and URAC (urac.org)). These sites assign quality scores to pages based on human review of their content; although importantly, they do not verify the correctness of any claims made on those sites. These labels, and other reliability signals, have been used for ranking within specialized websites [Gaudinat et al. 2006] or to predict escalations in concern following the review of Web content [White and Horvitz 2010b].

Models have been developed to automatically label purported treatments on the Web (e.g., in the case of cancer [Aphinyanaphongs and Aliferis 2007]). In addition to their use in internal search engine operation, measures of Web page quality have also be shown to be valuable in supporting result selection decisions [Schwarz and Morris 2011] and can impact searcher trust [Sillence et al. 2004]. Schwarz and Morris [2011] identify page features associated with the credibility of online content, and present methods to augment search-result presentation with credibility features to help people find more trustworthy information and make more reliable decisions. Sillence et al. [2004] studied the influence of design and content on the trust and mistrust of health Websites via an observational study with a small number of participants engaged in structured and unstructured search sessions. They found that aspects of the design could engender mistrust in the content, whereas the credibility of information and personalization of content served to engender trust.

2.5 Health-Related Search

Spink et al. [2004] characterized healthcare-related queries on Web search engines, and showed that users were gradually shifting from general-purpose search engines to specialized Web sites for medical- and health-related queries. Bhavnani et al. [2003] demonstrated that term co-occurrence counts for medical symptoms and disorders on Web pages can reasonably predict the degree of influence on search behavior. Ayers and Kronenfeld [2007] also utilized Web usage logs to perform a multiple regression analysis exploring the relationship between chronic medical conditions and frequency of Web use, as well as changes in health behavior due to frequency of Web use. Their findings suggest that it was not the presence of one particular chronic illness, but rather the total number of chronic conditions that determines the nature of Web use. They also found that the more frequently someone uses the Web as a source of health information, the more likely they are to alter their health behavior. Ofran et al. [2012] studied aspects of cancer searching using search logs, and showed that the information needs of cancer searchers transitioned between five discrete states wherein different types of information was sought. Other research on health search has considered the short- and long-term impact of Web content on searchers' medical concerns [White and Horvitz 2012] or transitions from Web search to the utilization of healthcare resources in the real world [White and Horvitz 2010a]. Lauckner and Hsieh [2013] studied the effect of health content on the emotional state of Web searchers posing queries for medical symptoms. They showed that presenting serious illnesses at higher ranked positions led to negative outcomes such as heightened searcher anxiety.

Researchers have studied the search behavior of medical domain experts [Bhavnani 2002; Hersh and Hickam 1998; Hersh et al. 2002; Wildemuth 2004] to better understand the behavior of those with specialist domain knowledge. Hersh and Hickam [1998] review research from medical informatics and information science on how physicians use search systems to support clinical question answering and decision making. They found that retrieval technology was inadequate for this purpose and generally retrieved less than half of the relevant articles on a topic. They follow up this review with a study of how medical and nurse practitioner students use MEDLINE to

gather evidence for clinical question-answering [Hersh et al. 2002]. They show that these participants were only moderately successful at answering clinical questions with the help of literature searching. Bhavnani [2002] observed healthcare and online shopping experts while they performed search tasks within and beyond their domains of expertise, and identified domain-specific search strategies in each domain, and that such search knowledge is not automatically acquired from general-purpose search engines. Wildemuth [2004] performed a longitudinal study on the tactics used by medical students searching a database in microbiology. She showed that over time changes in students' search tactics were observed as domain knowledge increased.

Recent research has developed search systems to support the retrieval of medical content, primarily for clinical settings [Limsopatham et al. 2012; Koopman et al. 2012; Cogley et al. 2013]. Hersh [2008] also provides an excellent overview of research and practice in medical information retrieval, as well as future directions. Other research has applied aggregated health search behavior mined from many searchers for public health purposes such as detecting adverse drug reactions and drug interactions [White et al. 2013; 2014; Yom-Tov and Gabrilovich 2013], monitoring diet and nutrition in populations [West et al. 2013], or tracking epidemics [Ginsberg et al. 2009].

2.6 Contributions over Previous Work

The research presented in this paper extends previous work in a number of ways. First, we introduce the notion of content bias (i.e., defined in terms of skew toward a particular (usually positive) answer and a deviation from reality). Second, to measure content bias we leverage a widely-recognized ground truth on the efficacy of medical interventions (rather than potentially less reliable sources such as those used in previous work [White 2013]), with queries from searchers with similar information needs. Third, we quantify the extent to which retrieved and indexed content is biased toward particular outcomes, and its impact on search behavior, result accuracy, and result skewness. Fourth, we examine the contribution of different ranking signals in toward content bias in search results. Fifth, we demonstrate the importance of query formulation in content bias and show how accuracy can be improved via global term substitution. Finally, given our findings, we offer specific recommendations to search providers to help mitigate content biases, and show early promise in automatic query alteration methods that may counteract the effects of content bias.

3. STUDY

We now present an overview of the study. This includes the research questions that we seek to answer, the ground truth, the data analyzed, and the mapping of search queries and results to ground truth, which allows us to measure content bias directly.

3.1 Research Questions

There were four questions that motivated our research:

— **RQ1:** To what extent do answer pages in results for question queries exhibit bias (deviations from the truth), and what is the impact on search outcomes?
— **RQ2:** What are some potential sources of bias in search results? Explanations for this include crawled content in the index and the use of feedback from aggregated user behavior, which may exhibit its own biases and be affected by content bias.
— **RQ3:** What is the effect of content biases on result examination behavior (a central component in search engine ranking algorithms), and can the accuracy of behavioral signals be improved by taking steps to reduce biases in logged search behavior?
— **RQ4:** What role does query formulation play in content biases, and how can the query-result relationship be leveraged to improve search-result accuracy?

In answering our four questions, we obtain insights about aspects of content bias in the retrieval process that we use to inform specific design recommendations for bias mitigation in Web search engines. We present these recommendations in Section 5.

3.2 Ground Truth

The definition of content bias introduced earlier is derived from that proposed by White [2013], which states that bias in search describes a situation "*where searchers seek or are presented with information that significantly deviates from the truth*."[2] As such, we required a dependable source of ground truth data against which to assess the degree of content bias in the results that were returned and indexed by the search engine. We focus on medical interventions (i.e., measures whose purpose is to improve health or alter the course of a disease). Given that they seek out evidence and answers, people searching for information about the efficacy of these interventions could be particularly susceptible to biases in information availability. In addition, accurate answers are important since health seekers frequently use search outcomes to inform healthcare utilization decisions [White and Horvitz 2009b; Fox and Duggan 2013].

There are many studies in the medical community on the efficacy of interventions (e.g., [Wright and Weinstein 1998]). For a given intervention, the best available clinical evidence may be summarized and analyzed by panels of experts, and presented as a Cochrane review. Cochrane reviews are systematic reviews of primary research in human health care and health policy. These are internationally recognized as the highest standard in evidence-based health care [Higgins 2008]. Cochrane reviews are used by physicians and healthcare practitioners (in combination with their own clinical expertise) in making evidence-based treatment decisions [Sackett et al. 1996]. They have been found to be more recent and rigorous than systematic reviews and meta-analyses published in paper-based journals [Jadad et al. 1998] or industry reviews, such as those involving pharmaceuticals [Jørgensen et al. 2006]. Direct analysis of the review quality in comparison to other systematic reviews has shown that Cochrane reviews are of superior quality than other review sources [Petticrew et al. 2002].

The reviews investigate the effect of interventions for prevention, treatment and rehabilitation. They also assess the accuracy of diagnostic tests for a given condition in a specific patient group and setting. Each review addresses a clearly formulated question, e.g., *Can melatonin prevent or treat jet lag?* During creation of the review, all existing primary research on a topic that meets certain criteria is sought and collated, and then it is assessed by a panel of medical experts using stringent guidelines to establish the existence of conclusive evidence about a specific treatment. The reviews are updated regularly to ensure that treatment decisions are based on the most recent and reliable evidence [Higgins 2008]. Abstracts of the reviews are available on the Cochrane library website (cochrane.org/cochrane-reviews). These comprise multiple sections, including title, background, objectives, methods, results, conclusions, and a plain language summary. Figure 2 shows the title, background, and plain language summary for one review. We use these three fields in later analysis.

We obtained 4906 abstracts from the Cochrane website for research purposes. The reviews discuss a range of treatment options, with titles including *Exercise for depression*, *Topical treatments for fungal infections of the skin and nails of the foot*, and *Cranberries for treating urinary tract infections*. We joined the content of these reviews

[2] Note that this is different from other related notions of bias that have been studied by the information retrieval community, which has focused on the failure to retrieve particular results or content [Azzopardi and Vinay 2008], but have largely ignored the impact of content skew on result accuracy, and the implications of that bias on search outcomes.

Title: *Melatonin for the prevention and treatment of jet lag*

Background: *Jet lag commonly affects air travelers who cross several time zones. It results from the body's internal rhythms being out of step with the day-night cycle at the destination. Melatonin is a pineal hormone that plays a central part in regulating bodily rhythms and has been used as a drug to re-align them with the outside world.*

Summary: *Melatonin is remarkably effective in preventing or reducing jet lag, and occasional short-term use appears to be safe. It should be recommended to adult travelers flying across five or more time zones, particularly in an easterly direction, and especially if they have experienced jet lag on previous journeys. Travelers crossing 2-4 time zones can also use it if need be.*

Figure 2. Title, background, and plain language summary from sample Cochrane review on melatonin for jet lag (label=*helps*).

against the queries using a multi-step matching methodology described in Section 3.4. Many of the Cochrane reviews were highly specific, focusing on detailed treatment options (e.g., one review has the title *Hypertonic saline solution administered via nebulizer for acute bronchiolitis in infants*), and sufficient matches with logged queries could not be obtained, even after replacing some of the more complex terminology with simpler variants. We ignored non-matching reviews in our analysis. For queries corresponding to the matching reviews, we could obtain their results (both SERP captions and the content of each of the search results) from the Microsoft Bing search engine, and search interactions from the behavioral logs on the same engine. The Cochrane reviews provided us with the ground truth data upon which to assess content bias in these results and its impact on people's search behavior.

3.3 Data

We utilized different data sources in our analysis of biases (specifically search logs, result scrapes, and human labels). We now describe the data sources in more detail.

3.3.1 Search Engine Logs

We automatically extracted question queries from a random sample of the logs of queries issued by over 10M users of the popular Microsoft Bing search engine during a three-month period from July to September 2013. The data includes user identifiers, timestamps, queries, result clicks, and the captions (titles, snippets, and URLs) of each of the top 10 results. To remove variability from cultural and linguistic variations in search behavior, we only include log entries from searchers in the English-speaking United States locale. In addition to queries and their associated results, we also extracted search result clicks, dwell times, and the position of the query in the session.

Given these search engine logs, we sought to extract queries which suggested that the searcher was seeking information on the efficacy of a medical intervention. To be more confident that they had that type of intent, we targeted cases where we observed searchers asking questions directly to Bing via queries. Questions started with words such as "can", "should", "does" and had significant overlap with the Cochrane reviews. To help ensure data quality, we performed the following additional filtering: (i) selected queries issued by at least five users, (ii) selected SERPs with same 10 results/captions and same result ordering across all instances of the query in the three-month period, and; (iii) focused on query instances that were either the only query in the session or the terminal query in the session with no preceding queries with term overlap. Filter #2 ensured consistent results and captions for our study of SERP examination behavior (Section 4.2). Filter #3 provided more certainty that users had terminated their search

with that query; useful for inferring answer attainment. Queries were also normalized: they were lowercased, surplus whitespace was trimmed, and punctuation was removed.

3.3.2 Search Engine Results

The logs described in the previous section only provided access to the top-10 results presented at query time. Since we were interested in probing much deeper in the ranking (e.g., to better understand biases in the index rather than only the top results) we scraped the content of the top-ranked results from the Microsoft Bing search engine using a publically-available API. For each query, we obtained the search results and the full-text content of up to the top 1000 results, or as many results as were available in the index, whichever was smaller. The content of the pages is required for some of our analysis of result ranking. Across all queries whose results were scraped, the average number of results per query was 769 (min=280, max=1000, median=783). This served as an estimate for all matching content in the index. While there may be queries for which the search engine returned more than 1000 results via the API, for 98.5% of queries in our set all page matches in the index were attained in less than 1000 results. We believe that this deep sampling method could therefore serve as a reasonable estimate for the contents of the search engine index, in the absence of access to the index directly (which we lacked in this study). Importantly, this approach is also reusable by researchers not affiliated directly with search providers.

The scrape of the search results happened immediately after the July-September timeframe used for search log mining. While we could not guarantee that the same results were returned by the API as were logged, we do not rely on the presence of direct overlap between the logged and the scraped result sets for any of the analysis reported in this article. We also obtained the static rank score for each result from the engine, which denotes the engine's query independent quality estimate for the page (See [Richardson et al. 2006] for more details on static rank in search engines).

3.3.3 Page Labels

Given that we had results returned by Bing and the logs of the results selected by searchers, we wanted to understand the nature of the answers contained within those pages. To do this, we created a human-intelligence task and recruited crowdsourced judges. The task was recognition oriented and presented judges with an intervention query (e.g., [does echinacea cure colds]), a Web page, and a set of response options to indicate the nature of the answer that appeared on the Web page if one were shown. The following five response options were provided to judges:[3]

— **Helps only:** Web page only states that the medical intervention is an effective treatment option.
— **Might help only:** Web page only states that the medical intervention is a potentially-effective treatment option.
— **Inconclusive:** Web page states that medical intervention is both an effective treatment option and that it is ineffective, or explicitly states that it is not known whether or not the intervention helps.
— **Might not help only:** Web page only states that the medical intervention is a potentially-ineffective treatment option.
— **Does not help only:** Web page only states that the medical intervention is an ineffective treatment option.

[3] Even though only three-level ratings were used in our later assessments of search engine bias, the labels collected were also used for other studies where the more granular rating scheme was necessary.

Figure 3. Example caption that was assigned the label *helps* and *does not help* per the definitions introduced earlier.

If both answers appeared on the page, irrespective of sequence order, then the judge was instructed to label the page as *Inconclusive*. Judges were offered two additional options: (i) **No answer:** the page shared terms with the question query but did not offer an answer, and (ii) **Error:** the judge could not load the page.

There were many answer pages for which we required human labels. Although methods exist for extracting answers automatically, e.g., [Abney et al. 2000], we were concerned about the reliability of these methods given page complexity. Considering overlap between query variants for the same Cochrane review, it would have been infeasible to judge all pages given the time and costs involved. In light of our research questions, we only required judgments for the top 10 results and an equal distribution of judgments from the top 1000. As such, we solicited judgments from crowdworkers for all clicked pages, all pages in the top 10, and pages at every 50 rank positions, i.e., $r \in \{1\text{-}10,50,100,\ldots,1000\}$. This ensured good coverage of the two aspects of the result ranking that were of particular interest in our analysis (i.e., the first page of results and the distribution across all available (matching) results in the search engine index).

Crowdsourced judges were provided to Microsoft Corporation under contract by Clickworker.com. They resided in the United States and were required to be fluent in English. They were compensated financially for each judgment that they provided. To avoid skewing data toward any one judge, we imposed a 100-label limit on each judge. For each query-URL pair we obtained at least three labels and at most five, until a simple majority out of five was reached (i.e., at least three judges agreeing on a label). Page access errors were encountered for 2.9% of pages. These pages were ignored in our analysis. We obtained a three-judge majority for 90.4% of pages (excluding errors): 83.1% of pages had agreement among three judges and did not require more judgments.

3.3.4 Caption Labels

To more fully understand people's SERP examination behavior, we needed to consider the content of the captions presented. Captions have been shown to have significant impact on result examination behavior [Clarke et al. 2007; Yue et al. 2010]. To this end, we performed labeling of the captions using the same definitions as were available for the pages (minus the *Error* response option). We also recruited crowdsourced judges to provide answer labels for captions. Judges were sourced from the same pool as used for the page judgments, but different judges to those used in the generation of the page labels. Given a question query, judges were asked to label the content of the caption using the same six labeling options as were available for the pages. Figure 3 provides example captions labeled as *helps* and *does not help* by judges.

The statistics for the caption judging were similar to that of the page judging. Agreement between page and caption judgments was also high (90%). Analysis of the judgments revealed that there was good agreement between judges for this task. Agreement between at least three judges was attained for 92.3% of query-caption pairs,

Table I. Example question queries resulting from the filtering process, per truth label.

Helps	Inconclusive	Does not help
does melatonin work for jet lag	do ear wax drops work	does magnesium stop cramps
do antibiotics cure whooping cough	do orthodontics help tmj	do insoles help back pain
does tramadol treat nerve pain	can acupuncture treat uterine fibroids	can hydroxyzine be used for anxiety
can folic acid help with depression	do antibiotics help with pneumonia	does cinnamon help for diabetes
can laxatives help hemorrhoids	can yoga fix seizures	can probiotics help eczema

with 86.4% of the pairs reaching a simple majority (three of five) with the minimum of three judges. While we might expect there to be more *inconclusive* labels for the captions (since they are shorter in length) this was not evident in the results attained. The captions were query-biased and the answers in the documents were typically short. This meant that the answer often appeared in the caption directly.

To estimate the reliability of the judgments for page and caption answer judging, we created a test set of 100 query-page and 100 query-caption pairs. We assigned each set to 10 judges, each of whom judged all 100 pairs in the set. The Fleiss' multi-rater κ [Fleiss 1971], capturing agreement between the two sets of 10 judges, was 0.844 and 0.858 for page- and caption-judging tasks respectively. This signified strong agreement and increased our confidence in the use of their labeled data for our bias estimates.

3.4 Mapping Question Queries to Reviews

We mapped the question queries from the search logs to the matching Cochrane reviews to obtain the ground truth answer for each question query. This allowed us to assess bias directly. In doing so, we followed this three-step process:

— **Overlap with titles:** The titles of Cochrane reviews are observed to follow the template <intervention> for <condition>. To match with a particular review, we required that both the intervention and the condition appear in the candidate query. To improve coverage, we used synonyms for both the intervention and the condition sourced from the Unified Medical Language System (UMLS) [Bodenreider 2004]. The UMLS is a well-known medical repository comprising over 60 families of biomedical vocabularies, and developed by the United States National Library of Medicine. It integrates over two million names for 900,000 health-related concepts. For each of the matching concepts appearing in search queries, we generated a variant of that query for each of the synonyms in the UMLS.

— **Sequence order of terms:** To avoid cases where there may be term overlap, but the order of query terms implied a different intent (e.g., [does the common cold increase zinc levels] matching against the review title "Zinc for the common cold") we imposed an order constraint on the terms in the queries. Specifically, we required that the intervention preceded the condition in the candidate query. We should note that the sequence order of query terms is highly language dependent. We focus only on queries from the U.S. English locale in our analysis.

— **Verification with human judges:** The previous two steps were automated, important to be able to handle large volumes of queries. They generated a filtered set of 2495 distinct queries that was small enough to verify manually. To ensure that the queries we selected were high quality, we created a human judgment task. Crowdworkers from Clickworker.com were used to verify that the candidate query matched the intent expressed in the Cochrane review. Judges were provided with the query and the title and background of the review (see Figure 2 from earlier for

an example of these fields). They were asked to indicate on a three-point scale—*yes*, *somewhat*, and *no*—whether the query had the same intent as the Cochrane review. Each query was reviewed by at least two judges and up to three judges to obtain a simple majority (two from three). We retained queries where the majority was *yes*.

There were 268 Cochrane reviews for which matches were found with this methodology, with 1342 distinct matching queries (and tens of thousands of matching query instances in our log data). Following these steps pruned our data significantly, but they were necessary to ensure that high quality queries were chosen, e.g., [do probiotics help colitis]. Table I lists a random sample of five queries generated via this approach for each of the three types of truth label described in the next section. We can see that the questions are generally phrased positively, as is the case for most questions in Web search settings and beyond [Wason 1960; White 2013]. This positive framing may have some impact on how the search engine interprets the query and we explore the effect of query formulation in more detail later in the article.

It is important to examine the question queries employed in this analysis since later in the article we show that query formulation can significantly affect the accuracy of the results retrieved by the engine. We therefore wanted to understand whether there were any differences in the nature of the three groups that may contribute to differences in performance. From visually inspecting the larger query set (and Table I) it is clear that queries in the different groups appear similar qualitatively. They are also of similar length (in terms of tokens): *helps* (mean (M)=5.92 terms, standard deviation (SD)=1.63); *inconclusive* (M=6.00, SD=1.64), and *does not help* (M=5.88, SD=1.57) (analysis of variance (ANOVA): $F(2,1339)=0.41$, $p=0.67$), and contain a similar fraction of terms that are shown to be important in later analysis (Tables VI and VII) (e.g., "help" was equally present in queries labeled *helps* (33.8%), *inconclusive* (35.4%), and *does not help* (36.1%); Chi-squared test $\chi^2(2)=0.71$, $p=0.70$). Before analyzing how search engines handle these queries, we needed an unambiguous label for the recommendation given by each Cochrane review.

3.5 Labeling Review Recommendations

Labeling review recommendations involved reading the Cochrane summary and assigning a label. This would have been a challenging task to crowdsource since it would require careful reading of the task description and consistent labeling across all 268 reviews. To address this concern, the authors of this article performed this task. Each of the authors reviewed the titles and plain language summary portion of each of the reviews independently and discussed disagreements, amending a small number of judgments in light of these discussions. Answers were provided on the following five-point scale: *helps*, *might help*, *inconclusive*, *might not help*, and *does not help* (the same scale as in Section 3.3.3). Table II shows the distribution of labels between judges, with the percentage representing the fraction of all judgments. The diagonal (dark shading) indicates exact agreement between the two judges.

The exact agreement between the judges was high (89.2%) and the Cohen's free-marginal κ was also strong (0.863), signifying almost perfect agreement. We use the free-marginal κ because we were not forced to assign a certain number of cases to each category [Brennan and Prediger 1981]. As shown in Table II, most disagreements were between *helps* and *might help*, and *might not help* and *does not help*. To improve label consistency we collapsed ratings into three groups: *helps*, *inconclusive*, and *does not help* (shown in light shading in Table II). This improved exact agreement to 97.4% (free-marginal $\kappa=0.959$). Although it may have been preferable to have physicians label the reviews, this would have been costly and time consuming. Since we observed

Table II. Label distribution for review recommendations between judges.
Key: H=Helps, MH=Might help, I=Inconclusive, MNH=Might not help, DNH=Does not help.
Dark shading=exact agreement, Light shading=merged in three-level grouping.

		Judge A				
		H	**MH**	**I**	**MNH**	**DNH**
Judge B	**H**	**19.4%**	3.7%	–	–	0.4%
	MH	–	**22.4%**	–	1.1%	–
	I	–	0.4%	**25.0%**	0.4%	–
	MNH	–	–	0.4%	**9.3%**	1.1%
	DNH	–	–	–	3.4%	**13.1%**

such high agreement we felt that it was acceptable to use the non-experts' recommendation labels. To reduce domain-specific terminology, which may require subject-matter expertise to interpret, we also used the plain language summary as the main source of evidence on intervention efficacy. Overall, 45.5% of the matching reviews were labeled *helps* (around half of the 45.5% were originally labeled as *might help*, but were assigned to *helps* upon label collapsing), 26.9% were labeled as *does not help* (around half of the 26.9% were *might not help* originally), and 25.0% were labeled as *inconclusive*. We only used reviews which had inter-rater agreement and ignored those remaining (2.6%). This label distribution provides our base rates from which we can assess bias generally. Before proceeding, we downsampled reviews and queries.

3.6 Downsampling Reviews and Queries

To be sure that the conclusions reached from our analysis were easily interpretable and reliable, we sought to create a balanced truth set comprising answers for the three outcomes. Given that *inconclusive* was the minority class (with 67 instances), we randomly downsampled the *helps* and *does not help* classes such that there were 67 reviews with each outcome; 201 reviews in total (75% of the full set of 268) and 33.3% on each. We use this set of reviews in the remainder of our analysis. Downsampling the queries from the 1342 reported earlier to the subset of 1062 that matched against the 201 reviews would result in an uneven distribution of queries per outcome. To prevent query-related bias toward particular outcomes, we also randomly sampled the queries within each of the categories so that there was an equal number of queries (four per review-answer pair) and a similar distribution of query terms (including question prefixes) for each of three answer labels. This resulted in a total of 804 distinct question queries upon which we base our analysis. Note that the downsampling had little impact on the statistics for query length or term presence reported in the previous section. The downsampling also addressed concerns with potential selection biases associated with intervention intent (e.g., are people simply more likely to search about interventions that help?), review authorship, or the query-review join.

3.7 Summary

In this section we have described the methodology employed to generate the dataset used to study content bias in search. We employed crowdsourced human labeling of results and associated captions, and evaluated the performance of the search engine using these labels to understand biases. We downsampled for interpretability and to counteract potential intrinsic biases in the data used for our experiments.

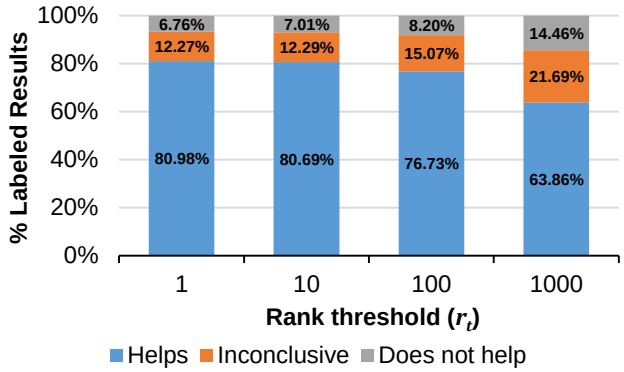


Figure 4. Distribution of answers in search results appearing at various rank thresholds.

4. FINDINGS

We now present the findings of the study, grouped by research question. We focus on the nature of the results returned, their accuracy, and other factors such as the effect of content biases on search behavior and impact of query formulation on these biases.

4.1 Biases in Result Rankings

To begin, we focus on result ranking. As shown earlier in Figure 1, the content of the search engine index is skewed positively (toward *helps*) and the ranking algorithm skews the answer pages provided even more positively. In this section, we seek to better understand the distribution of answer pages in the result set. We characterize the distribution of answers at different ranks, and the accuracy of the results returned. We also explore variations in these distributions with different ranking signals to better understand potential sources of any observed content bias in the results.

4.1.1 Answer Page Distribution

We used the deep scrape of search results from the search engine to study the answer page distribution at different rank positions. To avoid oversampling from queries with more results, we randomly selected one labeled page per query at rank $r \leq$ threshold r_t, where $r_t \in \{1,10,100,1000\}$. Since we had labels for the all of the top 10 and every 50 results, the random sampling for r_t=100 and r_t=1000 would have skewed the labeled data toward lower rank positions (i.e., closer to the top of the ranked list). To address this skew, for these two thresholds we sampled from a pool comprising the top rank position (r_t=1) and every 50 rank positions (i.e., 50, 100, 150, etc.). We generated such a distribution across the results for each of the queries in our set. We used the union of sampled answer pages over all queries at each r_t in our distributions. Since we had over 1000 distinct queries it was not necessary to generate multiple samples per query to obtain a sufficient sample size to assess content bias. Figure 4 shows the distribution of answer pages obtained at each of the four r_t values of interest using our sampling method. The proportions of *helps*, *inconclusive,* and *does not help* at r_t=10 and r_t=1000 match those in Figure 1. An answer page is offered at the top-ranked position (r=1) for 96.2% of queries. The fraction of pages with no answer to the question increases dramatically with rank from 3.8% at r_t=1 to 58.5% at r_t=1000. Pages at these higher rank positions may only match 1-2 query terms versus the top-ranked pages which may match against most or all of the terms.

Figure 4 offers a number of interesting insights about the answer pages:

Table III. Result accuracy at various rank thresholds. Also shown is accuracy broken out by truth, and correlations between availability and accuracy. Correlation significance using *t*-tests denoted as * $p < 0.05$, ** $p < 0.01$. † = excluding *no answer* as in Figure 4.

		r_t				Correlation with
		1	**10**	**100**	**1000**	**% *helps*-only**
All answers		0.3757	0.3790	0.4184	0.4576	–0.960*
Truth label	Helps	0.7842	0.7897	0.7658	0.6397	0.994**
	Inconclusive	0.2210	0.2072	0.3168	0.4887	–0.989*
	Does not help	0.1217	0.1279	0.1728	0.2444	–0.985*
% *helps*-only†		80.98%	80.69%	76.73%	63.86%	–

— The fraction of pages with *helps* answers only increases dramatically as we move closer to the top of the ranked list, with over 80% of the labeled pages at the first rank position containing *helps* only;
— Pages with *inconclusive* and *does not help* labels are fairly uncommon, across all rank positions, although the fraction that are *inconclusive* and *does not help* does increase (each by around 6% overall) at lower ranks. These larger values suggest that some negative/inconclusive answer pages may be present in the index, but the ranking algorithm positions them low in ranking (i.e., at ranks 100-1000), and;
— The distributions of the answers at each r_t significantly differs from the expected (our base rate) value of 33.3% per outcome (Chi-squared tests: all $\chi^2(2) \geq 14.43$, all $p < 0.001$), and clearly shows the extent of the content bias.[4]

In addition the quantity of answer pages in results, we also sought to understand ranking effects. To do this, we computed the relative ordering of the answer pages when all three labels were available. We computed this at r_t=10 since, as we mentioned earlier, we had labels for all of the top 10 results, removing the effect of missing labels. From the list we picked a random result with each of the three labels and recorded which result appeared higher in the list. We observed that *helps* appears first in the list for 70.1% of queries, *inconclusive* appears first for 18.3%, and *does not help* appears first for 11.6% ($p < 0.01$). A similar positive skew was noted when we focused on result lists with only two distinct labels. Beyond just answer page volume (in Figure 4), the search engine appears to favor positive information in how it orders answer pages. Considering biases in how people inspect results [Joachims et al. 2007], this rank bias may also contribute to biases in user-perceived beliefs about answer likelihoods.

An important consideration in addition to the nature of the results retrieved and their ordering is the *accuracy* of those results, defined as the fraction of times an answer page label matches the truth. This is important since our definition of bias depends on accuracy in terms of how well the truth is represented.

4.1.2 Answer Accuracy

Since searchers may rely on the ranking of the search engine to locate pages with answers [White and Horvitz 2009], it is important to understand the extent to which the search engine furnishes the correct answer. Given that we had page labels comprising the judgments for each of the pages and the truth for each of the Cochrane reviews that were analyzed, we could compute the answer accuracy of the search engine at different rank positions, when a label was available. For this purpose, we use the same sampling methodology as used for the analysis of answer distributions in

[4] Note that although some (around 1%) of the result sets contained fewer than 1000 results, the relative distributions of answer pages were highly similar to those where 1000 results were available.

Table IV. Result accuracy and percentage *helps* at top rank position for original ranking, and other re-ranking signals. Significance with tests of proportions denoted by * $p < 0.05$, *** $p < 0.001$.

Metric	Original ranking	Re-ranking signal		
		Content	Behavior	Quality
Accuracy (at r=1)	0.3757	0.3295	0.4501***	0.4166*
% Helps (at r=1)	80.98%	65.18%	74.28%	67.12%
ρ with orig. ranking	–	0.5971***	0.5196***	0.1937

the previous section. Table III shows accuracy at each r_t, and a breakdown by the truth. Also shown are the Pearson correlations between accuracy and *helps*-only volume.

Table III (row 1) shows the result accuracy is generally low at the higher ranks and near random (which would be 0.3334 given our truth data). When we move deeper in the ranking, lessening the impact of engine selection effects on the results we observe improved accuracy, e.g., 0.4576 at r_t=1000 (Z=5.53, p < 0.001). This change can be attributed to availability, i.e., % *helps*-only decreases with rank and there are strong correlations between accuracy and % *helps*-only (p < 0.05, even with the small n). Conditioning on the truth, Table III (rows 2-4) shows that accuracy for *helps* is higher than other answers, which are often inaccurate (all $\chi^2(2) \geq 10.01$, all p < 0.01). Search engines are likely to retrieve positive (*helps*) results, irrespective of the truth. This illustrates the challenge that searchers face when seeking answers: posing question queries where the truth is *helps* yields the correct answer often although perhaps not intentionally (simply due to content bias), but otherwise the result accuracy is poor.

4.1.3 Contribution of Ranking Signals

Many signals can contribute to content biases, including query-result content matching, authority information (e.g., inlinks from other websites [Brin and Page 1998; Kleinberg 1999]), and aggregated examination behavior across many searchers (often utilized as implicit feedback in ranking [Joachims 2002; Agichtein et al. 2006]). To obtain insight on the contribution of these signals during ranking and their effect on content biases, we explored the connection between them and the original ranking.

Given the variations reported thus far in our analysis, we wanted to understand the contributions of some of the primary features in ranking search results. Since we could not isolate individual features or feature classes from the production ranker itself, we instead used three sources of evidence known to contribute significantly to search engine rankings to re-rank the available (up to top 1000) results:

— **Content:** We obtained page content for each result in the list and applied the BM25 ranking model [Spärck-Jones et al. 2000]. BM25 is a competitive content-based ranking algorithm whose effectiveness has been demonstrated in numerous settings, including Web search [Craswell et al. 2003; Agichtein et al. 2006]. In implementing BM25, we used k_1=1.2 and b=0.75 (standard parameter settings), and the same stopword list as employed by the search engine.
— **User behavior:** All clicks for each of the search results across all queries in the one-year period preceding the July-September timeframe used in our analysis are mined from the search engine logs. These clicks are used to rank the results in descending order by click popularity. Since our queries were infrequent, we needed to use all instances of clicks on our results, irrespective of the query. Queries that share clicked URLs have been shown to be related [Beeferman and Berger 2000].
— **Quality:** Static rank used as quality estimate by search engine. It is based on query-independent page features, such as PageRank [Brin and Page 1998].

We had page content and page quality signals for all available results, up to 1000 per query. However, we did not have user behavior for all results. The average overlap between historical click data and returned results was on average 54.4 results (median=39). We used each signal to re-rank the results, and computed accuracy at the top rank position (r=1). To ensure a fair comparison between sources, we focused only on cases where all signals were available. We focus on the top-ranked result since for some queries (20.3%) 10 or more results were unavailable for user behavior. To better understand the relationship between the ranking from each signal and the original engine ranking, we computed the Spearman rank correlation coefficient (ρ) between the rankings, correcting for ties in scores and click frequencies as needed by assigning a rank equal to the average of their positions in ascending order of the values.

Table IV presents the accuracy and correlation values for each re-ranking signal. The results show that the correlation between the results and the quality scores is low (and non-significant), suggesting that quality alone may not be a primary contributor in the result ranking. Nonetheless, by considering the top search result from the quality-based re-ranking, we can observe that accuracy improves over the original ranking ($Z=2.01$, $p < 0.05$); suggesting that considering quality could be beneficial. The content signal is most correlated with the original ranking and also has the lowest accuracy. One explanation is that there is no consideration of the *quality* of answers in pages, which may be evident in other signals such as static rank and user behavior. We explore the contribution of content (as query terms) in more detail later. Although behavior is tightly coupled to results—engines learn from behavior, and behavior depends on what is shown—it is interesting to note that the correlation with the original ranking, although significant, is far from perfect, and that by ranking based on behavior, accuracy improves considerably ($Z=3.36$, $p < 0.001$). The effects of rank and quantity may direct attention away from accurate answers, but searchers may compensate somewhat with their click decisions.

Turning our attention to the fraction of the results at the top-rank position (r=1) that are positive (% *helps*-only), we can observe that the percentages greatly exceed the 33.3% that would be expected given the base rates and how we controlled for bias in our query sampling strategy (described in Section 3.6). The strong positive skew is apparent in each of the three sources, although more so in the ranking associated with user behavior ($\chi^2(2)=8.55$, $p=0.01$). We consider the relationship between content bias and searcher examination behavior in the next section.

4.1.4 Summary

We showed that for the queries examined in our study, there was a skew toward positive results, especially at high ranks, and this degraded the accuracy of the answer pages returned. The content-based features were most strongly correlated with the original ranking, but also had the lowest result accuracy. Informational question queries of this type are uncommon, so it is likely that the content match is more important than other signals such as historic search behavior. Later we explore the impact of content matching on ranking by considering the impact of individual query terms on result skew and accuracy. Search behavior is also correlated with the original ranking, but the accuracy is higher, perhaps because searchers mitigate some bias via SERP selection choices. We now explore examination behavior more detail.

4.2 Impact on Examination Behavior

We were interested in the impact of biased rankings on the results that people select. By controlling for this bias search engines could extract a cleaner behavioral signal from which engines could learn accurate answers. To do this, we consider the results

Table V. Fraction of clicks on each answer over all clicks, controlling for content biases (rank, quantity). Significance with tests of proportions denoted by ** $p < 0.01$, *** $p < 0.001$.

		Controlling for		
Caption label	**All clicks**	**Rank**	**Quantity**	**Rank *and* Quantity**
Helps	70.09%	66.81%	58.82%	54.95%
Inconclusive	14.42%	16.94%	20.87%	22.81%
Does not help	15.48%	17.25%	20.31%	22.24%
Answer accuracy	0.4477	0.4644	0.5253**	0.5615***
%Δ over all clicks	–	+3.73%	+17.33%	+25.42%

that people examined and then repeat this analysis, controlling for the effects of rank position. The results that people select reveal their preferences, but also their level of trust in the search engine response [Joachims et al. 2007], and are used by search engines to learn effective result rankings [Joachims 2002; Agichtein et al. 2006]. We mined the logs described in Section 3 (complete with logged SERPs containing the result URLs and captions) to study the results that people examined.[5] To better understand the impact of content bias on searchers' result examination behavior we focused our attention on two behavioral signals:

— **Clicks:** Examining result click occurrence irrespective of post-click behavior (using the caption labels from Section 3.3.4), and;
— **Answers:** Examining satisfied clicks (using the page labels from Section 3.3.3), defined as those with a dwell time ≥ 30 seconds or the last click in the session [Fox et al. 2005]. In the absence of confirmation from searchers, we used the last satisfied click as a proxy for the answer. We only know that the searcher reached the page and dwelled, not whether they located the answer. Analyses of gaze, cursor, and scrolling interactions could help make that determination [Guo and Agichtein 2012].

Considering each of these sources, we examine the nature of the clicks and the accuracy of the results found across the top 10 results. The "All Clicks" column in Table V shows these statistics. In addition, we also report on the effects of controlling for rank and quantity, both separately and in combination. The strong preference for positive information could be caused by the tendency of the search engine to rank pages containing positive answers above negative ones (as reported earlier and corroborated in prior work [White 2013]) and more *helps* pages generally. Given how searchers typically examine result lists (i.e., top-to-bottom [Joachims et al. 2007]), this could lead to an apparent preference for positive information caused by the search engine, even if one did not exist for searchers independent of the engine ranking. Unfortunately, since we performed this analysis retrospectively, we could not employ methods such as FairPairs [Radlinski and Joachims 2006] or interleaving [Chapelle et al. 2012] to counteract known biases in the logged data.

[5] Note that very little of the examination behavior was on Cochrane review links: across all queries, only one of the queries (0.09%) had a SERP (top-10) containing a Cochrane review result, and none of the clicks in our dataset were on results that were Cochrane reviews. Considering the top 1000 results from the search engine (our proxy for the search engine index) rather than the top 10 results, we find that there are 125 queries with a Cochrane review result in the top 1000. The failure to surface such authoritative information in the first page of results—even though it offers a definitive answer for the specific question query and is present in the search engine index for 10.95% of queries (125 of 1142)—highlights an aspect of the challenge for search systems in matching questions with answer pages.

To control for biases (i.e., the rank and quantity of answer pages in the search results), we did the following for each distinct query in our set (across all its instances):

— Control for *rank*:
  - Aggregate and count all observed result clicks (n_r) at each rank position r;
  - Find the r that minimizes n_r (i.e., $\min(n_r)$), and;
  - Randomly downsample clicks on search results at ranks other than r, so that all click counts equal $\min(n_r)$ and yield a uniform distribution of clicks across the results selected.

— Control for *quantity*:
  - Randomly downsample the clicks to yield a uniform distribution using a similar strategy as above, but based on the number of captions with *helps*, *inconclusive*, and *does not help* answers rather than click counts.

Table V reports the fraction of clicks on each answer type (caption labels), and the impact on answer accuracy (page labels) for different de-biasing strategies, including a combination of both rank and quantity de-biasing.

The results show a strong preference for positive (*helps*) content and that answer accuracy is fairly low (45%). These preferences may be amplified by result biases and drops significantly (from 70% *helps* to 55% *helps*) when controlling for the rank and quantity of answer pages (all $\chi^2(2) \geq 9.66$, all $p < 0.01$); with a larger decrease attained when controlling for quantity, given the large skew noted earlier. However, even after this attempt to de-bias the click data there is still a strong preference for positive (*helps*) content (over twice other answers), and more than expected given base rates of 33.3% per answer ($\chi^2(2)=9.86$, $p < 0.01$). This perhaps reflects an intrinsic searcher preference for positive information, also noted in earlier research [White 2013]. Importantly, controlling for these biases yields an accuracy improvement of 25.4% over the original behavioral signal ($p < 0.001$). Such de-biasing could yield cleaner implicit feedback to rank answer pages more accurately. Training machine-learned models that learn from less-biased implicit feedback signals to improve search-result accuracy is an important direction for future work, but is beyond the scope of this article.

4.3 Review and Query Formulation Effects

In Section 4.1.3 we demonstrated that the extent of content match was an important determinant of bias for question queries. We were also interested in understanding the impact of any review effects and the effect of question formulation decisions on content bias. Topic differences have been shown to impact search engine performance in offline settings such as the Text Retrieval Conference (TREC) [Harman and Buckley 2009]. In our context, it is important to understand the prevalence of the observed content biases across a range of different intents, albeit with a focus on medical interventions given the availability of the ground truth data that we required per our definition of bias. The specific query or even individual query terms that searchers select may also be important in determining the outcome of the search. In this section, we report on our analysis of both of these factors.

4.3.1 Variations in Search Engine Performance

To begin, we focus on variations in content bias in the search results given differences in query formulation. To capture intents for which there is a natural variation in their expression, we focused on the 63 reviews with three or more distinct queries noted in the log analysis described earlier, evenly split among the three answers labels; 21 per answer label. To understand the effect of specific query formulations on search engine performance we needed many query variants for the same review (more than were

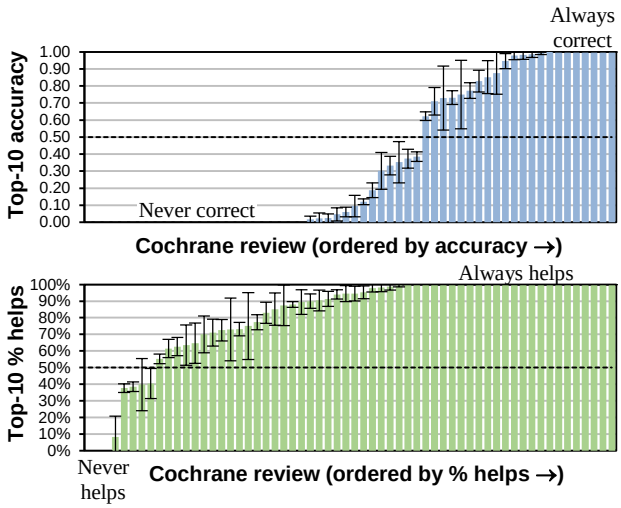


Figure 5. Within- and between-review variations in accuracy of the top-10 results / fraction answer pages labeled positive (*helps*). Error bars denote ± standard error of mean (SEM).

found in the search logs alone). To obtain this list, we ran a crowdsourcing study, provided judges with the Cochrane task descriptions (title and description; see Figure 2 from earlier for an example) and asked them to generate a question query for submission to a Web search engine. To improve representativeness, the task interface constrained the syntax of the query. Specifically, we required that queries started with a question word (e.g., "do", "can", etc.) and ended with a question mark. For each review, we recruited 20 judges and asked them to create a question. This yielded on average 13.62 distinct questions per review (median=16). These data allowed us to examine the impact of different variations of the same intent on the skew and accuracy of the top-ranked results from the search engine studied.

Figure 5 shows the accuracy and skew (measured in terms of the percentage of results in the top 10 results that have the label *helps*) statistics across all 63 Cochrane reviews, when ranked in ascending order by the measure of interest. Error bars denote the standard error of the mean and illustrate the extent of the variance in the measure value within each review. Although there were some review effects, the results were fairly consistent in pointing positive (*helps*) across most reviews. Some reviews (labeled as "always" or "never" in Figure 5) had no variation in accuracy or positive skew. Qualitative analysis of these cases revealed no systemic differences in the queries, the nature of the reviews, or the truth. Also, the consistency was not simply related to a small number of query variants for those reviews or high similarity between query variants in these reviews. The most likely explanation is the available evidence in the search engine index for or against a particular intervention, or lack thereof. Indeed, in analyzing the content of the index for these reviews we see that it is heavily skewed toward a particular outcome (e.g., for reviews under "Always helps" the percentage of *helps* content in the index is 94.5% vs. 63.9% (all reviews); $Z=6.93$, $p < 0.001$). When the ground truth is *inconclusive* or *does not help*, the index skew toward *helps* reduces the likelihood that documents with the correct answer will be retrieved. More research is needed to understand why this occurs and develop methods to ensure that pages with the appropriate answers are available to be ranked by the search engine. Related research in helping to ensure that content is crawled by search engines may be useful [Upstill et al. 2002; Dasgupta et al. 2007], but extensions are needed to also consider page quality in designing crawling and indexing strategies, e.g., [Tang et al. 2005].

Table VI. Example contingency table used to measure the impact of individual query terms on the metric (accuracy or skewness) across the top 10 search results.

		Change in measure	
		Increase	*Decrease*
Query Term	*Present*	*a*	*B*
	Absent	*c*	*d*

The within-review variance in engine performance suggests that even within the same review, how searchers phrase the question can affect the search engine response. The importance of query formulation has been studied in depth by the information retrieval community in various ways [Belkin et al. 1993; Ruthven 2003], but with a focus on relevance and not skew or accuracy. To better understand the within-review variance, we analyze the impact of particular query terms on engine performance.

4.3.2 Impact of Individual Terms

We used the same set of 63 reviews and their associated queries from the previous subsection, extracted all terms from all query variants within each review, and stemmed the terms using the Porter stemmer. We also excluded the terms appearing in the stopword list of the Microsoft Bing search engine since these were ignored by the engine in ranking. Note that some common terms such as "can" and "do" are not present in that stopword list. Given this set of terms, we could then understand the effect of each term on: (i) result accuracy in the top 10 and (ii) percentage of the results in top 10 with helps only (positive skew).

To perform this analysis, we computed the per-review average accuracy across all query terms, and then measured whether accuracy significantly increased or decreased (beyond the 95% confidence intervals) when each term was present or absent from the query. This formulation resulted in a 2×2 contingency table as shown in Table VI, with each cell containing the frequency counts for each of the four outcomes (denoted *a*-*d* in the table). For this analysis, we ignored cases where the accuracy did not change (increase or decrease) if the term was present or absent from the query.

A Chi-squared test (with one degree of freedom) was then applied to this table to determine the significance of any differences noted. Table VII presents the top 10 query terms ranked in descending order by their Chi-squared value. Also reported in the table is the directionality (i.e., whether the term increased or decreased top-10 result accuracy when it was present or absent). The table also displays the magnitude of the change in accuracy (% Change) that resulted from including each of these influential terms in the queries in our set (versus exclusion).

Table VII shows that there are particular query terms that yield large gains or losses in top-10 result accuracy. The most important terms ("can", "does", etc.) are often regarded as indiscriminative from a retrieval perspective, but they appear influential in this setting. The term "help" is likely to return *helps*-only pages due to the content match, and "can" denotes possibility and often matches against *helps*-only pages that also do so (e.g., a query such as [can exercise treat depression] would match pages with "exercise can treat depression"). Some of the query terms associated with result accuracy gains are recommendation-oriented (e.g., "recommend", "should", "people"), which may return pages describing others' experiences with the intervention (on sites like patientslikeme.com) and other terms target efficacy (e.g., "treat", "work").

Table VII. Query terms that are most influential on top-10 result accuracy, ranked in descending order by χ^2 value. Significance with $\chi^2(1)$ denoted by * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Query term	χ^2	Direction	% Change	Sig
help	15.137	↓	–32.33	***
can	14.502	↓	–29.57	***
does	13.339	↑	+28.44	***
recommend	12.580	↑	+29.98	***
cure	11.129	↓	–24.31	***
should	9.859	↑	+22.22	**
treat	9.277	↑	+20.95	**
people	8.026	↑	+19.47	**
do	6.671	↓	–21.63	**
work	5.153	↑	+17.80	*

Table VIII. Query terms that are most influential on % *helps* in the top 10 results, ranked in descending order by χ^2 value. Significance with $\chi^2(1)$ denoted by * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Query term	χ^2	Direction	% Change	Sig
good	17.323	↑	+44.39	***
prevent	14.888	↑	+23.30	***
lower	13.503	↑	+23.75	***
help	12.444	↑	+22.70	***
cure	11.653	↑	+16.47	***
effect	10.037	↓	–21.26	**
can	9.629	↑	+18.85	**
work	7.491	↓	–17.71	**
treat	5.214	↑	+17.45	*
aid	4.911	↓	–23.08	*

Next we turn our attention to the extent of positive skew in the search results, based on % *helps*, rather than accuracy. Table VIII shows the top-10 terms based on their $\chi^2(1)$ value. The table shows that query terms that specifically imply the potential efficacy of the intervention (vs. inquiring about its efficacy) (e.g., "good", "prevent", "lower") significantly increase the likelihood of *helps*-only results in the top 10. When the intervention does help, this can be useful, but when the intervention is not helpful, the inclusion of these terms and the resultant skew can degrade result accuracy. Terms that are perhaps more associated with seeking to assess efficacy (e.g., "effect", "work") are generally associated with a reduction in positive skew in the top-10 results.

Such knowledge of the impact of terms can have direct implications for search engines. To demonstrate the potential value in this analysis, let us consider a simplistic type of backend query alteration that is applicable by search engines. We re-computed the accuracy of the top-10 results for the global replacement of all instances of "help" in our queries with the synonymous term "treat" (e.g., the query [does zinc help the common cold] became [does zinc treat the common cold]), and separately replaced "can" with "does". The results of our analysis show that the accuracy of the answers appearing in pages in the top ten results increased by 16-20% via this simple global replacement (both $p < 0.05$). This also represents a reduction in content bias per our definition provided earlier. More nuanced versions of these rules could be developed to generate further improvements in accuracy. This could also be framed as a learning problem with the objective being to learn to alter the query such that bias in the results is minimized. In addition, rather than focusing only on result accuracy, a similar method could be applied to control for positive skew if, say, balanced perspectives and diversity were important in the results (as suggested in previous work [Kacimi and Gamper 2011; Yom-Tov et al. 2014]). The effectiveness of this

simplistic query-term substitution approach illustrates one way in which search engines could address content biases directly with low overhead.

4.4 Summary

Through our analysis of variance in accuracy and positive skew within and between Cochrane reviews, we observed between-review differences, but mostly positive skew and low accuracy. Our findings also demonstrated large within-review differences, depending on the specific query formulations employed; highlighting the importance of the query terms that are chosen. Some reviews had little or no variance (i.e., always *helps* or always *does not help*). These reviews were often missing index content related to particular outcomes. This is important, although may only be a real concern if the content currently indexed by the search engine is incorrect. We also found that there are query terms that contribute significantly to content biases. It is evident from our findings that by replacing troublesome terms with suitable synonyms (e.g., replace all "help" with "treat") we can mitigate some content biases in search engine results.

5. DISCUSSION AND IMPLICATIONS

We studied content biases in search results for queries associated with the efficacy of medical interventions, including their impact on search-result examination behavior. Controlling for the base rates per our experimental design, we showed that indexed content is skewed positively (toward *helps*), and that the search engine analyzed in our study may skew these results further given ranking signals, especially content match. Explanations for this skew, including authorship bias (e.g., commercial promotion of interventions) and other potential effects (e.g., the nature of the Cochrane data used as truth and potential implicit biases in the interventions reviewed) need to be further explored. We should also acknowledge that the definition of ground truth employed in this analysis is biased toward a particular view of medicine. The differences in opinion associated with different perspectives could contribute to some of the inaccuracies noted in this investigation. Previous studies have also observed skew in Web content concerning health-related outcomes, e.g., significant biases toward serious health conditions [White and Horvitz 2009] that can result in amplified anxiety among health information seekers [Lauckner and Hsieh 2013]. There are additional areas where positive skew in content is noted, e.g., the scientific literature contains more content reporting positive results [Fanelli 2012], and publication bias has been well studied in the medical community [Dickersin 1990; Easterbrook et al. 1991; Egger et al. 1997]. Positive publication bias may also exist in Web pages, and there may be more that is written online about scenarios where many people possess a particular opinion, irrespective of ground truth (e.g., on the viewpoint that there is a link between child vaccination and autism [Rochman 2011]), which could contribute to content biases.

We showed that there are particular terms in search queries that are important in determining skew and result accuracy. For reviews with little variance for different query formulations, we find that indexed content largely ignores one aspect. Along with the importance of content matching, query formulation appears critical in content bias, and our early findings suggest that result accuracy can be improved via query alterations. Particular terms may reflect authorship attributes (e.g., reliable pages may use "treat" more than "help") or how people express needs (e.g., initiating many question queries with "can"). People tend to frame questions positively [Wason 1960] and this has been observed previously in search engine query statements [White 2013], and anecdotally in this study. Although it is known that queries reflect comprised needs [Taylor 1968], bias affecting query formulation must be better understood. More generally, intervention-related search intentions may be skewed positively; as is

suggested by the skewed distribution in query matches (twice as many *helps* as the other outcomes) before the downsampling step described in Section 3.6. Search engines optimize content (including learned result rankings and crawling strategies) for the types of questions that searchers typically ask rather than normative base rates. So even after we created a uniform outcome distribution for our study we may still observe some residual effects from this optimization in the results analyzed in this study.

In our analysis, we assume that search engines will handle question queries via a ranked list of search results. However, search engines may also employ specialist support for handling such queries. Answers can be derived automatically from Web content [Dumais et al. 2002] or manually by humans [Bernstein et al. 2012], and presented to searchers directly on the SERP. These solutions may be appropriate for questions that assume a particular form (e.g., factoid questions such as "Who invented the telephone?"), and cases where there is a single (or commonly accepted) answer. For other types of questions such as those with many possible answers depending on the situation, search systems need to convey uncertainty. One way to do this is through result ranking, our focus in this study. Other presentation methods could be employed to visualize answer uncertainty and allow searchers to select the response that best matches their interests or even their beliefs (with appropriate warnings about biases).

We showed that result examination behavior was affected by content biases, but the behavioral signal was still more accurate than the results alone, and could be de-biased further to boost accuracy. The de-biasing method employed in this article simply downsampled clicks to handle potential biases introduced by skewed ranking of results toward *helps* and a similar skew in the availability of content. Employing this downsampling method led to a 25% increase in the accuracy of the derived click signal. In applying such a method in practice there may be issues with variance associated with the sampling, and more sophisticated methods for cleaning the behavioral signal in light of biased behaviors (e.g., [Yue et al. 2010]) need to be considered in this context. Evaluation methods such as FairPairs [Radlinski and Joachims 2006] or interleaving [Chapelle et al. 2012] can be employed, but these depend on being able to manipulate the search experience at the time the data are captured from searchers. The challenge of de-biasing such behavioral data *retrospectively* is non-trivial. However, we showed that it even with a simplistic technique involving downsampling logged clicks there can be gains in the accuracy of the implicit feedback signal. Further work is needed to understand the practical benefit of learning from features derived from this signal.

Even after controlling for rank and quantity, there is still a residual bias associated with aggregated search behavior. Specifically, we observed a strong preference for positive information (over twice other outcomes), that is similar to that observed in prior work [White 2013], and irrespective of the ground truth. The apparent tension between accuracy and preferences has been studied in the psychology community [Hart et al. 2009], and more work is needed on this *veracity-validation tradeoff* in retrieval scenarios. Search engines need to consider the extent to which they should focus on satisfying searcher needs (validation) versus providing them with the correct answer (veracity). This determination could be made globally (e.g., always prefer result accuracy), based on the subject matter (e.g., favor correctness for consequential topics such as healthcare or those where there is a known and accepted truth, but for more controversial topics such as politics or religion prefer satisfaction since there is no clear truth), or based on user models reflecting individual user preferences (e.g., searchers may only want their beliefs to be validated, irrespective of factual correctness). Although a long history of research that has shown that people strongly prefer content that validates their beliefs [Tversky and Kahneman 1973], there may also be value to

searchers in surfacing content that challenges their beliefs [Yom-Tov et al. 2014]. This may be most appropriate for controversial topics with no defined correct answer.

Despite our promising findings, there are some limitations of this study. We only used a fairly small number of Cochrane reviews (63 or 201, depending on the analysis), although the total number of query instances was high. We focused only on health queries, because this is important for many searchers [Fox and Duggan 2013] and we had ground truth for this domain, allowing us to measure the accuracy and the skew associated with the search results in more detail. However, before we can make claims about how these results could generalize to search engines more broadly, we need to investigate biases in other domains beyond health. Finally, we only used one search engine in our analysis (Microsoft Bing). Search engines crawl and index different parts of the Web, so our lens on bias is limited to the perspective of the particular search engine that we choose. Multi-engine analysis of biases is needed, as has been a central (even definitional) aspect of other studies of bias in search engines [Bharat and Broder 1998; Mowshowitz and Kawaguchi 2002a; 2002b]. Such analysis could be valuable to understand if and how different search engines are affected by content biases, with a view to addressing these biases by, for example, intelligently combining the results returned by different search engines.

Rather than relying on access to results from multiple engines to address possible content biases, which might be impractical for a number of reasons (including inter-engine coordination costs), search engines should consider a number of actions:

— **Define content bias in their setting:** Identify queries or query classes where content biases are important, obtain ground truth (known or accepted) from experts or by mining resources such as the medical literature, and label results accordingly.
— **Monitor content biases longitudinally:** Periodically perform analyses similar to that reported in this article to quantify content biases, and their potential effects on search behavior and search outcomes. If a large index bias is detected, search providers could proactively seek new sources, commission the creation of missing answer pages, or modify crawling algorithms to increase resource diversity.
— **Perform smarter crawling and indexing:** Consider base rates (or other attainable distributions to estimate reality), content distributions, and result quality during crawling and indexing (the latter per [Tang et al. 2005]). This could improve result accuracy by reducing the strong positive skew in both the results and in the indexed content (as observed in Figure 4);
— **Correct skewed queries:** Apply query alterations to replace problematic query terms with synonyms. This could be framed as an interesting learning problem by generating query features, including the presence or absence of particular terms, and optimizing for result accuracy and/or the degree of skew in retrieved results;
— **Reduce bias in implicit feedback:** Improve the quality of implicit feedback by learning from de-biased (or at least less biased) clicks by, say, downsampling based on rank/quantity. Search engines could use this as one of many veracity-oriented features to rank results more accurately and hence reduce content bias, and;
— **Consider user preferences:** Provide a means by which searchers can specify their preferences for information that supports their beliefs, versus information that is factually accurate (or at least commonly-accepted). Searchers could be given explicit control or their preferences could be estimated from search behavior, e.g., if they always seek documents expressing a particular view, skipping conflicting views.
— **Provide balanced perspectives for queries without known truth:** In this article we focus on question queries for which there is a known truth. However, there are a range of issues, such as those on morally- or politically-charged topics,

where there is no defined truth. For such question queries, search engines have an opportunity to challenge searchers' existing beliefs and promote civil discourse. For example, in response to questions with no defined truth, search engines could construct result lists that present a balanced set of perspectives from many sources.

— **Educate searchers about the potential for biases:** In addition to taking action to mitigate bias in search results, systems could alert or educate searchers about the risks involved when searching on consequential or controversial topics. A summary of the viewpoints expressed in the results could be presented on the SERP, accompanied with associated warnings if the results are biased. This support could be offered for all queries on a topic or only for queries likely to generate biased result lists (e.g. those containing "can" or "help", as highlighted in the previous section).

Before adopting these recommendations, it is important to consider the role of search engines in mitigating content biases, validating searchers' beliefs versus challenging them, and in broadening searchers' perspectives. The result sets currently offered by search engines are mostly influenced by the content available on the Web and objective measures of topical relevance. To our knowledge, no special attention is given to the veracity of the results retrieved, the perspectives that those results present, or the potential impact of those results on searcher beliefs. Before search engines take direct action to address biases, then detailed discussion is needed among the stakeholders including searchers, search providers, site owners, privacy advocates, and legislators, about the appropriateness and implications of attempts to intentionally manipulate search results; even if it is only to improve accuracy, as we argue for in this article. Our research underscores the need for the discussion by highlighting significant biases in results and their potential (negative) impact on decisions that searchers must make.

There are some interesting areas of future research related specifically to this study. We need to better understand the source of biases and how we could reliably measure them across different domains. There is also the important issue of commercial interest, which may affect the amount of content generated concerning particular interventions. Those interventions with greatest commercial interest may be associated with the greatest degree of positive (*helps*-only) content and appear more in the search results. Brand biases in search have been studied [Mowshowitz and Kawaguchi 2002b] but more research is needed to understand the role of level of commercial interest generally in content biases. In addition, although we focused on health search, there are a range of topics where people gather evidence to inform consequential decisions and there is often a truth or correct course of action (e.g., home or auto repair), where bias needs to be studied. The challenge in studying bias in these domains lies in attaining the ground truth data from which to measure biases, especially in terms of how we define it (i.e., as a deviation from the truth) when such knowledge may be tacit and undocumented in some cases. Moving forward, to drive research in this area the community needs test collections (comprising queries, ground truth (answers), and content labeled with answers), and experimental methodologies such as ours, through which researchers can build and study different bias measurement and mitigation methods.

6. CONCLUSIONS

Search engines are a trusted resource for a variety of search intents, including health search. We described a study of content bias in online health search, where we define bias in terms of deviations from the truth that adversely affect the accuracy of search results. Content biases are important when people pursue consequential information seeking tasks, such as those with a medical focus. Our study employed search log analysis using data from the Microsoft Bing search engine and the results from that

engine. To measure bias we used ground truth from authoritative Cochrane reviews and downsampled the reviews to create a uniform outcome distribution. This allowed us to measure deviations from the expected (true) distribution directly. We showed that search results are biased toward positive outcomes and that this bias relates to a number of factors, including skewed content in the engine index and content matching performed by the engine when answering search requests. We also showed that by controlling for various biases in search-result examination behavior (through downsampling clicks based on rank position and answers in captions) there is potential to improve the accuracy of behavioral signals. This has implications for the use of implicit feedback in ranking, which is now extremely common. More research is needed to understand the practical value of using this de-biased (or at least less biased) behavioral signal as a feature in applications of machine learning to result ranking. In addition, we observed large differences in search engine performance as a function of different formulations of the same intent. We showed that particular query terms contributed to this variance, and that by performing selective term substitution bias in the top-ranked search results could be reduced. Based on our findings, we also presented a set of actions that engines could take to mitigate and monitor content bias. Future work involves pursuing such directions in the health domain and beyond.

ACKNOWLEDGMENTS

The authors thank Susan Dumais and Eric Horvitz for feedback and discussions.

REFERENCES

Steven Abney, Michael Collins, and Amit Singhal. 2000. Answer extraction. *Proc. ANLC*, 296–301.

Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving web search ranking by incorporating user behavior information. *Proc. SIGIR*, 19–26.

Yin Aphinyanaphongs and Constantin Aliferis. 2007. Text categorization models for identifying unproven cancer treatments on the web. *Proc. 12th World Congress on Health (Medical) Informatics*, 968.

Rawia Awadallah, Maya Ramanath, and Gerhard Weikum. 2012. Harmony and dissonance: organizing the people's voices on political controversies. *Proc. WSDM*, 523–532.

Stephanie L. Ayers and Jennie J. Kronenfeld. 2007. Chronic illness and health-seeking information on the Internet. *Health*, 11(3): 327–347.

Leif Azzopardi and Ciaran Owens. 2009. Search engine predilection toward news media providers. *Proc. SIGIR*, 774–775.

Leif Azzopardi and Vishwa Vinay. 2008. Retrievability: An evaluation measure for higher order information access tasks. *Proc. CIKM*, 561–570.

Doug Beeferman and Adam Berger. 2000. Agglomerative clustering of a search engine query log. *Proc. SIGKDD*, 407–416.

Nicholas J. Belkin, Colleen Cool, W. Bruce Croft, and Jamie P. Callan. 1993. The effect of multiple query representations on information retrieval system performance. *Proc. SIGIR*, 339–346.

Mike Bengeri and Pierre Pluye. 2003. Shortcomings of health-related information on the internet. *Health Promotion International*, 18(4): 381–387.

Michael Bernstein, Jaime Teevan, Susan Dumais, Daniel Liebling, and Eric Horvitz. 2012. Direct answers for search queries in the long tail. *Proc. SIGCHI*, 237–246.

Krishna Bharat and Andrei Broder. 1998. A technique for measuring the relative size and overlap of public web search engines. *Computer Networks and ISDN Systems*, 30(1): 379–388.

Suresh K. Bhavnani. 2002. Domain-specific search strategies for the effective retrieval of healthcare and shopping information. *Proc. SIGCHI*, 610–611.

Suresh K. Bhavnani, Renju T. Jacob, Jennifer Nardine, and Frederick A. Peck. 2003. Exploring the distribution of online healthcare information. *Proc. SIGCHI*, 816–817.

Mikhail Bilenko and Ryen W. White. 2008. Mining the search trails of the surfing crowds: Identifying relevant websites from user activity. *Proc. WWW*, 51–60.

Oliver Bodenreider. 2004. *The Unified Medical Language System (UMLS)*. Oxford University Press.

Robert L. Brennan and Dale J. Prediger. 1981. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41(3): 687–699.

Sergei Brin and Larry Page. 1998. The anatomy of a large-scale hypertextual search engine. *Proc. WWW*.

Olivier Chapelle, Thorsten Joachims, Filip Radlinski, and Yisong Yue. 2012. Large scale validation and analysis of interleaved search evaluation. *ACM Transactions on Information Systems*, 30(1).

Lydia Chilton and Jaime Teevan. 2011. Addressing people's information needs directly in a web search result page. *Proc. WWW*, 27–36.
Junghoo Cho and Sourashis Roy. 2004. Impact of search engines on page popularity. *Proc. WWW*, 20–29.
Charles Clarke, Eugene Agichtein, Susan Dumais, and Ryen W. White. 2007. The influence of caption features on clickthrough patterns in Web search. *Proc. SIGIR*, 135–142.
Rebecca J.W. Cline and Katie M. Haynes. 2001. Consumer health information seeking on the internet: The state of the art. *Health Education Research*, 16(6): 671–692.
James Cogley, Nicola Stokes, and Joe Carthy. 2013. Exploring the effectiveness of medical entity recognition for clinical information retrieval. *Proceedings of the 7th International Workshop on Data and Text Mining in Biomedical Informatics*, 3–4.
Kevyn Collins-Thompson, Jamie Callan, Egidio Terra, and Charles L.A. Clarke. 2004. The effect of document retrieval quality on factoid question answering performance. *Proc. SIGIR*, 574–575.
Nick Craswell, David Hawking, Ross Wilkinson, and Mingfang Wu. 2003. Overview of the TREC 2003 Web Track. *Proc. TREC*.
Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. *Proc. WSDM*, 87–94.
Anirban Dasgupta, Arpita Gosh, Ravi Kumar, Christopher Olston, Sandeep Pandey, and Andrew Tomkins. 2007. The discoverability of the web. *Proc. WWW*, 421–430.
Kay Dickersin. 1990. The existence of publication bias and risk factors for its occurrence. *Journal of the American Medical Association*, 263(10): 1385–1389.
Shiri Dori-Hacohen and James Allan. 2013. Detecting controversy on the web. *Proc. CIKM*, 1845–1848.
Susan Dumais, Michele Banko, Eric Brill, Jimmy Lin, and Andrew Ng. 2002. Web question answering: Is more always better? *Proc. SIGIR*, 291–298.
Phillipa J. Easterbrook, R. Gopalan, J.A. Berlin, and David R. Matthews. 1991. Publication bias in clinical research. *The Lancet*, 337(8746): 867–872.
Matthias Egger, George Davey Smith, Martin Schneider, and Christoph Minder. 1997. Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315(7109): 629–634.
Rob Ennals, Beth Trushkowsky, and John M. Agosta. 2010. Highlighting disputed claims on the web. *Proc. WWW*, 341–350.
Gunther Eysenbach and Christian Köhler. 2002. How do consumers search for and appraise health information on the World Wide Web? Qualitative studies using focus groups, usability test, and in-depth interviews. *British Medical Journal*, 324: 573–577
Daniele Fanelli. 2012. Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3): 891–901.
Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5): 378–382.
Santo Fortunato, Alessandro Flammini, Fillipo Menczer, and Alessandro Vespignani. 2006. Topical interests and the mitigation of search engine bias. *Proc. of the National Academy of Sciences*, 103(34): 12684–12689.
Susannah Fox. 2006. *Online Health Search 2006*. Pew Internet and American Life Project. Accessed January 2014.
Susannah Fox and Maeve Duggan. 2013. *Health Topics*. Pew Internet and American Life Project. Accessed January 2014.
Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating implicit measures to improve Web search. *ACM Transactions on Information Systems*, 23(2): 147–168.
Arnaud Gaudinat, Patrick Ruch, Michel Joubert, Phillipe Uziel, P., Anne Strauss, Michèle Thonnet, Robert Baud, Stéphane Spahni, Patrick Weber, Juan Bonai, Celia Boyer, Marius Fieschi, and Antoine Geissbuhler. 2006. Health search engine with e-document analysis for reliable search results. *International Journal of Medical Informatics*. 75(1): 73–85.
Susan Gerhart. 2004. Do Web search engines suppress controversy? *First Monday*, 9: 1–5.
Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinksi, and Larry Brilliant. 2008. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232): 1012–1014.
Eric Goldman. 2006. Search engine bias and the demise of search utopianism. *Yale Journal of Law and* Technology, 188.
Qi Guo and Eugene Agichtein. 2012. Beyond dwell time: Estimating document relevance from cursor movements and other post-click searcher behavior. *Proc. WWW*, 569–578.
Donna Harman and Chris Buckley. 2009. Overview of the reliable information access workshop. *Information Retrieval*, 12: 614–641.
William Hart, Dolores Albarracín, Alice H. Eagly, Inge Brechan, Matthew J. Lindberg, and Lisa Merrill. 2009. Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, 135(4): 555–588.
William Hersh. 2008. *Information Retrieval: A Health and Biomedical Perspective*. Springer.

William R. Hersh and David H. Hickam. 1998. How well do physicians use electronic information retrieval systems? A frame work for investigation and systematic review. *Journal of the American Medical Informatics Association*, 280: 1347.

William R. Hersh, Katherine Crabtree, David H. Hickam, Lynetta Sacherek, Charles P. Friedman, Patricia Tidmarsh, Craig Mosbaek, and Dale Kraemer. 2002. Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions. *Journal of the American Medical Informatics Association*, 9: 283–293.

Julian P.T. Higgins (Ed.). 2008. *Cochrane Handbook for Systematic Reviews of Interventions* (Vol. 5). Chichester: Wiley-Blackwell.

Samuel Ieong, Nina Mishra, Eldar Sadikov, and Li Zhang. 2012. Domain bias in Web search. *Proc. WSDM*, 413–422.

Peter Ingwersen. 1994. Polyrepresentation of information needs and semantic entities: Elements of a cognitive theory of information retrieval interaction. *Proc. SIGIR*, 101–110.

Charles B. Inlander. 1993. *Good operations, Bad operations*. The People's Medical Society's Guide to Surgery. Viking Adult.

Alejandro R. Jadad, Deborah J. Cook, Alison Jones, Terry P. Klassen, Peter Tugwell, Michael Moher, and David Moher. 1998. Methodology and reports of systematic reviews and meta-analyses: a comparison of Cochrane reviews with articles published in paper-based journals. *Journal of the American Medical Association*, 280(3): 278–280.

Thorsten Joachims. 2002. Optimizing search engines using click-through data. *Proc. SIGKDD*, 132–142.

Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems*, 25(2).

Anders W. Jørgensen, Jørgen Hilden, and Peter C. Gøtzsche. 2006. Cochrane reviews compared with industry supported meta-analyses and other meta-analyses of the same drugs: systematic review. *British Medical Journal*, 333(7572), 782.

Mouna Kacimi and Johann Gamper. 2011. Diversifying search results of controversial queries. *Proc. CIKM*, 93–98.

Mouna Kacimi and Johann Gamper. 2012. MOUNA: mining opinions to unveil neglected arguments. *Proc. CIKM*, 2722–2724.

Diane Kelly, Xiao-jun Yuan, Nicholas J. Belkin, Vanessa Murdock, and W. Bruce Croft. 2002. Features of documents relevant to task- and fact-oriented questions. *Proc. CIKM*, 645–647.

Jon Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5): 604–632.

Bevan Koopman, Guido Zuccon, Peter Bruza, Laurianne Sitbon, and Michael Lawley. 2012. An evaluation of corpus-driven measures of medical concept similarity for information retrieval. *Proc. CIKM*, 2439–2442.

Carolyn Lauckner and Gary Hsieh. 2013. The presentation of health-related search results and its impact on negative emotional outcomes. *Proc. SIGCHI*, 333–342.

Hady W. Lauw, Ee-Peng Lim, and Ke Wang. 2006. Bias and controversy: Beyond the statistical deviation. *Proc. SIGKDD*, 625–630.

Steve Lawrence and C. Lee Giles. 1999. Accessibility of information on the web. *Nature*, 400(6740): 107.

Nut Limsopatham, Craig Macdonald, Richard McCreadie, and Iadh Ounis. 2012. Exploiting term dependence while handling negation in medical search. *Proc. SIGIR*, 1065–1066.

Abbe Mowshowitz. and Akira Kawaguchi. 2002a. Assessing bias in search engines. *Information Processing and Management*, 38(1): 141–156.

Abbe Mowshowitz and Akira Kawaguchi. 2002b. Bias on the Web. *CACM*, 45(9): 56–60.

Yishai Ofran, Ora Paltiel, Dan Pelleg, Jacob M. Rowe, and Elad Yom-Tov. 2012. Patterns of information-seeking for cancer on the internet: An analysis of real world data. *PloS One*, 7(9): e45921.

Bo Pang and Ravi Kumar. 2011. Search in the lost sense of "query": Question formulation in Web search queries and its temporal changes. *Proc. ACL*, 135–140.

Sandeep Pandey, Kedar Dhamdhere, and Christopher Olston. 2004. Wic: A general-purpose algorithm for monitoring web information sources. *Proc. VLDB*, 360–371.

Eli Pariser. 2011. *The Filter Bubble: What is the Internet Hiding from You?* Penguin Press.

Mark Petticrew, Paul Wilson, Kath Wright, and Fujian Song. 2002. Quality of Cochrane reviews: Quality of Cochrane reviews is better than that of non-Cochrane reviews. *British Medical Journal*, 324(7336): 545.

Mary C. Politi, Paul K.J. Han, and Nananda F. Col. 2007. Communicating the uncertainty of harms and benefits of medical interventions. *Medical Decision Making*, 27(5): 681–695.

Filip Radlinski and Thorsten Joachims. 2006. Minimally invasive randomization for collecting unbiased preferences from click-through logs. *Proc. AAAI*.

Matthew Richardson, Amit Prakash, and Eric Brill. 2006. Beyond PageRank: Machine learning for static ranking. *Proc. WWW*, 707–715.

Bonnie Rochman. 2011. Jenny McCarthy, vaccine expert? A quarter of parents trust celebrities. *Time*. Published 26 April 2011. Retrieved 6 May 2013.
Ian Ruthven. 2003. Re-examining the potential effectiveness of interactive query expansion. *Proc. SIGIR*, 213–220.
David L. Sackett, William Rosenberg, J. A. Gray, R. Brian Haynes, and W. Scott Richardson. 1996. Evidence based medicine: What it is and what it isn't. *British Medical Journal*, 312(7023): 71.
Julia Schwarz and Meredith Ringel Morris. 2011. Augmenting web pages and search results to help people find trustworthy information online. *Proc. SIGCHI*, 1245–1254.
Elizabeth Sillence, Pam Briggs, Lesley Fishwick, and Peter Harris. 2004. Trust and mistrust of online health sites. *Proc. SIGCHI*, 663–670.
Herbert A. Simon. 1991. Bounded rationality and organizational learning. *Organization Science*, 2(1): 125–134.
Amit Singhal, Chris Buckley and Mandar Mitra. 2006. Pivoted document length normalization. *Proc. SIGIR*, 21–29.
Karen Spärck-Jones, Steve Walker, and Stephen E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments. *Information Processing and Management*, 36(6): 779–840.
Amanda Spink, Yin Yang, Jim Jansen, Pirrko Nykanen, Daniel P. Lorence, Seda Ozmutlu, and H. Cenk Ozmutlu. 2004. A study of medical and health queries to Web search engines. *Health Information and Libraries Journal*, 21: 44–51.
Thanh Tin Tang, David Hawking, Nick Craswell, and Kathy Griffiths. 2005. Focused crawling for both topical relevance and quality of medical information. *Proc. CIKM*, 147–154.
Robert S. Taylor. 1968. Question-negotiation and information seeking in libraries. *College and Research Libraries*, 29: 178–194.
Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(1): 207–233.
Tristan Upstill, Nick Craswell, and David Hawking. 2002. Buying bestsellers online: A case study in search and searchability. *Proc. 7th Australasian Document Computing Symposium*.
Liwen Vaughn and Mike Thelwall. 2004. Search engine coverage bias: Evidence and possible causes. *Information Processing and Management*, 40(4): 693–707.
V.G. Vydiswaran, Cheng Xiang Zhai, Dan Roth, and Peter Pirolli. 2012. BiasTrust: Teaching biased users about controversial topics. *Proc. CIKM*, 1905–1909.
Peter C. Wason. 1960. On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12: 129–140.
Robert West, Ryen W. White, and Eric Horvitz. 2013. From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. *Proc. WWW*, 1399–1410.
Ryen W. White. 2013. Beliefs and biases in Web search. *Proc. SIGIR*, 3–10.
Ryen W. White, Rave Harpaz, Nigam H. Shah, William DuMouchel, and Eric Horvitz. 2014. Toward enhanced pharmacovigilance using patient-generated data on the Internet. *Nature Clinical Pharmacology and Therapeutics*, in press.
Ryen W. White and Eric Horvitz. 2009a. Cyberchondria: Studies of the escalation of medical concerns in web search. *ACM Transactions on Information Systems*, 27(4): 23.
Ryen W. White and Eric Horvitz. 2009b. Experiences with web search on medical concerns and self-diagnosis. *Proc. AMIA*, 696–700.
Ryen W. White and Eric Horvitz. 2010a. Web to World: Predicting transitions from self-diagnosis to the pursuit of local medical assistance in web search. *Proc. AMIA*, 882–886.
Ryen W. White and Eric Horvitz. 2010b. Predicting escalations of medical queries based on web page structure and content. *Proc. SIGIR*, 769–770.
Ryen W. White, Nicholas P. Tatonetti, Nigam H. Shah, Russ B. Altman, and Eric Horvitz. 2013. Web-scale pharmacovigilance: Listening to signals from the crowd. *Journal of the American Medical Informatics Association*, 20(3): 404–408.
Barbara M. Wildemuth. 2004. The effects of domain knowledge on search tactic information. *JASIST*, 55(3): 246–258.
Colin Wilkie and Leif Azzopardi. 2014. Best and fairest: An empirical analysis of retrieval system bias. *Proc. ECIR*, 13–25.
Janice C. Wright and Milton C. Weinstein. 1998. Gains in life expectancy from medical interventions: Standardizing data on outcomes. *New England Journal of Medicine*, 339(6): 380–386.
Elad Yom-Tov, Susan Dumais, and Qi Guo. 2014. Promoting civil discourse through search engine diversity. *Social Science Computer Review*, 32(2): 145–154.
Elad Yom-Tov and Evgeniy Gabrilovich. 2013. Post-market drug surveillance without trial costs: Discovery of adverse drug reactions through large-scale analysis of web search queries. *Journal of Medical Internet Research*, 15(6).
Yisong Yue, Rajan Patel, and Hein Roehrig. 2010. Beyond position bias: Examining result attractiveness

as a source of presentation bias in clickthrough data. *Proc. WWW*, 1011–1018.
Cheng Xiang Zhai, William W. Cohen, and John Lafferty. 2003. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. *Proc. SIGIR*, 10–17.