Captions and Biases in Diagnostic Search

RYEN W. WHITE, Microsoft Research
ERIC HORVITZ, Microsoft Research

People frequently turn to the Web with the goal of diagnosing medical symptoms. Studies have shown that diagnostic search can often lead to anxiety about the possibility that symptoms are explained by the presence of rare, serious medical disorders, rather than far more common benign syndromes. We study the influence of the appearance of potentially-alarming content, such as severe illnesses or serious treatment options associated with the queried for symptoms, in captions comprising titles, snippets, and URLs. We explore whether users are drawn to results with potentially-alarming caption content, and if so, the implications of such attraction for the design of search engines. We specifically study the influence of the content of search result captions shown in response to symptom searches on search-result click-through behavior. We show that users are significantly more likely to examine and click on captions containing potentially-alarming medical terminology such as "heart attack" or "medical emergency" independent of result rank position and well-known positional biases in users' search examination behaviors. The findings provide insights about the possible effects of displaying implicit correlates of searchers' goals in search-result captions, such as unexpressed concerns and fears. As an illustration of the potential utility of these results, we developed and evaluated an enhanced click prediction model that incorporates potentially-alarming caption features and show that it significantly outperforms models that ignore caption content. Beyond providing additional understanding of the effects of Web content on medical concerns, the methods and findings have implications for search engine design. As part of our discussion on the implications of this research, we propose procedures for generating more representative captions that may be less likely to cause alarm, as well as methods for learning to more appropriately rank search results from logged search behavior, e.g., by also considering the presence of potentially-alarming content in the captions that motivate observed clicks and down-weighting clicks seemingly driven by searchers' health anxieties.

Categories and Subject Descriptors: **H.3.3 [Information Storage and Retrieval]**: Information Search and Retrieval.

General Terms: Experimentation, Human Factors

Additional Keywords and Phrases: Captions; Biases; Diagnostic search; Cyberchondria

1. INTRODUCTION

People frequently turn to the Web to find information about their medical concerns. A recent study found that 80% of U.S. Web users have performed online medical searches [Fox 2011]. Diagnostic search, where people query about the potential causes of symptoms that they notice, is a popular type of health search task. Another recent study found that 35% of U.S. adults had used the Web to perform diagnosis of medical conditions either for themselves or on behalf of another person [Fox & Duggan 2013]. Symptoms occur in as many as 40% of the medical queries that search engines receive [White & Horvitz 2012]. The view that search engines provide on medical content can affect searchers' beliefs and behaviors around medical matters, including decisions involving diagnosis and treatment. In addition, approximately 25% of Web searchers have reported interpreting the ranked ordering of search results returned in symptom searches as an ordering of diseases by occurrence likelihood [White & Horvitz 2009a]. However, search engine ranking algorithms can exhibit biases in the information that they cover [Gerhart, 2004; Vaughan & Thelwall 2004; Goldman, 2006] and how they choose to order their results [Mowshowitz & Kawaguchi, 2002a, 2002b], have limited access to information about a searcher's situation and background probabilities on conditions, and the trust that people place in search engine rankings can lead to erroneous beliefs and negative emotional outcomes [Lauckner & Hsieh 2013].

Beyond ranking, the presentation of results on search engine result pages (SERPs) has been studied to understand what aspects of result captions motivate users to select particular results [Clarke et al. 2007; Yue et al. 2010]. In diagnostic search, decisions about what content to view can have direct implications on the

Fig. 1. Top three search result captions for [chest pain].
Potentially-alarming caption content is highlighted.

wellbeing of searchers and influence decisions about self-treatment and healthcare utilization [Ayers & Kronenfeld 2007]. Figure 1 shows the top three result captions from the Microsoft Bing search engine for query [chest pain]. The snippet content in two of the top three captions shown on the SERP (at rank positions one and three) contain *potentially-alarming content*, which may lead to heightened concern and focus from searchers. The first caption describes the severity of conditions associated with chest pain and suggests that emergency treatment should be sought. The third caption includes multiple serious disorders linked to chest pain, all of which are pretty rare. In addition, diagnostic searchers may be in a heightened state of anxiety and therefore more attracted and receptive to concerning content [Asmundson, Taylor & Cox 2001]. We hypothesize that captions with potentially-concerning or potentially-alarming content, such as the mention of serious ailments or severe treatment options, can draw people's focus of attention to particular search results, independent of rank position or result relevance. Results with attractive captions can create feedback loops, where there associated search results are clicked on frequently (regardless of relevance) and hence ranked most highly by the search engine for future queries [Cho & Roy, 2004; Yue et al. 2010].

Selection choices may be influenced by multiple aspects of human judgment and decision making, including biases long studied in cognitive psychology, such as *base-rate neglect*, *availability bias,* and *confirmation bias* [Tversky & Kahneman 1974]. Such cognitive biases have been highlighted as playing a significant role in the unfounded escalation of concerns that common symptomology is caused by rare, serious illnesses, following searchers' review of search results and online literature [White & Horvitz 2009b]. Previous work in this area has shown that symptoms and escalatory terminology occur in high-ranked search results returned on symptom searches more frequently than expected given prevalence of disorders that users often quickly transition from symptoms to related serious conditions, and have explored the relationship between Web page content and structure on successive queries [White & Horvitz 2010]. In this article, we extend our prior research to focus on understanding *the relationship between potentially-alarming content in search-result captions and searcher engagement with the search results*, primarily search-result clicks but also cursor hovers on the search results (as an additional proxy for searcher attention). In distinction to our prior work on the influence of multiple aspects of viewed content on reformulations and Web page access [White & Horvitz 2010], the study of captions provides opportunities for characterizing searchers' reactions to specific displayed

content with a crisp metric of search-result click-through. That is, we can observe directly if and when searchers are drawn to potentially-alarming content in the search results, allowing us to better understand and tailor search support to address aspects of searchers' unexpressed concerns and fears during diagnostic search.

We perform a controlled study of the effect of potentially-alarming caption content in diagnostic search. As we discussed earlier, diagnostic search is common activity, with profound implications for the health and wellbeing of Web searchers as well as associated anxieties and concerns that make it fertile ground for potentially-alarming content to influence selection choices. Our main contributions are:

- We demonstrate comprehensively that the presence of terminology in the captions associated with alarming outcomes attracts greater attention and click-through on search results. For example, the mention of serious conditions such as "cancer" or "tumor" in captions leads to more lengthy examination of those captions and significant increases in click-through on the results associated with them.
- We show that we can more accurately model users' diagnostic search behavior by considering the effect of potentially-alarming caption content. Including features of that content and their importance weights in click-through predictions leads to significant gains in prediction accuracy. To our knowledge this is the first demonstration of the benefit of adding features of caption content associated with likely emotive responses to Web-search click prediction (focused on potentially-alarming content in our case), rather than just syntactic features of the captions, which have been used with some success previously [Kang et al. 2011].
- We demonstrate that search results for diagnostic queries with captions containing serious conditions are typically ranked near the top of the search result list. In previous work we studied this issue broadly in terms co-occurrences in the content top-100 results [White & Horvitz 2009a]. In this paper we are focus on the top-most results (at rank positions 1-10, which users are most likely to see), on the captions of those results in addition to their content (which users engage with directly), and illustrate the presence of a significant mismatch between captions and content which allows us to attribute much of the promotion of these results to signals learned from aggregated caption click-through behavior.
- We discuss the implications of this research for the design of search systems. Given their widespread adoption as a primary mechanism through which people discover and learn about health issues, we consider the argument that search engines have a duty of care (as has been suggested in related research on search personalization [Pariser 2011]) and should surface information to their users responsibly. As such, we suggest that mechanisms are needed for caption generation and bias-sensitive ranking that can consider base rates, as well as downweight SERP clicks motivated by effects such as health anxiety, since those may influence learned rankings.

The remainder of this article is structured as follows. Section 2 describes relevant research in the areas of health information seeking, result examination behavior, caption attractiveness, and modeling search behavior to predict result click-through behavior. Section 3 describes initial analysis of caption content and how users examine result captions with different content; specifically, potentially-reassuring benign conditions and potentially-alarming serious illnesses. We then study the attractiveness of captions associated with each of these content types in Section 4. Section 5 incorporates caption attractiveness features into click prediction models resulting in more accurate models of search behavior. We discuss our findings and their implications in Section 6, and conclude in Section 7.

2. RELATED WORK

Research in a number of areas is relevant to the work presented here: (1) prior work on health search and health anxiety, (2) studies of SERP examination behavior, (3) research on snippet attractiveness, and (4) the development of click prediction models.

2.1 Health Search and Health Anxiety

Prior studies have explored the effect of Web content on health search behavior. Bhavnani and colleagues [Bhavnani et al. 2003] demonstrated that Webpage term co-occurrence of medical symptoms and disorders can reasonably predict the degree of influence on search behavior. Spink and colleagues [Spink et al. 2004] characterized healthcare-related queries issued to 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. Eysenbach and Kohler [Eysenbach & Kohler 2002] reviewed several studies and concluded that health-related Web content is often of poor quality. Other research has shown that people do not sufficiently consider the reliability of the source of the health-related content they examine [Sillence et al. 2007]. White and Horvitz [White & Horvitz 2009a] used a log-based methodology to study escalations in medical concerns during Web search (shifts from search on common symptoms to searches on serious ailments). Their work highlighted the potential influence of several biases of judgment demonstrated by people and by search engines themselves, including base-rate neglect and availability. Cartright and colleagues [Cartright, White & Horvitz 2011] studied differences in search behavior associated with diagnosis vs. more general health-related information seeking. They decomposed health information seeking into evidence-based (search on relevance of signs and symptoms) and hypothesis-based (search on conditions and treatments), and studied how medical foci evolve during exploratory health search.

The medical community has studied the effects of health anxiety disorders, including hypochondriasis, over time [Asmundson, Taylor & Cox 2001], but not in the context of Web search. Health anxiety is problematic as it is often maladaptive (i.e., out of proportion with the degree of medical risk) and amplified by the fact that those affected are often undiscerning about the source of their medical information [Taylor & Asmundson 2004; Kring et al. 2007]. Such anxiety usually persists even after an evaluation by a physician and reassurance that concerns about symptoms lack an underlying serious health basis. Beyond interactions with medical professionals, patients' health concerns may manifest in other ways such as search behavior, an assertion supported by [White & Horvitz 2009b], which showed that those who self-identified as hypochondriacs searched the Web more often for health information than average Web searchers.

2.2 SERP Examination Behavior

SERP clicks provide a signal that users are attracted to a search result. Joachims and colleagues [Joachims et al. 2007] analyzed users' decision processes via gaze tracking and compared implicit feedback from search-result clicks against manual relevance judgments. They found that clicks are informative but biased, yet relative result preferences derived from clicks mirror searchers' true preferences. Agichtein and colleagues [Agichtein et al. 2006] used search and browsing data from a Web search engine to predict search result preferences. They also generalize their approach to model post-query browsing behavior, resulting in more accurate predictions about preferences. Cutrell and Guan [Cutrell & Guan 2007] manipulated the quantity of information displayed in snippets and found that adding information significantly improved searcher performance for informational tasks but degraded performance for

navigational tasks. Buscher and colleagues [Buscher et al. 2010] systematically varied the type of search task (informational or navigational), the quality of online advertisements (relevant or irrelevant to the query), and the sequence in which advertisements of different quality were presented on the SERP. They demonstrated that the amount of visual attention that people devote to search results depends on both task type and advertisement quality. The amount of visual attention that people devote to advertisements depends on their quality, but not the type of task. Guo and Agichtein [Guo & Agichtein 2010] used interactions such as cursor movement, hovers, and scrolling to accurately infer search intent and interest in search results. Huang and colleagues [Huang, White & Dumais 2011] developed a scalable version of cursor tracking that was deployed in the Bing search engine. They showed that many of the gaze patterns observed in the gaze tracking studies above were also observed in an analysis of cursor behavior on the SERP.

2.3 Caption Attractiveness

Captions help users decide on whether they should click on a particular search result. Studies have investigated the value of query-dependent summarization in a non-Web setting [Tombros & Sanderson 1998]. Snippet attractiveness has been studied via search engine log data. Clarke and colleagues [Clarke et al. 2007] introduced click inversions, a methodology which we employ in this paper, to study features of the captions that increase caption attractiveness. Agichtein and colleagues [Agichtein et al. 2006] examined features of the overlap between the query and different SERP elements: titles, snippets, and URLs. More recent work has sought to leverage caption features to build better models of search behavior. Yue and colleagues [Yue et al. 2010] studied the effect of caption attractiveness, defined for their study as the presence and absence of bolded terms in the titles and snippets of the caption. They show via experiments conducted on the Google Web search engine substantial evidence of presentation bias in clicks towards results with more attractive titles (those with more bolded terms). Later in the article we explore the use of features related to caption content to more accurately model diagnostic search behavior. Rather than how terms are presented, our work focuses on the potential that caption content has to alarm or reassure searchers and the effect that has on search behavior.

Beyond traditional click prediction in Web search, Shaparenko and colleagues [Shaparenko et al. 2009] leveraged word-pair features between the query and terms appear in advertisements for click prediction in sponsored search. Kang and colleagues [Kang et al. 2011] use a set of snippet features to model the perceived relevance of searchers. They do so for queries without click data and show that their model can effectively predict relevance and improve search performance. Others have used behavior related to captions and other document representations as implicit feedback, and taken content from studied captions to enhance relevance [Murata et al. 2009; White, Ruthven & Jose 2002]. Ieong and colleagues [Ieong et al. 2012] studied the effect of domain biases, whereby a result is believed to be more relevant because of its source domain. They show that this bias exists in click behaviors as well as human judgments, show that domain can flip user preference in a caption around a quarter of the time, and as we do here, show that their findings are independent of rank or relevance.

2.4 Click Prediction Models

Searcher models (e.g., [Chapelle & Zhang 2009; Craswell et al. 2008; Wang et al. 2010]) track the user's state as they examine search results and use observed events (e.g., search result click-through) to infer search result attractiveness and document

relevance. The examination hypothesis [Dupret & Liao 2010] states that the likelihood that the user will click on a particular search result is influenced only by (1) whether the user examined the search result snippet, and (2) its attractiveness. Since users would rather select search results that are higher ranked [Joachims et al. 2007], the examination hypothesis is used to isolate a search result's attractiveness from its position. The cascade hypothesis [Craswell et al. 2008] assumes that a user always examines results sequentially from top-to-bottom, and is used to determine whether a user examined the result. Under this assumption, a user decides whether to click a result before examining the next result, overlooking scenarios where the user returns to a higher-ranked search result after skipping over it. If users do not examine a particular search result, it is assumed that they will not examine search results below it. Extensions of the cascade hypothesis allow for query sessions to comprise multiple clicks or represent the probability that a user abandons a query session without clicking [Chapelle & Zhang 2009; Guo et al. 2009].

2.5 Primary Contributions over Prior Work

This research extends previous work in a number of ways. First, although there has been some related work on snippet attractiveness, we focus specifically on diagnostic search to identify how the appearance of potentially-alarming content that evoke fear and anxiety can influence search behavior and information access. Second, we target SERP click-through rather than query reformulation or post-query browsing since it lets us establish a clear link between content and subsequent activity. Third, we show behavioral differences, both in terms of caption examination and result click-through, depending on the presence of potentially-alarming content in captions. Fourth, we show that the presence of this content affects the position of these results in the result ranking of search engines. Fifth, we develop click prediction models that directly incorporate caption features into the computation of search result attractiveness, leading to gains in prediction accuracy. This modeling provides a demonstration of the value in considering such biases extending beyond modifications in caption generation methods. Finally, we discuss the implications of potentially-alarming caption content on search engine design, including de-biasing search engine rankings.

3. EFFECT OF CAPTION CONTENT

We showed in a prior study that the content of the top 100 results from a Web search engine for searches on common symptoms contain a significant overexpression of co-occurrences of symptoms and serious illnesses when prevalence rates are taken into consideration (e.g., 37% of the top-100 results for symptom [chest pain] mentioned "heart attack" even when chances of a heart attack are low given input symptoms) [White & Horvitz 2009a]. That work was focused on the *potential* for escalation (i.e., shifts from search on common symptoms to searches on serious ailments) and as such did not study searcher engagement with the results (i.e., no log analyses of clicks or user studies). Escalations were also identified via keyword spotting methods that looked for transitions from symptoms to associated serious conditions in users' query statements. That research also looked broadly at the top 100 results for symptom queries, which many users will never see since SERP pagination occurs rarely in Web search, and did not consider the effect of rank, which is critical since searchers typically only examine the top captions [Joachims et al. 2007]. It also did not consider the effect of captions on rankings learned from result click-through. We now seek to establish: (1) whether snippets presented by search engines exhibit a bias toward potentially-alarming content, and (2) show how the display of such concerning terminology affects caption examination and search result click-through decisions.

Ultimately, the effects of such content on people's behaviors across many users could influence search engine ranking if the engine learns from those users' click-through behavior [Joachims 2002; Agichtein et al. 2006]. Later we provide evidence suggesting a positive relationship between the presence of potentially-alarming content in the captions of results and the ranking of those results in the search engine result list.

3.1 Data

We used two months of log data from the Microsoft Bing commercial Web search engine, from 17 April 2012 to 11 June 2012. The logs contained queries, clicks, and captions (titles, snippets, URLs) of the top ten search results that were shown to the user for each query. Overall, the logs contained billions of queries from millions of users. Users were identified using a unique cookie-based identifier which was unlikely to change during a query session. To remove variability caused by geographic and linguistic variation in search behavior, we only include entries generated in the English speaking United States locale. Further analysis is needed to understand how different populations of users search. Previous work has suggested a relationship between demographics and search behavior [Weber & Castillo 2010]. Different regions of the world have different attitudes and access to healthcare and patient utilization of healthcare also differs between countries [Anderson & Hussey 2001]. Factors such as these could affect the extent to which people chose to pursue health-related content online.

We also ignored any queries containing complex or non-alphanumeric terms (e.g., operators and phrases) and normalized the remaining queries by lowercasing and trimming whitespace. We filtered logs based on a symptom list from the online version of the Merck medical dictionary (see http://www.merckmanuals.com/professional/full-symptoms.html). Starting with the Merck list, we removed duplicates (e.g., multiple references to the same condition with different cohorts), and split pairs of symptoms into singletons (e.g., "*Nausea and Vomiting in Adults*" and "*Nausea and Vomiting in Infants and Children*" became "nausea" and "vomiting"). The final list contained 60 symptoms, including common symptoms such as "chest pain," "headache," and "twitching." The same list has been used in previous analysis of search behavior to identify medical sessions [Cartright, White & Horvitz 2011; White & Horvitz 2012].

In the log-centric analysis, we also used synonyms of symptoms and conditions to increase coverage (e.g., including "tiredness" in addition to "fatigue"). Synonyms for each symptom or condition were identified via a two-step walk on the search engine click-graph using an approach similar to that described from Beeferman and Berger [Beeferman & Berger 2000]. The automatically generated lists of synonyms were reviewed by the authors to remove erroneous list entries (e.g., all army-related synonyms were removed for the symptom "fatigue"). This procedure resulted in a list of 1,408 symptom-synonym pairs. Example synonyms for the symptom abdominal pain included: "sore stomach," "belly ache," and "pain in abdomen." The correctness of the list was verified by one of the authors (EH) who received an MD/PhD degree, and a separate practicing physician working under contract.

To count as a match on one or more of the concern types, a query needed to be an exact match against the symptoms or a synonym. We avoided substring matches to ensure high precision in the query labeling. The filtering yielded 189,346 symptom queries from 37,642 users. We refer to this set as S.

3.2 Content and Result Ordering

We begin by focusing on the presence and absence of serious illnesses and benign explanations in the top 10 captions and the full text of their corresponding results. To

develop a better sense of the distribution of serious illnesses and benign explanations among the top results, we used keyword spotting to automatically label each SERP caption depending on whether it contained the following:

- **Benign explanations:** List of commonly-occurring conditions, defined in [White & Horvitz 2009a] and used in that study for a log-based analysis of online search behavior. The wordlist comprises a set of conditions selected from across the *International Classification of Diseases 10th Edition* (ICD-10) published by the World Health Organization[1]. We selected a range of conditions that were well known and likely to be observed in our log data. Examples of the conditions chosen include "caffeine withdrawal," "common cold," and "pregnancy."
- **Serious illnesses:** List of serious conditions defined in [White & Horvitz 2009a], again based on the ICD-10. Again these were well-known conditions that we believed were likely to appear in our logs. Examples of serious illnesses selected included "heart failure," "multiple sclerosis," and "hepatitis."

The lists of benign explanations and serious illnesses that were used for our analysis are presented in Table II of [White & Horvitz 2009a]. Note that conditions in these lists were chosen independent of any relationship to the symptoms that we study.

We found on average that 1.34 captions per SERP contained mention of a serious illness (standard deviation, SD=1.11) and 1.53 captions per SERP with a benign explanation (SD=1.29). In total 40% of the SERPs for the symptom queries in *S* contained at least one caption with a serious illness and 52.6% of the SERPs contained at least one caption with a benign explanation. 38.2% of SERPs mentioned both types (in the same caption or otherwise) and 19.9% of SERPs did not mention either. With serious conditions appearing on SERPs for 57% of queries, there is potential for users to be exposed to potentially-alarming content. However, since the information is only shown in one or two captions, the rank position of those captions is likely to be an important determinant of whether they notice these terms [Joachims et al. 2007].

To better understand *where* in the top 10 results the captions of interest appeared, we also computed the distribution of captions with all four combinations of serious illnesses and benign explanations, including neither. Figure 2 (overleaf) depicts the distribution of each combination. The sum for each line across all ten rank positions is 100%. If the conditions are evenly distributed in result captions, the percentage would be stable as a function of rank position (i.e., at 10%). The blue line in the figure suggests that this is the case for snippets that do not contain serious illnesses nor benign explanations. More importantly, the results also show that the captions with serious illnesses are more likely to be located near the top of the list. This phenomenon may be caused by potentially-alarming content-based features, including anchor text in links referencing those pages, as well as behavioral features learned from user engagement with that content on the SERP.

To better understand the extent to which the caption skew in the top ten results is related to content-based features of the results, we studied two sources gleaned from a snapshot of the search index taken at the same time as when the captions were shown: (1) result content (i.e., the full-text content of each result) and (2) anchor text (i.e., the visible, clickable text in hyperlinks pointing to each result). Anchor text has been used extensively for Web search ranking since its utility was demonstrated by Brin and Page [Brin & Page 1998]. We computed the same distributions over the top

[1] http://www.who.int/classifications/icd/en/

ten results, but this time for content and anchor text rather than for the captions, to understand how they relate to the ranking.[2]

[2] Ideally it would be possible to review the relative feature weights used in the Bing ranking algorithm for these queries, but we did not have access to these weights retrospectively.

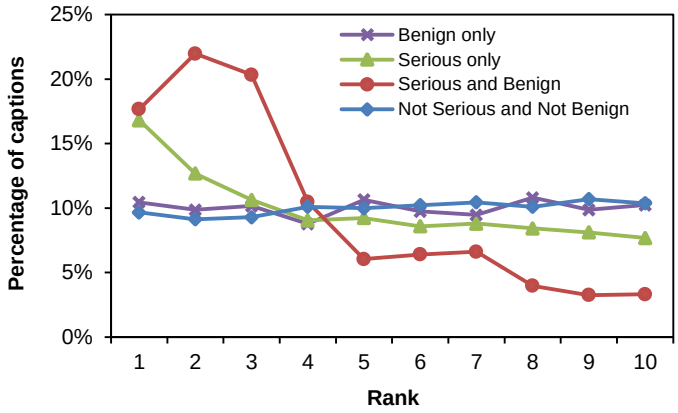

Fig. 2. Distribution of serious illnesses and benign explanations appearing in the *captions* of top ranked search results.

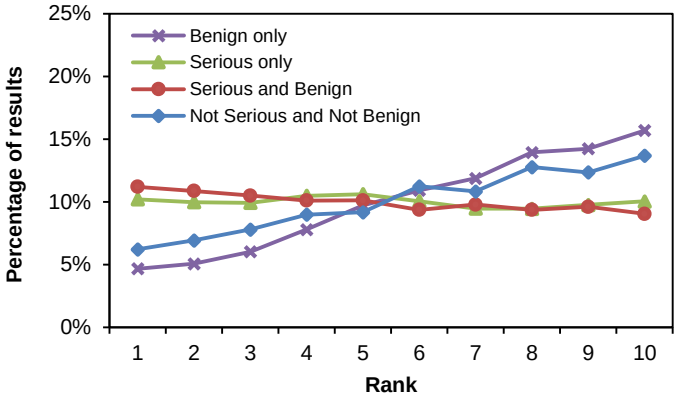

Fig. 3. Distribution of serious illnesses and benign explanations appearing in the *content* of top-ranked search results.

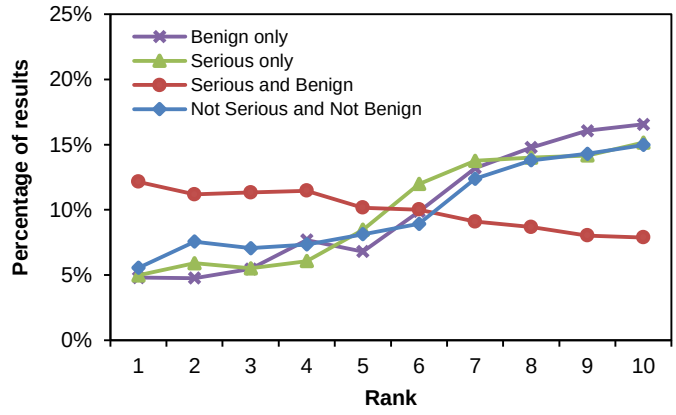

Fig. 4. Distribution of serious illnesses and benign explanations appearing in the *anchor text* of top-ranked search results.

We found that, on average, eight of the top 10 results returned for our symptom queries contained at least one serious condition. The distribution for result content is

Table I. Serious illness presence/absence in caption versus order of occurrence on the landing page.

Order on landing page	Caption has serious illness	
	Yes	No
Serious first	44.3%	19.2%
Benign first	55.7%	80.8%

shown in Figure 3. There are interesting differences compared with the distributions for the captions. First, if landing pages contain only benign explanations or contain neither benign nor serious conditions, they are more likely to be located toward the bottom of the ranking. Second, the distribution of results with serious illnesses remains relatively constant over the rank positions. Focusing on the anchor text, we found that on average seven of the top ten results had anchor text that contained at least one serious condition. The distribution of anchor text content across the top results is shown in Figure 4. The figure illustrates that there is a general decrease in the presence of anchor text that only refers to serious illnesses and anchor text only referring to benign explanations with rank position. Pages associated with anchor text referring to them in terms of both serious and benign explanations are slightly more likely to occur near the top of the ranking. This result may not be surprising since the links are made by page authors, whose actions are also influenced by their own perceptions about the content of the Web pages that they are linking to and their beliefs and expectations about information that future visitors to their site will be interested in reviewing.

These results in both Figures 3 and 4 differ markedly from those in Figure 2, which shows large increases in the likelihood of serious illnesses for captions appearing early in the ranking, and little variation in benign explanations with rank. A potential explanation for these differences is that results linked to captions with content that may alarm users are clicked on more frequently, and the ranker learns this behavioral signal and places these results higher, regardless of the content in the results themselves. Several studies have shown that search engines leverage logged clicks in this way (e.g., [Agichtein et al. 2006]), making their result rankings susceptible to biases in click behaviors associated with captions with potentially-alarming content. Although Figure 4 reveals a noticeable increase in both benign and serious content with rank, the increases are significantly smaller than the changes seen with lowering rank in captions. The anchor text is also more balanced than the captions (i.e., only *serious and benign* increases with rank, not *serious only* or *benign only*). One explanation is that anchor text may not be as affected by biased behaviors in authoring as ranking might be from caption biases affecting click-through. Note that there were no dominant medical conditions comprising the serious and benign conditions in Figures 2–4.

3.2.1. Relative Ordering on Landing Pages. Moving beyond the order of the conditions in the result ranking, previous research has shown that the relative ordering of serious illnesses and benign explanations within the content of a Web page is a strong predictor of escalation in medical concerns [White and Horvitz 2010]. Specifically, the work showed that when mention of a serious condition precedes a benign explanation, immediate escalations in queries occur about 70% of the time versus 30% for the reverse ordering. We can use this observation to help us understand the relationship between the content of captions and the likelihood that a searcher will escalate their search by querying next on serious, rare conditions. We computed the frequency with which captions contained serious illnesses and the relative ordering of those illnesses compared to benign conditions on the landing pages. The results of this analysis are presented in Table I.

The results in the table show that landing pages linked to captions containing serious illnesses are more likely to contain mentions of those serious illnesses before benign explanations ($\chi^2(1) = 88105.52$, $p < 0.001$). Nearly 45% of pages linked to snippets containing concerning terms will contain content with concerning terms appearing before benign explanations. This result can be coupled with the result reported in prior studies that users will escalate their successive searches in 70% of views of such pages [White & Horvitz 2010]. We can assert via the chain law of probability that clicking on a caption containing concerning terms will lead to escalations in more than 30% of click-through. Overall, the presence of potentially-alarming content in prominent positions on the SERP is concerning, especially if some users assume the results are ranked by likelihood [White & Horvitz 2009b], and user engagement with these captions is worth studying further to understand the degree of influence that potentially-alarming caption content has on search behavior.

We shall now study users' caption examination behavior, independent of rank, for each of the four combinations in presence/absence of serious and benign conditions.

3.3 Effect on Examination Behavior

To study examination behavior at scale we used logs recently collected via cursor tracking instrumentation on the Microsoft Bing SERP. Data were collected over almost four weeks from a group comprising a randomly assigned 1% of Bing users. Users remained in the group for the full 25 days. In addition to traditional search behavior such as queries and clicks, we also recorded users' mouse cursor behavior.

3.3.1. Capturing Cursor Behavior at Scale. To record user interactions on the SERP at scale without the need to install any browser plugins, we used an efficient and scalable approach developed by Buscher and colleagues [Buscher et al. 2012]. The method uses entirely JavaScript-based logging functions that were embedded into the HTML source code of the SERP for the Microsoft Bing search engine. To obtain a detailed understanding of user interactions with the SERP, we deployed methods to measure and record a variety of interactions with the page as well as page characteristics, such as the layout of elements on the SERP. We recorded information on cursor movements, clicks, scrolling, as well as bounding boxes of *areas of interest* (AOIs) on the SERP, such as each of the result captions, as well as the Web browser's viewport size (e.g., the dimensions of the browser's view on the source Web page). We periodically checked the cursor's x- and y-coordinates within the Web page relative to its top-left corner of the page every 250 milliseconds. Whenever the cursor had been moved more than eight pixels away from its previously logged position, its new coordinates were sent to a backend server at Microsoft. Eight pixels corresponds to approximately half a line of text on the SERP.

Cursor logs were gathered during a 25-day period that fully overlapped with the duration of the logs described earlier during an external experiment on a small fraction of user traffic from the U.S. English geographic locale. The data were sampled by user, storing every query from each participating user. We filtered the data for symptom queries described earlier. The resultant data comprised 2,070 symptom queries from 714 users. We used these data to explore caption examination.

3.3.2. Differences in Hovers. From the cursor tracking logs, we extracted hovers on the captions of each of the search results. As in previous work [Huang, White & Dumais 2011], we used caption hover events as a proxy for the user having examined the caption. We defined a hover as time spent with the cursor placed inside the AOI of the caption. To reduce noise we required that a hover last at least one second (4× the

Table II. Features of mouse cursor behavior for snippets with and without serious illnesses and/or benign explanations. Table shows mean and standard error (parenthesized). *N* = number of hovers.

Caption has benign explanation	Caption has serious illness: Yes		Caption has serious illness: No	
Yes	*N*=321		*N*=348	
	Number of hovers	1.05 (0.02)	Number of hovers	1.02 (0.02)
	Time per hover (secs)	4.17 (0.32)	Time per hover (secs)	4.16 (0.28)
	HTime/char	0.024 (0.003)	HTime/char	0.025 (0.003)
	AOI time (secs)	4.31 (0.44)	AOI time (secs)	5.88 (0.47)
	P(Click | Hover)	0.241	*P*(Click | Hover)	0.094
No	*N*=443		*N*=1018	
	Number of hovers	1.36 (0.03)	Number of hovers	1.05 (0.01)
	Time per hover (secs)	5.07 (0.22)	Time per hover (secs)	3.87 (0.19)
	HTime/char	0.030 (0.003)	HTime/char	0.022 (0.002)
	AOI time (secs)	9.07 (0.53)	AOI time	4.81 (0.25)
	P(Click | Hover)	0.285	*P*(Click | Hover)	0.106

sampling rate) and less than 30 seconds. The 30-second threshold was meant to remove outliers unassociated with user attention (e.g., user parking cursor over caption and becoming distracted, meaning that they stopped attending to the mouse position). The cursor could move within the caption AOI during the hover; the hover terminated if they left the AOI or on timeout.

The next step was to determine whether there were any differences in the caption examination behavior depending on the presence and absence of serious illnesses and benign explanations, we calculated the following features for each combination:

- **Number of hovers:** Average number of distinct hovers per caption, given that at least one hover is observed.
- **Time per hover:** Average hover time over caption.
- **Hover time per character (HTime/char):** Average hover time normalized by the number of characters in the caption.
- **AOI time:** Total time spent hovering on caption. This can span many distinct hovers and also sums the duration of all non-hovers (i.e., those under one second).

Taking the cursor position as a reasonable proxy for attention, these hover features can provide us with insights about searcher engagement with the captions, even if they do not click on any of the result hyperlinks. The number of hovers captures repeat visits to the caption (perhaps reflecting heightened interest), and the temporal features capture different aspects of the amount of time spent with the cursor in the caption (e.g., time when they may be examining the caption content). There may be variations in hover time due to caption length for the simple reason that longer captions take more time to read. We calculate *HTime/char* for each caption as a way to address varying caption lengths. Other features could be computed that are based on users' search behavior while the user is attending to the caption, e.g., to monitor for evidence of reading behavior [Rodden et al. 2008] or to gather text selections to estimate searcher interests [White & Buscher 2012].

We computed these features across four hover groups: (1) hovers over captions mentioning serious illnesses, (2) captions with benign explanations, (3) captions with both serious illnesses and benign explanations, and (4) captions with neither serious illnesses nor benign explanations. We also computed the normalized hover time per character to counter the influence of a larger amount of text on longer hovers.

As mentioned earlier, order effects have been shown to have a marked influence on how people examine SERPs [Joachims et al. 2007]. If we simply used all hovers we would be unable to attribute any observed differences in examination behavior to the

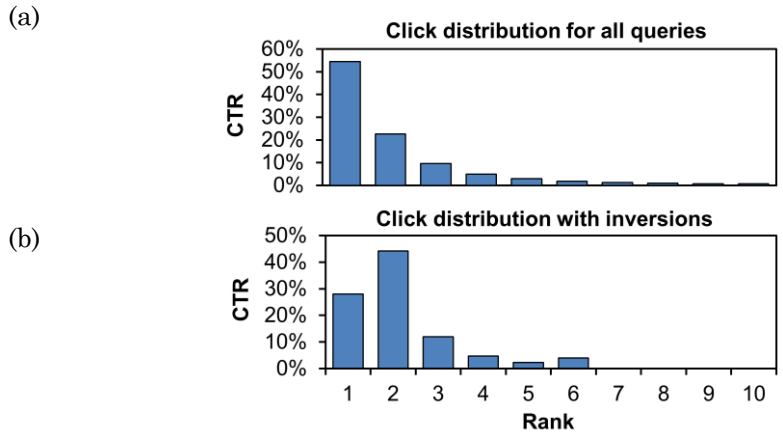

Fig. 5. Click-through curves across the top-10 rank positions for (a) all queries, and (b) the symptom query [stomach pain] with click-through inversions at rank positions two and six.

content of the caption. Since we were performing this study retrospectively, we did not have an opportunity to instrument the SERP to gather unbiased clicks using a method such as FairPairs [Radlinski & Joachims 2006]. To isolate the hover features from the rank position, for each of the groups we sampled hovers uniformly across all of the top 10 rank positions. This means that a hover on a result at position 10 had as much chance as being included as a hover at position 1. Down-sampling in this way allowed us to control for rank, but also means that there was an upper bound that was the minimum number of hovers, usually observed at rank position 10. In doing so, we also preserved all hovers for each query session, allowing us to also compute the total number of hovers on each of the captions on a per-query basis. This method resulted in around 200 hovers per rank position. Table II has the contingency table with the mean average and standard error for each feature.

The findings in presented in Table II appear to show differences related to the presence and absence of serious illnesses and benign explanations. We applied two-way analyses of variance (ANOVAs) between each of the groups for the three hover features. To reduce the chance of Type I errors due to multiple comparisons, we used a Bonferroni correction to adjust α to 0.0125. The ANOVAs showed differences for each of the four hover features (all $F(1, 2126) \geq 7.72$, $p \leq 0.006$). Results from Tukey-Kramer post-hoc testing showed that users hover on captions with serious illnesses more often ($p = 0.001$), average time per hover is longer ($p = 0.003$) (even when normalized for caption length ($p = 0.004$)), and total time in the caption AOI is higher ($p < 0.001$). This finding suggests that users are examining captions with serious illnesses in more detail than other types and supports our hypothesis that concerning content in snippets influences examination behavior. However, examination via hovers only provides limited insight into SERP engagement and we also seek to understand whether content biases in captions influence *click-through* behavior.

3.3.3. Clicks Conditioned on Hovers. We studied the SERP click-through behavior using the same data as in the previous section. We focus on cases where we observed a hover followed by a click. This allowed us to be more confident that the user had examined the caption prior to clicking (we remove this requirement in the detailed analysis we perform in the next section). We computed $P(\text{Click} \mid \text{Hover})$ for each of the four groups and report the results of this analysis in Table II. The findings show that when at least one serious illness is in the caption, the click probability is higher ($F(1, 2126) = 10.66$, $p = 0.001$; Tukey-Kramer: all $p < 0.001$). Not only are users more likely to

examine captions when they contain potentially-alarming content they are also more likely to engage with them and transition to the landing page via hyperlink clicks.

Overall, our findings support our hypothesis that the presence and absence of potentially-concerning medical conditions in captions (titles, snippets, and/or URLs) influences click-through behavior. However, we cannot guarantee from these findings that it is the content in the snippets that causes people to examine the captions in more detail. Other factors could influence how people attend to caption content (e.g., the other terms in the snippet co-occurring with the potentially-alarming content, users' perceptions of the relevance of the landing page). To more fully establish a relationship between the presence of either potentially-alarming or potentially-reassuring terminology, and click-through (as well as other terms as mentioned above), we needed to understand the extent to which various caption features may contribute significantly to clicks. With that goal in mind, we study click inversions [Clarke et al. 2007] on symptom SERPs. Click inversions let us examine the effect of specific caption features on click-through behavior given the presence of terms in lower-ranked clicked captions and their absence in higher-ranked unclicked captions.

4. CLICK INVERSIONS

We now focus on features of the captions that may motivate users to click on them more than expected given the rank position. We approach this with an analysis of click inversions, introduced in our previous work [Clarke et al. 2007]. Inversions occur when the click-through rate (CTR) for a result is higher than the result directly above, therefore overcoming the position biases affecting clicks and caption examination [Joachims et al. 2007]. Figure 5a shows the expected click-through curve for the rank position computed across all queries. Figure 5b shows the curve for the query [stomach pain] which has inversions at the second and the sixth rank positions. We use the click inversions methodology to study effects of potentially-alarming captions.

4.1 Extracting Click Inversions

4.1.1. Data. Using the data described in Section 3.1, we seek a consistent ordering of results and consistency in the content of captions over which the CTR distribution was computed. Since the result order and captions may change during the two-month period it is not possible to simply create a single top-10 for each query. We did three things to address this challenge: (1) we assigned all unique SERPs for each query (in terms of results, result rankings, and captions) an identifier and treated this separately in the remainder of our analysis. There were approximately five different SERP arrangements for one of the symptom queries over the duration of the logs (some with inversions and some without); (2) we retained click-through for a specific combination of a query and a result only if this result appears in a consistent position for at least 50% of the click-through. Click-through for the same result when it appeared at other positions were discarded; and (3) if we did not observe at least ten clicks for a particular query during the sampling period, no clicks for that query were retained.

When identifying clicks, we consider only the first click-through action taken by a user after entering a query and viewing the result page. By focusing on the initial click-through, we hope to capture a user's impression of the relative relevance within a caption pair when first encountered. If the user later clicks on other results or re-issues the same query, we ignore these actions. Any preference captured by a click-through inversion is therefore a preference among a group of users issuing a particular query, rather than a preference on the part of a single user.

Table III. Comparison of relevance of results (A = more highly ranked by search engine).

Relationship	Number	Percent
rel(A) < rel(B)	668	29.32%
rel(A) = rel(B)	982	43.11%
rel(A) > rel(B)	628	27.57%

Following these steps, the data comprises a set of records with each record describing the clicks for a given query/result combination. Each record includes a query, a rank position, a caption, the number of clicks for this result, and the total number of clicks for this query. We process this set to generate click-through curves and identify inversions. In total, 193 unique symptom queries and 902 unique query-{result list} combinations met these criteria.

As suggested in [Clarke et al. 2007], there may be several reasons for inversion in a click-through curve. The search engine may have failed to rank more relevant results below less relevant results. Even when the relative ranking is appropriate, a caption may not reflect the content of the underlying page with respect to the query (as was suggested by our earlier analysis comparing captions), leading the user to make an incorrect judgment. Before turning to the second case, we address the first, and examine the extent to which relevance alone may explain these inversions.

4.1.2. Association with Relevance. For each click-through inversion, we have two results of interest: result A, which is more highly ranked by the search engine, and result B, which the search engine ranks lower. To determine the relevance of the result at the higher position, A, and the result at the lower position, B, we used trained human judges, recruited as part of an internal relevance assessment effort. Judges assigned labels on a four-point relevance scale—*excellent*, *good*, *fair*, and *bad*—to each URL for each query. Each query-URL pair was assessed by at least three judges to obtain consensus and by at most five judges. Table III shows the results for queries where three of the judges agreed on the relevance of the URL. We dropped the other query-URL pairs from this analysis because we was sufficient disagreement between judges for us to be concerned about label reliability. If inversions were only attributable to relevance we would expect B to frequently be more relevant than A.

The results show little difference in the relevance between A and B. Relevance is generally equal and only slightly in favor of B, but not often enough to account for the many click inversions in the labeled data. Having demonstrated that click-through inversions cannot always be explained by relevance, we explore caption features that may lead users to prefer one result over another.

4.2 Methodology

We extracted two sets of caption pairs from S. The first is a set of 2,278 click-through inversions, extracted according to the procedure described earlier in this paper (Section 3.1). The second is a corresponding set of caption pairs that do not exhibit click-through inversions. In other words, for pairs in this set, the result at the higher rank (caption A) received more click-through than the result at the lower rank (caption B). To the greatest extent possible, each pair in the second set was selected to correspond to a pair in the first set, in terms of result position and number of clicks on each result. For the remainder of this analysis, we shall refer to the first set, containing inversions, as the INV set; we refer to the second set, containing caption pairs for which the click-through are consistent with their rank order, as the CON set.

We extracted a number of features characterizing captions (described in detail in the next section) and compare the presence of each feature in the INV and CON sets.

Table IV. Features measured in caption pairs (caption A and caption B), with caption A as the higher ranked result. Features are expressed from perspective of prevalent relationship predicted for click-through inversions.

Category	Feature Tag	Description
Course	Acute	caption B (but not A) contains the term "acute"
	Chronic	caption B (but not A) contains the term "chronic"
Degree	Severe	caption B (but not A) contains the term "severe" (or variants, e.g., "serious", "terrible")
	Mild	caption B (but not A) contains the term "mild" (or variants, e.g., "moderate")
Tendency	Malignant	caption B (but not A) contains the term "malignant"
	Benign	caption B (but not A) contains the term "benign"
Prognosis	Deadly	caption B (but not A) contains the term "deadly" (or variants, e.g., "fatal", "grave")
	Nonfatal	caption B (but not A) contains the term "nonfatal" (or variants, e.g., "harmless")
Transition	Escalations	caption B (but not A) contains an serious illness related to the symptom in query
	NonEscalations	caption B (but not A) contains an benign explanation related to the symptom in query
Condition	AnySeriousCondition	caption B (but not A) contains any serious illness
	AnyBenignCondition	caption B (but not A) contains any benign explanation
	Cancer	caption B (but not A) contains the term "cancer" (with stemming)
	Pregnancy	caption B (but not A) contains the term "pregnancy" (with stemming)
Healthcare utilization	MedicalFacility	caption B (but not A) contains a medical facility
	MedicalSpecialist	caption B (but not A) contains a medical specialist
	MedicalProfessional	caption B (but not A) contains a medical professional such as a physician
Source	MayoClinic	title or snippet or URL of caption B (but not A) contains the term "mayo clinic"
	WebMD	title or snippet or URL of caption B (but not A) contains the term "webmd"
	MedlinePlus	title or snippet or URL of caption B (but not A) contains the term "medlineplus"
	PubMed	title or snippet or URL of caption B (but not A) contains the term "pubmed"
Snippet	MissingSnippet	snippet missing in caption A and present in caption B
	SnippetShort	short snippet in caption A (< 25 characters) with long snippet (> 100 characters) in caption B
Term match	TermMatchTitle	title of caption A contains matches to fewer query terms than the title of caption B
	TermMatchTS	title+snippet of caption A contains matches to fewer query terms than caption B
	TermMatchTSU	title+snippet+URL of caption A contains matches to fewer query terms than caption B
	TitleStartQuery	title of caption B (but not A) starts with a phrase match to the query
	QueryPhraseMatch	title+snippet+url contains the query as a phrase match
URL	URLQuery	caption B URL takes the form *www.query.com*, where the query matches exactly minus spaces
	URLSlashes	caption A URL contains more slashes (i.e. a longer path length) than the caption B URL
	URLLenDIff	caption A URL is longer than the caption B URL
Readability	Readable	caption B (but not A) passes a simple readability test

We describe the features as a hypothesized preference (e.g., a preference for captions containing the name of a serious illness). Thus, in either set, a given feature may be present in one of two forms: favoring the higher ranked caption (caption A) or favoring the lower ranked caption (caption B). For example, the absence of a serious illness in caption A favors caption B, and the absence of a serious illness in caption B favors caption A. When the feature favors caption B (consistent with a click-through inversion) we refer to the caption pair as a *positive pair*. When the feature favors caption A, we refer to it as a *negative pair*. For serious illnesses, a positive pair has a serious illness mentioned in caption B (but not A) and a negative pair has a serious illness mentioned in A (but not B).

Therefore, for each feature we built four subsets: (1) INV+, the set of positive pairs from INV; (2) INV−, the set of negative pairs from INV; (3) CON+, the set of positive pairs from CON; and (4) CON− the set of negative pairs from CON. The sets INV+, INV−, CON+, and CON− will contain different subsets of INV and CON for each feature. When stating a feature corresponding to a hypothesized user preference, we follow the practice of stating the feature with the expectation that the size of INV+ relative to the size of INV− should be greater than the size of CON+ relative to the size of CON−. For example, we state the serious illness feature as "a serious illness missing in caption A and present in caption B". This methodology allows us to create a contingency table for each feature, with INV as the experimental group and CON the control group. Given those tables, we then applied Pearson's Chi-square test to compute the significance of the differences between the two groups.

Table V. Results corresponding to the features listed in Table IV with χ^2 and p-values (df = 1). Features related to inversions and supported at 95% confidence level are bold. In rows with any cell count < 5 we use a Fisher's exact test.

Category	Feature Tag	INV+	INV–	%+	CON+	CON–	%+	Diff	χ^2	p-value
Course	**Acute**	38	13	74.51	23	45	33.82	+40.69	19.309	< .0001
	Chronic	48	54	47.06	61	43	58.65	–11.59	2.7787	0.0955
Degree	**Severe**	105	65	61.76	71	99	41.76	+20.00	13.6170	0.0002
	Mild	13	52	20.00	14	7	66.67	–46.67	16.0483	< .0001
Tendency	**Malignant**	72	33	68.57	45	55	45.00	+23.57	10.6700	0.0011
	Benign	29	29	50.00	53	37	58.88	–8.88	0.8	0.3711
Prognosis	**Deadly**	22	6	78.57	12	15	44.44	+34.13	6.7824	0.0092
	Nonfatal	4	5	44.44	7	7	50.00	+5.55		0.2469
Transition	**Escalations**	111	54	67.27	42	46	47.73	+19.54	9.1725	0.0025
	NonEscalations	90	70	56.25	118	104	53.15	+3.10	0.3596	0.5486
Condition	**AnySeriousCondition**	274	189	59.18	236	246	48.96	+10.22	9.9223	0.0016
	AnyBenignCondition	329	302	52.14	310	336	47.99	+4.15	2.04	0.1532
	Cancer	31	19	62.00	16	40	28.57	+33.43	11.9605	0.0005
	Pregnancy	28	22	56.00	27	27	50.00	+6.00	0.1729	0.6801
Healthcare utilization	MedicalFacility	101	105	49.03	131	143	47.81	+1.22	0.06996	0.7914
	MedicalSpecialist	6	5	54.55	13	2	86.67	–32.12		0.0847
	MedicalProfessional	115	145	44.23	153	84	64.56	–20.33	20.6167	< .0001
Source	**MayoClinic**	75	66	53.19	90	123	42.25	+10.94	4.0788	0.0434
	WebMD	81	30	72.97	47	48	49.47	+23.50	12.0149	0.0005
	MedlinePlus	32	60	34.78	69	40	63.30	–28.52	16.2328	< .0001
	PubMed	3	10	23.08	12	4	75.00	–51.92		0.0073
Snippet	MissingSnippet	14	20	41.18	3	9	25.00	+16.18		0.2614
	SnippetShort	6	2	75.00	13	20	39.39	+35.61		0.0078
Term match	TermMatchTitle	7	3	70.00	12	13	48.00	+22.00		0.2117
	TermMatchTS	131	127	50.78	192	136	58.54	–7.76	3.5165	0.0608
	TermMatchTSU	82	94	46.59	112	81	58.03	–11.44	4.8319	0.0279
	TitleStartQuery	446	348	56.17	450	414	52.08	+4.09	2.7840	0.0952
	QueryPhraseMatch	213	154	58.04	233	233	50.00	+8.04	5.3329	0.0209
URL	**URLQuery**	16	11	59.26	13	26	33.33	+25.93	4.3535	0.0369
	URLSlashes	833	644	56.4	718	861	45.47	+10.93	36.4513	< .0001
	URLLenDiff	1471	753	66.14	1166	1218	48.91	+17.23	139.5928	< .0001
Readability	Readable	22	30	42.31	22	24	47.83	–5.52	0.3004	0.5836

4.3 Features

We devised features associated with potentially-alarming content, and variants which may not be likely to cause such alarm. We selected features that explicitly captured different aspects of clinical and diagnostic procedure and were sufficiently popular to be observed appear in the caption text. The features are listed in Table IV, grouped in the following categories:

- **Course:** The duration of a condition and/or the nature of its onset (e.g., "acute" may be associated with a condition with short duration and rapid onset).
- **Degree:** The extent or severity of the condition (e.g., "severe" may be associated with an extreme symptom or condition). The non-serious variant in this case was "mild" or "moderate", rather than "none", since the symptom (e.g., "mild back pain") needed to be observed to at least some extent by the searcher.
- **Tendency:** The trajectory of a condition over time (e.g., "malignant" may be used to describe a severe, progressively-worsening disease most commonly associated with cancer).
- **Prognosis:** The likely outcome of a medical condition. The term "deadly" (and its variants) could be associated with terminal conditions. In contrast, the term "nonfatal" could be associated with non-life threatening conditions.
- **Transition:** The nature of the transitions, if any, between the symptom query and the conditions in the caption. For this we used the list of symptom-condition pairs from previous work [White and Horvitz 2009a] (e.g., an escalation for the symptom "chest pain" is "heart attack" or "myocardial infraction", whereas a non-

escalation is "indigestion"). These were initially validated by the authors using Web research and their own medical knowledge (received through training in the case of EH). Prior to using these transitions in the analysis here, they were also validated by a physician who was not part of the research team. This was the same expert who judged the symptom-synonym pairs described earlier.

- **Condition:** The presence of any serious or benign conditions in the captions, as well as specific serious and benign conditions—cancer and pregnancy—which appear frequently in the logs.
- **Healthcare utilization:** Whether the caption relates to the use of in-world medical care (e.g., hospital, neurologist, and physician for facility, specialist, and professional respectively). Transitions from searches on conditions to searches on professional care have been shown to be common in previous work [White & Horvitz 2010]. Such a transition may reflect an escalation in concern and we wanted to capture evidence of such transitions in the SERP clicks studied here.
- **Source:** Site containing landing page, with consumer (Mayo Clinic, WebMD) and professional (MedlinePlus, PubMed). In previous work, searchers were shown to prefer particular Web domains [Ieong et al. 2012], and we believed that such preferences may also be evident in the click inversions studied in this article.

Within many of these feature categories, a likely reassuring variant is associated with possible lowered concern and a likely alarming variant is associated with possible heightened concern. Although this list of categories is not exhaustive, it allows us to analyze searcher behavior with respect to captions for a range of different health seeking intentions, and study whether potentially-alarming content affects behavior. If our hypothesis about the attractiveness of potentially-alarming content is correct, we should see significant increases in click inversions for cases where the alarming variant is present and a reduction (or no change) in inversions for captions where the reassuring variant is present.

For comparison, we include some of the features from our previous work [Clarke et al. 2007] to understand whether these are informative for symptom queries and to study the nature of any differences between the feature classes. These features are highlighted with a gray background at the bottom of Table IV. Most of these features can be understood from the descriptions provided in the table. Readability is implemented in a simple way as it was in the previous study of click inversions. That is, based on whether more than 40% of one caption comprises one of the top-100 most common English words and the other caption comprises fewer than 10%.

4.4 Findings

In this section we provide results focused on two aspects of the analysis: the *feature level* (i.e., the features from Table IV that are most strongly associated with click inversions) and the *term level* (i.e., the individual terms that have a positive or negative affect on click-through).

4.4.1. Feature Level. We begin by looking at the inversion statistics for each feature. The results are presented in Table V. In order to reject the null hypothesis, the positive percentage (+%) should be significantly greater for INV than CON. If the difference is significantly positive, it means that the feature is associated with inversions more than would be expected by chance (if significantly negative, it negatively affects click-through). For cases where the counts in the cells were all at least five, we applied the Chi-squared test of independence to these sizes, with p-values shown in the last

Table VI. Terms exhibiting the greatest positive (↑) and negative (↓) influence on click-through patterns.

Rank	Term	χ^2	Influence
1	encyclopedia	125.9786	↓
2	wikipedia	93.2633	↓
3	causes	89.1112	↑
4	free	82.5041	↓
5	symptoms	79.1907	↑
6	treatment	66.1347	↑
7	webmd	62.2521	↑
8	severe	52.1414	↑
9	learn	47.3026	↑
10	tumor	47.1595	↑

column. For cases where at least one of the counts was less than five we used the Fisher's exact test. Features supported at the 95% confidence level and favoring click inversions are shown in bold.

The findings show that the presence of more potentially-alarming content is associated with increased click likelihoods across many of the categories. For example, the presence of "malignant" significantly increases the likelihood of a click whereas "benign" slightly reduces the likelihood of click-through. There are also differences attributable to the source of the information, with consumer sites like MayoClinic.com and WebMD.com having a positive influence on click-through, whereas technical content (e.g., from PubMed) was likely to dissuade users from clicking. This apparent preference for particular online sources may be related to domain bias described earlier [Ieong et al. 2012], including factors such as the perceived credibility of the site [Schwarz & Morris 2011]. The only category of features that did not show a significant relationship with inversions was healthcare utilization (HU). One explanation for this is that pursuing medical facilities and medical professionals is not necessarily a sign of alarm and there can be other reasons for doing so, including visiting patients. Interestingly, of the HU features, it is *MedicalSpecialist* that is closest to significant (p=0.08). One might expect the pursuit of information about specialist care to perhaps be the most likely of the HU types to be associated with medical concern (e.g., concerns about cancer leading to searches for oncologists), at least more frequently than with general practitioners and facilities. In addition, of the features from previous work [Clarke et al. 2007], similar trends were observed although some of the differences were not significant, perhaps because of small counts (as in *MissingSnippet* and *TermMatchTitle*) or our focus on diagnostic queries rather than the random sample used previously, which may explain the no difference in *Readable* since many captions had a similar reading level.

4.4.2. Term Level. As an additional experiment, we wanted to understand whether there were any particular terms that were likely to be associated with higher click-through during diagnostic search. To perform this experiment, we treated each of the terms appearing in the INV and CON sets as a separate feature (case normalized and removing stopwords such as "of" and "the"), and ranked them by their χ^2 values. The top 10 terms are shown in Table VI. Because we use the χ^2 statistic as a divergence measure rather than as a significance test, no p-values are given in the table. The final column of the table indicates the directionality of the influence on click-through of the presence of the term (e.g., up arrow means more clicks when the term is present).

The table shows that the presence of "wikipedia" negatively affects click-through, as do the terms "encyclopedia" and "free" which appear in the titles of Wikipedia

articles. This concurs with the findings of our previous study [Clarke et al. 2007]. Turning our attention to medical terminology we see that the terms "tumor" and "severe" increased click-through and support our claims about potentially-alarming terminology attracting more clicks. Interestingly, we also observe that the presence of "learn" and "causes" in the most influential suggests that health searchers may be interested in understanding and diagnosing the symptoms they seek; captions including those terms appear to increase clicks.

Now that we have established that the presence of potentially-alarming content in the captions for symptom queries can influence click-through behavior, we seek to understand whether such knowledge could help us to better model diagnostic search behavior. Modeling the search process and predicting clicks is a well-studied research area [Chapelle & Zhang 2009; Craswell et al. 2008; Zhang et al. 2010], but the role of caption content in these models is not well understood. We do not seek to develop an optimal caption-based click prediction model with our research, but rather seek to demonstrate that caption features associated with potential user alarm can yield prediction gains for diagnostic search. This is important for demonstrating predictive value in these features that could potentially be leveraged in ranking. For example, by weighting result clicks by the likelihood that they were caused by potentially-alarming caption content and not landing-page relevance.

5. CLICK PREDICTION

We now describe the click prediction model that we constructed using the features from Section 4, and then describe its evaluation and the findings.

5.1 Click Model

We replicate the Dynamic Bayesian Network (DBN) model [Chapelle & Zhang 2009] because it fares well compared to other models (e.g., [Craswell et al. 2008; Zhang et al. 2010]), making it a solid baseline. The DBN model is a graphical model where the nodes represent states of the user examining the search results. Users examine results from top to bottom, assessing at each result whether or not it is attractive enough to click (cascade hypothesis), which depends only on the attractiveness of the link a_u estimated based on clicks (examination hypothesis). If a user clicks, there is some probability s_u of satisfaction and a cessation of search; given a continuation of search, they either return to the SERP to examine the next search result with probability γ, or abandon the search. We re-implement the DBN model with $\gamma = 1$, labeled Algorithm 1 in Chapelle and Zhang [Chapelle & Zhang 2009]. We chose this algorithm and $\gamma = 1$ in order to simplify the inference of latent variables. Then we modify the attractiveness of result a_u and generate an enhanced attractiveness value a'_u that incorporates caption features and weights assigned using the chi-squared values from Table V. Attractiveness provides an estimate of whether users are likely to be drawn to the result (not just notice it, but be attracted to it). a_u is initially computed based on a prediction of the click-through rate at the first rank position for the current query, as in [Chapelle & Zhang 2009]. This prediction is based on a held out set of four months of Bing query logs from 1 January 2012 until 1 April 2012, and before the timespan used for the other analysis described in this article. The revised attractiveness estimate a'_u is computed based on the weights arising from the analysis in the previous section and the following,

$$w_u = \sum_{f \in F_u} \text{sign}(d_f) \cdot \log(c_f + 1) \qquad a'_u = a_u + \alpha . n(w_u) \quad \text{where,} \qquad (1,2)$$

$$\text{sign}(x) = \begin{cases} -1, x < 0 \\ 1, \quad x \geq 0 \end{cases} \qquad n(w_u) = \frac{w_u - w_{MIN}}{w_{MAX} - w_{MIN}} \qquad \alpha = \mathbb{1}_{w_u \geq 0} - a_u \qquad (3,4,5)$$

The sign is computed for each feature f in all features F_u based on the difference (d_f) in the percentage positive clicks between the INV and the CON sets for that feature. With this approach, we capture whether a feature is likely to have a positive or negative effect on click-through. We use the notation c_f for the value of the Chi-squared test for the feature f in Table V. We add one to c_f to avoid negatives and take the log to make the value more stable. The sign and this value are multiplied together to yield w_u, which is the raw attractiveness weight of the caption. w_u is then rescaled in 0 and 1 via $n(w_u)$ (Eq. 4) with respect to the other captions on the SERP, and α is used to ensure that the value of a'_u does not exceed one.

5.2 Click Perplexity

Click perplexity measures how "surprised" a click prediction model is upon observing a click on a result [Dupret & Piwoworksi 2008]. The perplexity over a set of binary observations is estimated via the geometric average of the predicted probability of the observations. It reflects the average number of times that an experiment needs to be repeated to observe a correct prediction. As a result, the lowest attainable perplexity (of the perfect deterministic model) is one, meaning the trained model perfectly predicted the test data, while a larger perplexity means the model was less accurate.

Click perplexity has been employed in a number of similar studies [Chapelle & Zhang 2009; Guo et al. 2009; Zhang et al. 2010] as a measure of predicting click-through rates. Our evaluation is similar to the previous studies: query sessions were divided evenly into training and test sets, each comprising at least five query sessions; we only accepted one query session from each user for a particular query to prevent a small number of users from dominating the data. The weights in Table V are based on clicks over the full two months of logs. To hide knowledge about the future in the training of the models, we used two weeks of logs immediately following the other time period. χ^2 values were determined based on the preceding two months, but were applied in the model on the new two-week period. There were 438 symptom queries that were issued by at least 10 users. This removed queries with insufficient data.

We compared the DBN model using only click data, with the DBN model using the updated attractiveness score based on click inversions in addition to the click data. These datasets were used to train the searcher model, and the trained model was used to predict clicks in the test set. Better prediction of clicks in the test set implies that the searcher model (and its inferred parameters) better reflects the result examination process. Click perplexity quantifies how much the test data surprises the trained model; it is computed for each combination of query and position as,

$$p_i = 2^{-\frac{1}{N}\sum_{n=1}^{N}\left(C_i^n \log_2 q_i^n + (1-C_i^n)\log_2(1-q_i^n)\right)} \qquad (6)$$

where p_i is the perplexity in the ith position, N is the number of links, q_i^n is the predicted click probability for the nth query session, and C_i^n denotes whether the user clicked the search result. The exponent represents the cross-entropy estimated from a probability distribution. Because the lower bound of the perplexity depends on click-

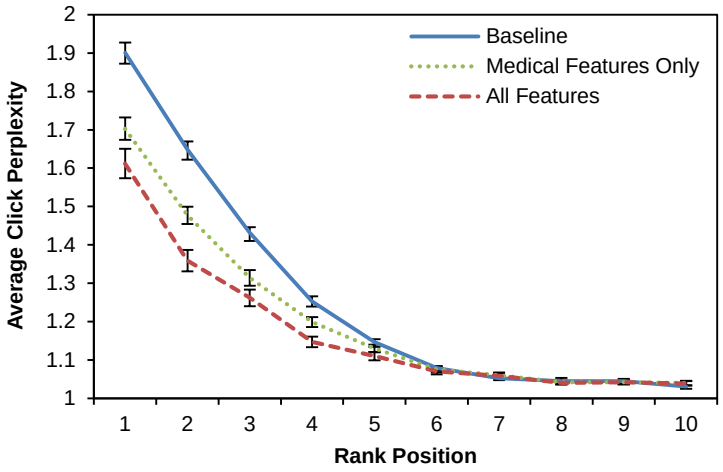

Fig. 4. Perplexity curves for DBN-model variants.
Lower perplexity represents better prediction. Error bars denote standard error.

through rate, perplexity varies substantially depending with rank. Thus, we computed a separate perplexity value for each of the top ten rank positions.

5.3 Results

Figure 4 shows the computed perplexities for each rank position on the SERP for each of three DBN models: the baseline model using only clicks, a modified model using the updated a_u based on all features with a χ^2 value in Table V, and a model using the updated a_u based on all features except those from previous work [Clarke et al. 2007]. The latter was included to measure the prediction performance if we only examined those features associated with potential medical alarm.

The findings show that our click prediction model is significantly improved by the addition of the caption attractiveness features. Paired t-tests were employed at each rank position to compare the performance of our experimental models with that of the baseline. In both cases, significant differences from the baseline were noted at the first four rank positions (all $t(436) \geq 3.62$, all $p < 0.001$). We also see that although the performance of the full model is the best, the other experimental model which only uses the weights from the medical features still has significant gains over the baseline model (all $t(436) \geq 3.11$, all $p \leq 0.003$). This is encouraging as it suggests that the performance gains are not simply attributable to non-medical caption features from earlier work, but rather that there are additional signals related to the presence of potentially-alarming content that can be observed and learned.

6. DISCUSSION AND IMPLICATIONS

This article describes a study of the role of potentially-alarming caption content in search engines and the influence that content can have on search behavior. The study demonstrates that it may be possible to learn more about searchers' unexpressed anxieties, concerns, and fears by studying cursor hovers and click-through on captions containing potentially-alarming content.

We showed that users examined the captions results containing serious illnesses in more detail, and that attributes of the captions containing terms reflecting serious concerns appear to make these results more likely to be selected, as measured using click inversions. We found that results with serious illnesses in their captions were typically ranked more highly in search-result lists than those without, when the same

ranking effect is not observed when we consider result content or anchor text pointing to pages. This suggests that *non-content* features are influencing the result order, primarily (we suspect) large numbers of clicks on captions with potentially-alarming content. To demonstrate the value of considering potentially-alarming caption content in modeling user behavior, we modified a click prediction model and found significant improvements in prediction accuracy by considering the presence of potentially-alarming content in SERP captions.

Methods for identifying and addressing potentially-alarming content in SERPs frame a broader discussion on the role of search engines in society and the interactions among people's intentions, hopes, fears, anxieties, and biases of judgment and content in the ranking of search results. Search engines serve as the gateway to the Web for the majority of online users and searchers make inferences about likelihoods of explanations and outcomes from the rank returned pages. For example, in a survey distributed as part of previous work, we showed that 25% of people interpreted the ranking in the search results as a direct ranking of medical outcomes by likelihoods [White & Horvitz 2009a]. While search engines certainly cannot control the content that is displayed on the pages that they index, they can decide whether they include them, they can control the generation methods that create the captions used by people in making decisions about reading more detailed content, and they can influence how ranking algorithms learn from this implicit feedback. Implicit feedback can lead to a "rich get richer" scenario, where pages that are popular end up being more highly ranked, making it challenging for other pages to break into the first page of search results [Cho & Roy 2004]. However, contradictory evidence suggesting that the combination of retrieval by search engines and search behavior mitigates the attraction of popular pages, directing traffic to less popular sites, even more than what would be expected from users randomly surfing the Web [Fortunato et al. 2006].

The results have implications for many aspects of search system development, primarily the design of captions, ranking algorithms that learn from clickthrough, and learning from logged search activity more broadly. In a recent study by Lauckner and Hsieh [Lauckner & Hsieh 2013] on health search, SERPs generated in response to searches on symptoms were modified manually and tasks were sent to over 300 remote participants. The authors showed that the presence of serious illnesses at the start of the SERP, and in high frequency throughout the SERP, lead to increased perceptions of threat and resulted in negative emotions. Their findings support our conjectures about potential negative outcomes from including potentially-alarming content on SERPs. In the remainder of this section, we focus on steps search engines can take to reduce the amount of potentially-alarming content in SERPs—through both caption generation and ranking—and reduce unwarranted emotional distress.

To reduce the presence of potentially-alarming content in snippets, caption generation algorithms could be modified to consider base rates when generating caption content for presentation to searchers. One implication of this would be that they only show serious illnesses in captions if those conditions are likely to actually be as a result of the symptom query input by the searcher. Another option is that if serious and benign conditions are mentioned in the content of a search result, the caption-generation algorithm could be designed to provide a balanced perspective, showing both aspects in the caption. Finally, the search engine could augment the presentation of the captions, with indicators of the likelihoods of included conditions given the current query, to help users be more informed before making the decision to click a particular result. Beyond caption augmentation, engines could also augment the SERP more broadly to provide users with a choice about whether to select the

alarming or reassuring route, with pages assigned to each route using the conditional probabilities of the outcomes that they mention given the current symptom query.

Search engines can also consider base rates when ranking search results for diagnostic queries. Given a query, they can use background information from medical resources (e.g., prevalence rates from organizations such as the Centers for Disease Control) to estimate condition likelihoods and factor that information into result ranking for health queries. In doing so, careful consideration needs to be given to ways of gathering, representing, and using base rate information (e.g., search engines could compute query-dependent correctness features for each result in the filter set and use that information in result ranking).

As mentioned earlier, an explanation for the promotion of results with captions that contain serious illnesses is that search engines use previous clicks as a signal in result ranking [Joachims 2002; Agichtein et al. 2006]. If click-through is encouraged by potentially-alarming caption content as our findings suggest, then over time the most serious content may be pushed to the top of the ranking by clicks, leading to the creation of a reinforcing cycle whereby articles with the most concerning content are pushed to the top of the list, are clicked on most frequently, and so on [Cho & Roy 2004]. Careful consideration of the caption content associated with the click—and not just the click itself—may help search engines leverage clicks more objectively. For example, rather than ignoring the caption associated with the result click (as is common practice in current learning-to-rank methods), the likely effect that content had on behavior could be quantified (as we did in Section 4) and used to weight click frequencies by the presence or absence of potentially-alarming content in clicked captions. A click that is estimated to be driven by health anxiety (given a symptom query, a serious condition in the clicked caption, and a low symptom-condition prevalence in the real world) may count for less in ranking than a click with no detectable associated biases.

One limitation of our study is the limited lens afforded by log analysis. Search logs provide an incredible opportunity to study the behavior of large user populations in naturalistic settings. While we can observe online search behaviors, we cannot truly understand people's rationales behind those behaviors without working with them directly. Complementary methods such as surveys and user studies could be useful in understanding some of these motivations, as well as emotional factors and cognitive biases, albeit likely with different populations of users, or in artificial settings. There is a need to further validation of our findings via follow-up investigations with health seekers directly at search time. We have done this effectively in some of our recent research using *in-situ* surveys to understand observed actions such as SERP abandonment [Diriye et al. 2012] and search engine switching [Guo et al. 2011]. Other research on understanding search satisfaction has also used this method to associate behaviors with satisfaction estimates [Fox et al. 2005]. A similar methodology could be effective to study the rationales behind health seeking decisions; users would be presented with a survey (perhaps via a pop-up dialog as has been done previously) at the point of the health search and asked to provide more information, such as the goals of their search, and their affective and cognitive state.

We have focused on diagnostic search in this study because it is a common activity on the Web (as mentioned earlier, 35% of U.S. adults reported using the Web for medical diagnosis [Fox & Duggan, 2013]), and is important to users given the gravity of the outcomes. Focusing on a single domain also afforded us more control over the sources of alarm and the types of features that we studied in our investigations of hovers, inversions, and predictive models. Although the health domain is important, there are other diagnostic scenarios where people make consequential decisions (e.g.,

a car owner may wish to diagnose an engine noise and decide whether to take the car to a dealership at high cost). Before we can make broad claims about the association between potentially-alarming content and click-through we need to explore this relationship in other domains.

Further work is also necessary to understand the relationship between people's domain knowledge and the influence of captions on click-through. Related work on the influence of the source website (where the presence of consumer sites led to increase clicks) suggest that users' domain knowledge may be an important factor in determining which results to select [Ieong et al. 2012]. Indeed, prior research has shown that medical domain experts tend to prefer different sites than novices [White, Dumais & Teevan 2009]. Given prior research, we assume that users (like search engines) ignore base rates and base judgments on information availability [Tversky & Kahneman 1974]. However, this may not be true to the same extent for domain experts and medical professionals. We also need to focus on particular user cohorts to better understand behaviors. For example, those who exhibit health anxiety might be more likely to be drawn toward concerning content and as a result could benefit from the removal of such content from SERPs on an individual basis. Long-term models of search behavior could also be helpful in automatically identifying these users.

7. CONCLUSIONS

We presented a study of the influences of potentially-alarming content in search-result captions on the examination of search results. We demonstrated that search engines rank pages with concerning captions more highly and that this promotion is likely to be associated with increased click-through. Through experimentation with click logs and large-scale cursor tracking, we demonstrated that, even when we control for rank position and relevance (through down-sampling and click inversions), users are still significantly more likely to be drawn to examine and click on captions containing potentially-alarming content. We discussed how this result can be used to develop more sophisticated click prediction models and demonstrated the value of features derived from the presence of potentially-alarming content for improved click predictions. Our findings have implications for how search engines generate and present captions, how they rank search results for diagnostic queries, and how they could leverage higher quality data for training rankers by considering the content of the captions behind the clicks they use. Captions could be designed with care to bias and to be made more balanced, especially for emotive or consequential topics. More generally, we wish to understand how affect, emotion, and biases of judgment influence searching and browsing, moving beyond concerns and fears in health diagnosis to fears, hopes, expectations, and desires more broadly in other domains.

REFERENCES

AGICHTEIN, E., BRILL, E., AND DUMAIS, S. 2006. Improving Web search ranking by incorporating user behavior. *SIGIR*, 19–26.

AGICHTEIN, E., BRILL, E., DUMAIS, S., AND RAGNO, R. 2006. Learning user interaction models for predicting web search result preferences. *SIGIR*, 3–10.

ANDERSON, G. AND HUSSEY, P.S. (2001). Comparing health system performance in OECD countries. *Health Affairs*, 20(3): 219–232.

ASMUNDSON, G.J.C., TAYLOR, S., AND COX, B.J. 2001. *Health Anxiety: Clinical and Research Perspectives on Hypochondriasis and Related Conditions*. Wiley.

AYERS, S. AND KRONENFELD. J. 2007. Chronic illness and health-seeking information on the Internet. *Health*, 11, 3.

BEEFERMAN, D. AND BERGER, A. 2000. Agglomerative clustering of a search engine query log. *KDD*, 407–416.

BHAVNANI, S.K., JACOB, R.T., NARDINE, J., AND PECK, F.A. 2003. Exploring the distribution of online healthcare information. *SIGCHI*, 816–817.
BRIN, S. AND PAGE, L. 1998. Anatomy of a large-scale hypertextual Web search engine. *WWW*.
BUSCHER, G., WHITE, R.W., DUMAIS, S.T., AND HUANG, J. 2012. Large-scale analysis of individual and task differences in search result page examination strategies. *WSDM*, 373–382.
BUSCHER, G., DUMAIS, S.T., AND CUTRELL, E. 2010. The good, the bad, and the random: an eye-tracking study of ad quality in web search. *SIGIR*, 42–49.
CARTRIGHT, M., WHITE, R.W., AND HORVITZ, E. 2011. Intentions and attention in exploratory health search. *SIGIR*, 65–74
CHAPELLE, O. AND ZHANG, Y. 2009. A dynamic Bayesian network click model for Web search ranking. *WWW*, 1–10.
CHO, J. AND ROY, S. 2004. Impact of search engines on page popularity. *WWW*, 20–29.
CLARKE, C., AGICHTEIN, E., DUMAIS, S.T., AND WHITE, R.W. 2007. The influence of caption features on click-through patterns in Web search. *SIGIR*, 135–142.
CRASWELL, N., ZOETER, O., TAYLOR, M., AND RAMSEY, B. 2008. An experimental comparison of click position-bias models. *WSDM*, 87–94.
CUTRELL, E. AND GUAN, Z. 2007. What are you looking for? An eye-tracking study of information usage in web search. *SIGCHI*, 407–416.
DIRIYE, A., WHITE, R.W., BUSCHER, G., AND DUMAIS, S.T. 2012. Leaving so soon? Understanding and predicting Web search abandonment. *CIKM*, 1025–1034.
DUPRET, G.E. AND LIAO, C. 2010. A model to estimate intrinsic document relevance from the click-through logs of a web search engine. *WSDM*, 181–190.
DUPRET, G.E. AND PIWOWARSKI, B. 2008. A user browsing model to predict search engine click data from past observations. *SIGIR*, 331–338.
EYSENBACH, G. AND KOHLER, K. 2002. How do consumers search for and appraise health information on the World Wide Web? *British Medical Journal*, 324: 573–577.
FORTUNATO, S., FLAMMINI, A., MENCZER, F., AND VESPIGNANI, A. 2006. Topical interests and the mitigation of search engine bias. *PNAS*, 103, 34, 12684-12689.
FOX, S. 2011. Health topics. *Pew Internet & Amer. Life Project*. http://pewinternet.org/Reports/2011/Health-online.aspx
FOX, S. AND DUGGAN, M. 2013. Health topics. *Pew Internet & Amer. Life Project*. http://pewinternet.org/Reports/2013/Health-online.aspx
FOX, S., KARNAWAT, K., MYDLAND, M., DUMAIS, S., AND WHITE, T. 2005. Evaluating implicit measures to improve the search experience. *ACM TOIS*, 23, 2: 147–168.
GERHART, S. 2004. Do Web search engines suppress controversy? *First Monday*, 9, 1–5.
GOLDMAN, E. 2006. Search engine bias and the demise of search utopianism. *Yale Journal of Law and Technology*, 188.
GUO, F., LIU, C., KANNAN. A., MINKA, T., TAYLOR, M.J., WANG, Y.M., AND FALOUTSOS, C. 2009. Click chain model in Web search. *WWW*, 11–20.
GUO, F., LIU, C., AND WANG, Y.M. 2009. Efficient multiple-click models in web search. *WSDM*, 124–131.
GUO, Q. AND AGICHTEIN, E. 2010. Ready to buy or just browsing? Detecting web searcher goals from interaction data. *SIGIR*, 130–137.
GUO, Q., WHITE, R.W., ZHANG, Q., ANDERSON, B., AND DUMAIS, S. 2011. Why searchers switch: understanding and predicting engine switching rationales. *SIGIR*, 335–344.
HUANG, J., WHITE, R.W., AND DUMAIS, S.T. 2011. No clicks, no problem: Using cursor movements to understand and improve search. *SIGCHI*, 1225–1234.
IEONG, S., MISHRA, N., SADIKOV, E. AND ZHANG, L. 2012. Domain bias in Web search. *WSDM*, 413–422.
JOACHIMS, T. (2002). Optimizing search engines using click-through data. *SIGKDD*, 132–142.
JOACHIMS, T., GRANKA, L.A., PAN, B., HEMBROOKE, H., RADLINSKI, F., AND GAY, G. 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *TOIS*, 25, 2.
KANG, C., LIN, X., WANG, X., CHANG, Y., AND TSENG, B. 2011. Modeling perceived relevance for tail queries without click-through data. CoRR abs/1110.1112.
KRING, A.M., JOHNSON, S., DAVISON, G.C. AND NEALE, J.M. 2007. *Abnormal Psychology*. 10th edition. Wiley.
LAUCKNER, C. AND HSIEH, G. 2013. The presentation of health-related search results and its impact on negative emotional outcomes. *SIGCHI*, in press.
MOWSHOWITZ, A. AND KAWAGUCHI, A. 2002. Assessing bias in search engines. *IP&M*, 38, 1, 141-156.
MOWSHOWITZ, A. AND KAWAGUCHI, A. 2002. Bias on the Web. *CACM*, 45, 9, 56-60.
MURATA, M., TODA, H., MARSUURA, Y. AND KATAOKA, R. 2009. Query-page intention matching using clicked titles and snippets to boost search rankings. *JCDL*, 105–114.

RADLINSKI, F. AND JOACHIMS, T. 2006. Minimally invasive randomization for collecting unbiased preferences from click-through logs. *AAAI*.
RODDEN, K., FU, X., AULA, A. AND SPIRO, I. (2008). Eye-mouse coordination patterns on web search results pages. *CHI Extended Abstracts*, 2997–3002.
SCHWARZ, J. AND MORRIS, M.R. 2011. Augmenting Web pages and search results to help people find trustworthy information online. *SIGCHI*, 1245–1254.
SHAPARENKO, B., CETIN, O., AND IYER, R. 2009. Data-driven text features for sponsored search click prediction. *ADKDD*.
SILLENCE, E., BRIGGS, P., HARRIS, P.R., AND FISHWICK, L. 2007. How do patients evaluate and make use of online health information? *Social Science & Medicine*, 64, 9: 1853–1862.
SPINK, A., YANG, Y., JANSEN, J., NYKANEN, P., LORENCE, D.P., OZMUTLU, S., AND OZMUTLU, H.C. 2004. A study of medical and health queries to Web search engines. *Health Info. and Lib. J.*, 21: 44–51.
TAYLOR, S. AND ASMUNDSON, G.J.C. 2004. *Treating Health Anxiety: A Cognitive-Behavioral Approach*. Guilford Press.
TOMBROS, A. AND SANDERSON, M. 1998. Advantages of query biased summaries in information retrieval. *SIGIR*, 2–10.
TVERSKY, A. AND KAHNEMAN, D. 1974. Judgment under uncertainty: heuristics and biases. *Science*, 185, 4157.
VAUGHN, L. AND THELWALL, M. 2004. Search engine coverage bias: evidence and possible causes. *IP&M*, 40, 4: 693-707.
WANG, K., GLOY, N., AND LI, X. 2010. Inferring search behaviors using partially observable Markov (POM) model. *WSDM*, 211–220.
WEBER, I. AND CASTILLO, C. 2010. The demographics of Web search. *SIGIR*, 523–530.
WHITE, R.W. AND BUSCHER, G. 2012. Text selections as implicit relevance feedback. *SIGIR*, 1151–1152.
WHITE, R.W., DUMAIS, S.T., AND TEEVAN, J. 2009. Characterizing the influence of domain expertise on Web search behavior. *WSDM*, 132–141.
WHITE, R.W. AND HORVITZ, E. 2009a. Cyberchondria: Studies of the escalation of medical concerns in web search. *TOIS*, 27, 4: 23.
WHITE, R.W. AND HORVITZ, E. 2009b. Experiences with web search on medical concerns and self-diagnosis. *AMIA*, 696–700.
WHITE, R.W. AND HORVITZ, E. 2012. Studies on the onset and persistence of medical concerns in search logs. *SIGIR*, 265–274.
WHITE, R.W. AND HORVITZ, E. 2010. Predicting escalations of medical queries based on web page structure and content. *SIGIR*, 769–770.
WHITE, R.W., RUTHVEN, I., AND JOSE, J.M. 2002. Finding relevant documents using top ranking sentences: An evaluation of two alternative schemes. *SIGIR*, 57–64.
YUE, Y., PATEL, R., AND ROEHRIG, H. 2010. Beyond position bias: examining result attractiveness as a source of presentation bias in click-through data. *WWW*, 1011–1018.
ZHANG, Y., WANG, D., WANG, G., CHEN, W., ZHANG, Z., HU, B., AND ZHANG, L. 2010. Learning click models via probit Bayesian inference. *CIKM*, 439–448.

Received November 2012; revised March 2013; accepted May 2013.