Search and Breast Cancer: On Episodic Shifts of Attention over Life Histories of an Illness

MICHAEL J. PAUL, Johns Hopkins University
RYEN W. WHITE, Microsoft Research
ERIC HORVITZ, Microsoft Research

We seek to understand the evolving needs of people who are faced with a life-changing medical diagnosis based on analyses of queries extracted from an anonymized search query log. Focusing on breast cancer, we manually tag a set of Web searchers as showing patterns of search behavior consistent with someone grappling with the screening, diagnosis, and treatment of breast cancer. We build and apply probabilistic classifiers to detect these searchers from multiple sessions and to identify the timing of diagnosis using temporal and statistical features. We explore the changes in information-seeking over time before and after an inferred diagnosis of breast cancer by aligning multiple searchers by the estimated time of diagnosis. We employ the classifier to automatically identify 1700 candidate searchers with an estimated 90% precision, and we predict the day of diagnosis within 15 days with an 88% accuracy. We show that the geographic and demographic attributes of searchers identified with high probability are strongly correlated with ground truth of reported incidence rates. We then analyze the content of queries over time for inferred cancer patients, using a detailed ontology of cancer-related search terms. The analysis reveals the rich temporal structure of the evolving queries of people likely diagnosed with breast cancer. Finally, we focus on subtypes of illness based on inferred stages of cancer and show clinically relevant dynamics of information seeking based on the dominant stage expressed by searchers.

Categories and Subject Descriptors: H.2.8 [**Database management**]: Database applications

Additional Key Words and Phrases: Medical search, cancer, behavior analysis

ACM Reference Format:
Michael J. Paul, Ryen W. White, and Eric Horvitz, 2016. Search and breast cancer: On episodic shifts of attention over life histories of an illness. *ACM Trans. Web* 0, 0, Article 0 (2016), 27 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

When faced with significant life-changing events such as the onset of a serious illness, people often turn to search engines to better understand their situation and to collect information to guide decisions. Receiving a diagnosis of a serious cancer is shocking and life changing. Patients are immediately faced with medical, psychological, financial, cosmetic, and social challenges. On the medical side, patients are quickly immersed in new terminology about diagnosis, prognosis, and multiple critical and potentially time-sensitive decisions about alternative courses of treatment. Patients and their loved ones seeking understanding and guidance increasingly rely on Web search for locating helpful information [Castleton et al. 2011; Helft 2012; Ofran et al. 2012; Satterlund et al. 2003].

Work conducted during a Microsoft Research internship. Author's addresses: M. J. Paul, University of Colorado Boulder, 315 UCB, Boulder, CO 80309; R. W. White and E. Horvitz, Microsoft Research, One Microsoft Way, Redmond, WA 98052. Email addresses: michael.j.paul@colorado.edu; {ryenw,horvitz}@microsoft.com

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

Disruptive changes, such as being diagnosed with a life-threatening illness, may lead to characteristic patterns of search over extended timelines. Aligning and aggregating common patterns of search and retrieval over time across many searchers can serve as a lens for understanding the interests and intentions of searchers [Richardson 2009; White and Horvitz 2013b; Fourney et al. 2015]. We present such an analysis in this paper, focusing on breast cancer as a sensor of human behavior and attention for understanding the information needs of searchers over time. Observed patterns of interest and concern in search logs can strongly correlate with expected questions and informational needs associated with the diagnosis and treatment of individuals who have been faced with the disruptive news about cancer [Ofran et al. 2012; Paul et al. 2015].

To study the evolving and episodic nature of search in the context of breast cancer, we examine anonymized search logs from the Microsoft Bing Web search engine. We leverage these logs to learn to detect and understand disruptive shifts in the focus of attention of searchers, and to track the evolving informational needs and corresponding search patterns. For our retrospective analysis with anonymized logs, we focus on searchers who demonstrate intensive and long-lived shifts in attention to breast cancer, and who subsequently behave as expected given the life history of the illness and its treatment. After a broad filtering of search logs for general interest in breast cancer, three annotators (including one with formal medical training, who framed and guided the annotation process) manually tagged a subset of searchers as having likely been diagnosed with breast cancer, based on noting a disruptive shift of focus and the timeline of changing information needs. Beyond identifying such searchers, the annotators noted the online search session that appeared to be closest to the time that a diagnosis had been received. These identifications were based on the sudden appearance of a flood of detailed pathology and staging queries coming after queries on screening and then biopsy, appearing in accordance with the rhythm of the life history of breast cancer [Vandergrift et al. 2013; Pérez-Stable et al. 2013]. We use these labeled cases to build classifiers capable of identifying searchers with similar characteristics in large-scale log data.

As an additional verification of classifier accuracy, we correlate rates of breast cancer estimated by considering the geographical (U.S. state) and demographic (gender and age) distribution of searchers, attained via the use of additional meta-data captured in the logs, with national incidence rates provided by the National Cancer Institute. We show that search statistics from those assigned a high probability of having been diagnosed with breast cancer using the classifier provide estimates of incidence that correlate strongly and significantly with ground truth incidence rates. The correlation marks a tenfold increase from that obtained with searchers assigned lower probabilities of having been diagnosed.

The availability of accurate prediction methods facilitates rich analysis of aggregated information-seeking behavior over a population of searchers. Given a set of searchers inferred as experiencing a cancer diagnosis, we align multiple life histories around common points to identify shared patterns of information-seeking over the course of an illness. We analyze aggregate search patterns over time using a large set of relevant search terms organized into an ontology constructed specifically for this study. The results provide insights about the dynamics of information needs and highlight the promise of using search histories extracted from anonymized logs to better understand the attentional dynamics and information challenges that people may face when handling significant life events.

2. BACKGROUND AND RELATED WORK

2.1. Web search for health information

The Web is a central source of health-related information, with 81% of Americans using the Internet to find health information according to a recent survey [Fox and Duggan 2013]. Internet-based health information acquisition enables broader and faster information access. However, concerns have arisen about the quality and clarity of online health information [Cline and Haynes 2001] and that information can be misunderstood and misused [Benigeri and Pluye 2003], leading to problems such as the unnecessary escalation of medical concerns following queries about typically benign symptoms [White and Horvitz 2012]. Several researchers have sought to gain an understanding of health information-seeking on the Internet, using interviews and focus groups [Eysenbach and Kohler 2002; Peterson et al. 2003], surveys [Trotter and Morgan 2008], and more recently, large-scale analysis of search engine logs [Ayers and Kronenfeld 2007; Cartright et al. 2011; White and Horvitz 2013a].

2.2. Information-seeking by cancer patients

Information access plays an important role for cancer patients [Ziebland et al. 2004], most especially during treatment [Rutten et al. 2005], when decisions with difficult tradeoffs and unclear answers must be made. Almost all cancer patients want access to all relevant information [Gaston and Mitchell 2005], and doctors often under-estimate how much information patients need [Fallowfield 2001]. A majority of breast cancer patients prefer to have a role in decision making [Degner et al. 1997], and breast cancer patients who take an active role in decisions report having a higher quality of life than those who deferred decisions to others [Hack et al. 2006].

These reasons, coupled with the rise of the Web as a common information source for cancer patients [Satterlund et al. 2003], have led to further study of online information-seeking for cancer [Castleton et al. 2011; Helft 2012]. In a recent study, resonating with motivations of this paper, Ofran et al. [2012] analyzed cancer-related search engine queries to infer general patterns of cancer information seeking. The latter study relied on query volume rather than constructing and employing classifiers for identifying searchers experiencing a diagnosis and the likely timing of the diagnosis as we do.

Most related to the current study is our own recent paper [Paul et al. 2015], which employed the classifiers described here in a study of prostate cancer-related decision-making in search logs. However, that study focused specifically on search related to medical treatment options and relied heavily on manual annotation. In the current study, we more broadly analyze different categories of search, and use automated classifiers to align search histories over time.

2.3. Search log analysis

Our study involving large-scale search logs forms part of a more general body of work that has demonstrated that search query logs can be an effective source of data for learning about human behavior. Search logs have been used to study how people use search engines [White and Drucker 2007], to predict future search actions and interests [Lau and Horvitz 1999; Downey et al. 2007; Dupret and Piwowarski 2008], and to detect real-world events and activities [Richardson 2009].

In the health and medical domain, much research has demonstrated the ability to understand real-world activity from search data including the detection of influenza [Eysenbach 2006; Ginsberg et al. 2008; Santillana et al. 2014], norovirus [Desai et al. 2012], and dengue [Chan et al. 2011] outbreaks, the discovery of side effects of medications [Yom-Tov and Gabrilovich 2013; White et al. 2013], insights into healthcare

utilization [White and Horvitz 2013b], the study of dietary patterns [West et al. 2013; Kusmierczyk et al. 2015], and measuring the effectiveness of public health awareness campaigns [Glynn et al. 2011; Ayers et al. 2012].

Studies in the information retrieval (IR) community have also examined search tasks that span multiple sessions [Kotov et al. 2011], with a view to supporting task resumption over time.

2.4. Query and session classification

Many studies have used classifiers to label search queries and sessions, similar to our use of classifiers to identify relevant search histories. Classification is typically used to identify the *intent* of a query or session, such as whether the search activity is used to support a particular goal or task. A challenge in building such classifiers is obtaining quality training data. Most research relies on third-party annotations of existing log data, as we do in this study, [Rose and Levinson 2004; Cartright et al. 2011; Hassan et al. 2014; Raman et al. 2014], as well as external resources, such as labels from website directories [White et al. 2010; Broder et al. 2007], or through user studies [Guo and Agichtein 2010]. In-situ labeling methods are also feasible (e.g., using popup surveys) and can obtain accurate labels [Fox et al. 2005; Hassan et al. 2011]. However, these rely on the provision of labels at search time, which is intrusive and is also infeasible for the retrospective log analysis described in this article.

3. SURVEY ON CANCER-RELATED WEB ACTIVITY

To give additional context and motivation for our log-based analysis, we present results of a survey that we conducted asking a random sample of U.S. Microsoft employees about their Web activity and experience following a cancer diagnosis. We collected 867 anonymous responses using an internal Web-based survey system. We note that this sample is not representative of the general population, or even the Internet-using population, but it is a useful starting point.

We asked respondents if they or someone they know had been diagnosed with cancer in the past five years. 36.7% answered Yes, among whom a plurality said a parent was diagnosed (30.5%) followed by a non-immediate family member or relative (23.3%). 6.0% of respondents had been diagnosed personally. Of those who specified a type of cancer, breast cancer was the most common, with 13.5% of responses.

89.4% of respondents who were personally diagnosed reported that they had searched for information about breast cancer on the Web. This percentage is 76.8% for those with an immediate family member diagnosed, and 55.7% with another relative or friend diagnosed. We found that the likelihood of searching for information increases with the closeness of the searcher to the person diagnosed, as would be expected. These findings suggest that a substantial number of people search the Web regarding a recent cancer diagnosis. This serves as important motivation for our research and suggests that search logs may contain valuable insights about the intentions and actions associated with breast cancer searching over time.

Quality of Information. Although people commonly use the Web for cancer information, many find the information to be of poor quality. In total, 41.2% of respondents answered Yes to the question, "Did you find certain information or resources contradictory or confusing?"

We allowed for free-form responses to provide participants with an option to explain what was confusing, and separately to describe any conflicts that arose specifically between advice received from a physician and information they had found on the Web. These sample responses illustrate difficulties that people have with using the Web to access information about cancer and its treatment:

Table I. Percentage of survey respondents reporting a diagnosis with cancer who searched for various types of information, as well as the ratio of values of those diagnosed and all other respondents.

Category of search content	Searched	Ratio
Information about the type of cancer	100.0%	1.60
Information about cancer staging / grading	88.2%	1.47
Prognosis (survival rates or other statistics)	82.4%	1.09
Information about treatment options	76.5%	1.09
Side effects of treatment	70.6%	1.26
Information about the diagnostic process	58.8%	1.12
Explanations of a pathology report	52.9%	2.62
Advances in treatment and other research	52.9%	1.36
Healthcare providers	47.1%	2.06
Symptoms and signs of cancer or metastasis	41.2%	0.86
Info. about diet, exercise, and lifestyle issues	35.3%	1.17
Health insurance and financial issues	17.6%	1.48
Support groups and online communities	11.8%	0.76
Cancer awareness and outreach	5.9%	0.42
Stories from cancer survivors / celebrities	5.9%	0.30

— *I don't like checking the web—it's too depressing, confusing, overwhelming and contradictory. I like to follow a doctor's advice and go with it.*
— *Some of the websites out there can do more harm than good. The diagnosis is really devastating, the last thing you want is some idiot's opinion.*

Multiple responses said the information was "overwhelming". Another common complaint was there were few comprehensive sources of information, so one would have to read many different sources to form a complete picture. Others complained of difficulties discerning legitimate websites from lower quality sources. Information was also said to vary depending on whether the source endorsed Western or alternative medicine, as well as whether the source was from North America or Europe. Two respondents said they were explicitly told not to use the Web by their doctors, while another respondent said they worked with the doctor to identify a list of trustworthy websites.

Clearly, many people have found difficulties and dangers with using the Web as a resource, yet there is also a clear desire to seek information from the Web despite the challenges. Respondents expressed clear advantages of accessing information from the Web:

— *I found that doctors and nurses did not always sync in the information they provided. So I would validate or do further research on the net.*
— *Actually, the information I found confirmed and let me better understand what I have heard from the doctor. This was particularly important because the treatment was conducted abroad.*

The survey results suggest that information from the Web can be a useful complement to information provided by healthcare professionals. However, there are challenges with finding information that is reliable, comprehensive, and relevant to searchers. An important step toward evaluating and enhancing the value of cancer-related Web search is to understand the information needs and behavioral dynamics of search users. The log-based analysis presented in the remainder of this article provides rich insights into these issues, in greater depth than can be gleaned from survey responses alone.

Content of Interest. In order to understand *what* information is important to those diagnosed with cancer, we asked respondents to state the types of content that were searched (from a checkbox list where they could provide more than one response), e.g.,

information about prognosis or treatments. Table I shows the percentage of respondents who were personally diagnosed that searched for different categories of search content.

We calculated the ratio of the percentage of each category among those personally diagnosed to the percentage among all other respondents, shown in the right column of the table. We see that the diagnosed respondents were much more likely to search about their pathology reports (by a factor of 2.62), and also more likely to search for healthcare providers and insurance, information about cancer staging and grading, advances in treatment, and treatment side effects. Knowing these associations with diagnosed searchers can help inform our classification of search histories that are characteristic of those who are actually experiencing cancer, versus only possessing an interest in the topic.

4. DATA COLLECTION

4.1. Data Source

The primary source of behavioral data for this study is a proprietary data set composed of the anonymized logs of consenting users of a widely distributed Web browser add-on associated with the Microsoft Internet Explorer Web browser. The data set was gathered over an 18-month period from February 2012 through July 2013. It consists of billions of queries, issued by millions of searchers to the Bing search engine, represented as tuples including a unique user identifier, a timestamp for each query, and the text of the query issued. User location information in the logs is used for later comparisons between counts of searchers in the logs who are classified as having breast cancer with incidence rates provided by US federal agencies. We do not consider users' IP addresses directly, only geographic location information derived from them (city and state). All log entries resolving to the same town or city were assigned the same latitude and longitude. To remove variability caused by cultural and linguistic variation in search behavior, we only include log entries from the English-speaking United States locale.

We also used data provided under contract by the Web analytics company, comScore. comScore recruits millions of consumer panelists who give explicit permission to passively measure their online activities using monitoring software installed on their computers in exchange for incentives such as software, cash, and prizes. The demographic distribution of these users has been validated to be representative of the online population and projectable to the United States population [Fulgoni 2005]. The comScore data comprise unfiltered search queries on major Web search engines, collected over a two-year period from mid-2011 to mid-2013. Events in the logs contain text of queries, search result clicks, and the time that the events occur. Person identifiers in these logs are associated with the user's age and gender. We employ the comScore data in section 6.2.2, to compare the demographic distributions of cancer searchers in the data to known incidence rates.

4.2. Corpus Creation

Our goal is to analyze the content and timing of breast cancer-related search. The first task is to create a corpus of search histories that appear to refer to a breast cancer diagnosis with experiential search sessions. Given the terms of use under which the Bing log data were collected, user identifying information was removed from the logs at source. As such, we did not have a way to contact searchers directly to determine whether diagnosis occurred. We therefore performed manual labeling of the logs to generate data for training and evaluating our classifiers.

Table II. An example of episodic queries by a fictitious user consistent with many users in our dataset. In this example, Dec 12 would have been labeled as the day of diagnosis.

Time	Query
Nov 13 2013 7:40pm	feels like lump in breast
Dec 1 2013 11:21am	pain after biopsy
Dec 1 2013 11:31am	what happens after breast biopsy
Dec 9 2013 6:33pm	how often are breast lumps cancer
Dec 9 2013 6:45pm	does cancer make you thirsty
Dec 9 2013 6:49pm	how long does it take for biopsy results
Dec 12 2013 12:08pm	stage 2a breast cancer
Dec 12 2013 12:15pm	invasive ductal carcinoma
Dec 12 2013 12:17pm	poorly differentiated idc breast cancer
Dec 12 2013 12:29pm	breast cancer survival rate
Dec 12 2013 12:32pm	stage 2 breast cancer survival rate
Dec 12 2013 7:44pm	breast reconstruction surgery
Dec 12 2013 7:46pm	breast reconstruction after cancer
Dec 13 2013 8:05am	breast cancer treatment
Dec 13 2013 8:16am	recovering from breast cancer
Dec 15 2013 09:20am	breast cancer surgeon
Dec 15 2013 10:22am	full mastectomy
Dec 15 2013 10:23am	mastectomy pros and cons
Dec 15 2013 10:29am	do you need chemo after mastectomy

We began by collecting data from users who issued a query containing the string "breast cancer" in at least three separate search sessions and whose histories spanned at least 20 days. These criteria were used as heuristics for identifying searchers with enough data for our analysis. 138,306 users met these criteria. Then, we set out to manually tag searchers whose histories were consistent with a cancer diagnosis. We collected tags on 480 of these users drawn randomly from the larger set. The three co-authors[1] independently provided labels with two types of information:

(1) **DX classification:** We labeled whether the search history spans the time when a diagnosis (DX) of breast cancer has occurred. We first labeled search histories as whether (P) or not (N) the history showed a sustained focus of attention on breast cancer, relative to other medical searches. Searchers with many queries about many diseases (which may arise for example if the searcher is a medical professional) would be treated as negative instances for DX. Of the users who do have a sustained breast cancer focus, we labeled whether this focus of attention began during the search activity contained in the data (PP), or whether the focus of attention is strong throughout the entire history (PN). The latter is also treated as a negative instance because the time of the attentional shift to breast cancer (e.g., a new diagnosis) does not happen within the period of history included in the data. The positive examples have characteristics that are consistent with a patient (or loved ones searching on her behalf) who learned of a diagnosis during the search history, although in the absence of ground truth, we have no guarantees about how many of these users are grappling with a new cancer diagnosis. Even though we cannot construct a data set with guaranteed quality, we can at least filter out histories that do not plausibly express an experience of diagnosis.

(2) **DDX identification:** If a shift of focus of attention to breast cancer, consistent with a new diagnosis, was labeled to have occurred during the available period of search, we note the likely day of the diagnosis, referred to as DDX. Searchers would issue sets of searches over a period of days, resonating with a real-world sequence

[1] One of the authors (E.H.) has formal medical training and he guided the annotation process.

of queries on mammography (e.g., revealing in the logs that they had obtained information that a screening was suspicious and needed to be followed up), followed by biopsy, and, in many cases, onto searches on pathology and staging information. Table II gives an example search history. We chose the label to be consistent with the time when an actual patient would have learned of a diagnosis: the first day that search queries indicate a confirmation from laboratory results, per the specifics shared on pathology, stage, and grade as is often included by physicians in discussion and/or via a diagnostic report shared with patients at the time of diagnosis.

The annotators were shown all queries in sessions containing relevant terms (the terms in the term ontology described in the next section) as well as timestamp information (but not the query content, to respect privacy) of remaining sessions. Annotator 3 has a medical background (an MD) and provided guidelines on the criteria to apply during judging. Each searcher was assigned one of three labels described in (1) above, and in cases of ambiguity, annotators were asked to provide multiple labels in order of likelihood.

Annotators 2 and 3 each labeled disjoint sets of 150 searchers, while Annotator 1 labeled all 300. Annotators 1 and 2 agreed on the top choice of label on 69% of users ($\kappa = 0.51$). Annotators 1 and 3 agreed on 77% ($\kappa = 0.61$). Such scores are considered to represent fair to good agreement [Fleiss 1981]. A plurality (40%) of disagreements were between the PN and PP labels, while a minority (28%) of disagreements were on N versus PP. Disagreements on N versus PP were typically due to the presence of bursts in medical terminology that looked indicative of PP, but that were also suggestive of medical expertise and thus indicative of N.

Disagreements were resolved by taking a majority vote whenever the two annotators had included a label somewhere in the list, even if it was not the top choice. This accounted for 83% of users. To be conservative, the remaining users were assigned the most negative label (PN over PP and N over PN) that either annotator included in the list of possible labels, so as to avoid using ambiguous cases as positive examples. To increase the amount of training data, Annotator 1 also labeled an additional 180 searchers, for a total of 480. Again, for ambiguous users with multiple labels, the most negative label was selected. This annotator also revised the annotations after discussing some general disagreements with the other two annotators to improve consistency. Following this labeling procedure, 56% of the 480 searchers were labeled N, and 22% for both PN and PP.

Finally, 105 PP searchers were given DDX labels by two annotators. Annotators 1 and 2 agreed on the exact day in 46% of the searcher timelines, with an average disagreement of 15.3 days on the remaining searchers. Annotators 1 and 3 agreed on the exact day in 63% of the timelines, with an average disagreement of 5.7 days on the remaining searchers. The annotators discussed and resolved all disagreements larger than 7 days, which accounted for 15% of the searchers. Of the 31% of searcher timelines where disagreements were greater than 0 but less than or equal to 7 days, we automatically set the label as the later day of the two annotations, so that the label is more likely to fall on a day after the diagnosis had been officially confirmed with pathology reports, which could be difficult to identify and thus a source of disagreement.

We note that, while we do not have ground truth information about the searchers in our dataset, our methodology is consistent with previous work on query and session classification of searcher intent, in which data is annotated based on the inferences of the annotators, even though searchers cannot be contacted to verify the annotations [Rose and Levinson 2004; Cartright et al. 2011; Hassan et al. 2014; Raman et al. 2014].

Table III. Sample of ontology categories and terms. Spelling variants of the terms are also included, but not exhaustively shown in this table. ( ) indicates optional characters, { } indicates sets, and [ ] indicates ranges.

Category			
Level 1	**Level 2**	**Level 3**	**Terms (partial list)**
Cosmetic	Post-Surgery	Post-Surgery	{cosmetic,plastic} {surgery,surgeon}, prosthetic(s)
Cosmetic	Hair Loss	Hair Loss	wig(s), head {scarf,scarves,covering(s)}, hair (re)grow(th)
Description	Type	Cancer Type	DCIS, LCIS, IDC, ILC, lobular, ductal, in situ, inflammatory
Description	Staging/Grading	Staging/Grading	what stage, stages, staging, what grade, grades, grading
Description	Staging/Grading	Stage	early stage, stage {[0–4],zero–four,[I–IV]}({a,b,c})
Description	Staging/Grading	Grade	grade {[1–3],[I–III]}, {low,moderate,intermediate,high} grade
Diagnosis	Diagnosis	Diagnosis	diagnosis, diagnosed
Diagnosis	Diagnostics	Biopsy	biopsy, biopsies
Diagnosis	Screening	Mammagraphy	mammogram(s), mammography
Diagnosis	Screening	Ultrasound	ultrasound(s)
Lifestyle	Lifestyle	Diet	diet(s), eat(ing), food(s), vitamin(s), supplements, nutrition
Lifestyle	Lifestyle	Fitness	fitness, exercise(s), yoga
Professional	Healthcare	Provider	clinic(s), hospital(s), cancer center(s)
Professional	Healthcare	Doctor	doctor(s), physician(s)
Professional	Healthcare	Oncologist	oncologist(s)
Treatment	Treatment	Treatment	treatment(s), medication(s), meds
Treatment	Treatment	Side Effects	side effect(s)
Treatment	Chemotherapy	Chemotherapy	chemotherapy, chemo, cemo, kemo
Treatment	Chemotherapy	Side Effects	hair loss, hair fall(ing), {lose,losing} {my,your} hair

4.3. Term Ontology

We manually created a large ontology of health- and cancer-related keywords and phrases. The purpose of the lexicon is twofold: the ontology categories and terms can be used as features (described below) in learning classifiers for predicting whether and when a searcher may have been diagnosed with cancer, and the ontology will assist in our analysis of searcher histories.

We created a three-level hierarchy of categories for a wide variety of topics that cancer patients might search for via inspection of sessions, review of informational resources for breast cancer, and reflection about the needs of newly diagnosed patients. Classes include cancer diagnostics, healthcare, treatment, information on types and causes of cancer, coping and social support, and many others. The ontology was primarily created by one author (M.P.) after reviewing the relevant literature and a sample of search logs, and the categorization was iteratively refined following discussions with the co-authors.

Table III shows a sample of categories and the associated terms. The full ontology contains 19 top-level, 47 mid-level, and 127 bottom-level categories covering 1963 terms.[2] The ontology includes some constraints such that terms are only considered part of the ontology if they co-occur with other terms. For example, the term “mass” is highly ambiguous and is only considered a relevant term if it occurs in the same query with terms about cancer, anatomy, healthcare, or diagnosis. Occasionally we created constraints that a term must *not* co-occur with other terms (for example “ribs” is not considered an anatomical reference if the term co-occurs with “bbq” or “pork”).

A category serving an important role in our feature set (described in Section 5) is named EXPERT. This category contains terms that we found to be commonly used by searchers who show strong signs of being healthcare professionals doing exploration on breast cancer, rather than patients, e.g., users who appear to be or are studying to become nurses and doctors. The motivation for creating this category is that it could be a useful feature for identifying such searchers, which comprise a significant portion of false instances in the corpus.

[2]The ontology is available at: http://research.microsoft.com/apps/pubs/?id=260512

We also include categories called CANCER (under DISEASES) and BREAST (under ANATOMY) which simply contain the strings "cancer" and "breast" and spelling variants thereof.

Some of the term categories were created by accessing lists appearing in external resources or by generating phrases from patterns that we specified. We created a large set of strings to match geographic locations in the United States, called the GEOGRAPHIC category, populated from a gazetteer from the U.S. Geological Survey (`http://geonames.usgs.gov`) containing 185,800 cities. The motivation is that cancer- or healthcare-related queries are more likely to be experiential rather than exploratory if they appear with a geographic location. For example, someone searching "oncologist" might simply be looking up the definition of this word, while "oncologist in seattle" suggests an intention to visit an oncologist [White and Horvitz 2010].

We also created a set of strings that capture the mention of the age of a person in queries (the AGE category, including strings such as "at 55", "age 55", and "55 year(s) old", "55 y/o", as well as the phrases "in {my/your/her} 50s" for all decades from 20s through 90s. This category has the same motivation as GEOGRAPHIC—users expending effort to modify general search terms with personal details, e.g., issuing the query "breast cancer at age 55," are more likely to be experiencing cancer than someone who simply searches "breast cancer".

4.3.1. Symptoms and Diseases. We created categories with symptom and disease words and phrases, composed by White and Horvitz [White and Horvitz 2009]. After removing ambiguous terms, our SYMPTOMS category contained 62 terms, and after removing various cancers (which we distinguish from other diseases), our DISEASES category contained 249 terms. We created a separate category with 109 types of cancer listed by the U.S. National Cancer Institute (NCI) (`cancer.gov`). Also using a list from NCI, we created a DRUGS category containing 57 brand and generic names of drugs approved for breast cancer in the United States.

We also downloaded a list of 109 types of cancer from the U.S. National Cancer Institute,[3] which we split into two groups: COMORBIDCANCERS which contains the cancers that are the most common sites[4] for breast cancer to spread—bone, brain, liver, and lung—and OTHERCANCERS for the remainder. (These disease and cancer categories are children of a top-level DISEASES category.)

4.3.2. Breast Cancer Medications. We accessed a list of medications from the National Cancer Institute[5] of drugs approved by the U.S. government for breast cancer treatment. The list, which forms the DRUGS category (under TREATMENT), contains 57 terms including both brand names and generic chemical names.

4.3.3. Celebrities with Cancer. We included a CELEBRITIES category which includes a list of well-known people who have or had breast cancer, scraped from the English Wikipedia article, "List of breast cancer patients by survival status." We processed the list to add alternate spellings of names: nicknames instead of real names, with and without middle names. This class includes 493 name strings.

4.3.4. Health Insurance. We created an INSURANCE category which contains 42 names of health insurance companies, taken from the data set compiled by [Paul et al. 2013], among other terms.

[3] http://www.cancer.gov/cancertopics/types/alphalist
[4] http://www.cancer.gov/cancertopics/factsheet/Sites-Types/metastatic
[5] http://www.cancer.gov/cancertopics/druginfo/breastcancer

4.4. Experiential Categories

We identified a subset of ontology categories that capture terms used in information-seeking that are particularly evocative for experiential searching rather than exploratory searching. We designate the following subset of ontology categories as *experiential* categories:

— COSMETIC,
— DESCRIPTION,
— DIAGNOSIS → {DIAGNOSTICS, SCREENING},
— INVOLVEMENT → {LYMPHNODES, METASTASIS},
— PATHOLOGY,
— TREATMENT → {CHEMOTHERAPY, RADIATION, DRUGS},
— STATISTICS → PROGNOSIS.

The arrow symbols indicate deeper levels in the ontology. For example, DIAGNOSIS → DIAGNOSTICS is a level-2 ontology category, as seen in Table III.

The appearance of terms in these categories more strongly suggests that a searcher is personally experiencing cancer than seeing terms from such categories as PROFESSIONAL and LIFESTYLE, which someone may search for a wider variety of reasons. For example, there is no strong reason to believe that someone searching "hospital" has breast cancer without additional evidence, but it is more likely that someone searching "tamoxifen," a medication used to treat certain types of breast cancer, has or knows someone who has breast cancer.

The experiential designation of these categories is used when constructing features for experiential classification, described in the next section.

5. DIAGNOSIS CLASSIFICATION

5.1. DX Classifier (Searcher Model)

A key sub-goal of this work is to determine whether the focus of attention and pattern of queries over time is consistent with the searcher having breast cancer, and, if so, whether the diagnosis appears to have occurred during the observation period so as to allow for alignment of multiple life histories. This subsection describes the features used to classify searchers into these categories.

5.1.1. Query Features. We extracted the set of terms and categories in Table III appearing in each query within a searcher's history. Counts of these terms and categories constitute standard lexical "bag of words" features with additional features that group terms into more general categories. We also created features of conjunctions of up to three lexical features (indicating counts of the number of queries and sessions within which these features co-occurred).

In addition to the terms and categories, the following binary features, which may indicate that the search is experiential, are extracted from each query:

— Does the query contain a first or second person pronoun ("I", "me", "you") or possessive pronoun ("my", "your"). This could indicate that the searcher is searching about her own experience.
— Does the query contain a phrase that indicates the searcher is personally experiencing cancer; specifically if the query contains one of the strings "i have", "you have", "my", "your", "i was diagnosed", "i have been diagnosed", "recently diagnosed with", and other similar phrases followed by a cancer phrase. Cancer phrases can be any string in the CANCERTYPE, STAGE, or GRADE classes (Table III), or the strings "cancer" or "breast cancer".

— Does the query contain a phrase that indicates a personal healthcare experience; specifically if the query contains “my”, “your”, “{ask,tell,talk to} {the,a}”, or “question(s) for(a, the)” followed by any phrase in the HEALTHCARE class.
— Does the query begin with a question word (e.g., “what”, “why”).

5.1.2. Volume Pattern Features. We included a variety of features that are aggregated across searchers’ entire histories, including search volume from sessions that do not contain ontology terms, as described in Section 4.2. These features often includes counts of categories from Table III, but we also created features that count only the subset of these categories that are specific to cancer (i.e., we exclude categories pertaining to general healthcare or other life matters). We will describe these two different sets of categories as the “ontology/cancer” categories. We extract the following features from a searcher’s full history:

— Percentage of the searcher’s sessions which contain ontology/cancer terms or the term “cancer” (three features).
— Of the sessions that contain ontology terms, the percentages of those sessions/queries which contain experiential terms, CANCER, and the EXPERT category.
— Number of different ontology/cancer categories ever searched, and the distribution over categories.
— Distribution over the lowest-level ontology categories.That is, each class is associated with a feature whose value is the number of queries containing this class, normalized to sum to one across all classes.
— Ratios of the average length (in number of queries) of sessions with ontology/cancer terms or “cancer” over the average length of all sessions, and similar ratios for the number of such sessions per day.
— Ratios of the number of ontology, experiential, and CANCER sessions per day over the number of all sessions per day.
— Number of different disease terms or symptom terms ever searched for. Searchers who search a large number of different diseases are often healthcare professionals or anxious searchers who are searching in an exploratory rather than experiential manner. Similarly, we have a feature for the number of different SYMPTOMS terms ever searched.

5.1.3. Temporal Pattern Features. Finally, we include various features that extract temporal patterns from search histories. Features that characterize temporal patterns in search histories could be informative as searchers experiencing the illness follow certain timelines, as illustrated in Table II, and higher density of cancer-related queries is also suggestive of experiential usage.

— Number of days from the beginning of the searcher’s search activity to the first query containing an ontology/cancer term or “cancer”, and similarly from the last such query to the end of activity. If the searcher never searches for an experiential or CANCER category, then this is set to the total number of days of searcher history.
— Average, minimum, and maximum number of days between sessions containing an ontology/cancer terms or “cancer”, normalized by the average/min/max number of days between all sessions of the searcher, and the average number of 7+ or 30+ day gaps between such sessions.
— Number of times gaps of 7+and 30+ days exist between consecutive sessions containing ontology, experiential, and CANCER terms, normalized by the number of 7- or 30-day gaps that exist between all consecutive sessions in the searcher’s history.

— Largest number of sessions containing ontology/cancer terms or "cancer" that appear within a 7-day and 30-day periods in the search history, normalized by the searcher's average number of sessions per day.
— Shortest number of days that span 5 sessions and 10 sessions containing ontology, experiential, and CANCER terms.
— Difference in number of days between the first query for one ontology category and the first query for another, for all pairs of top-level categories, as well as the average difference between all such queries. For example, if the first query for DIAGNOSTICS is on March 1 and the first query for DRUGS is March 20, the difference between DIAGNOSTICS and DRUGS is -19 days. This models our intuition about the chronological sequencing of searches in light of real world events. Specifically, that afflicted searchers will tend to search for categories in a certain order; e.g., searches about diagnostics precede searches about treatment options. We set the feature values to zero for pairs containing classes the searcher never searched.

5.2. Time of Diagnosis Classifier (Timeline Model)

A second critical task is to predict the point in time when it appears a searcher receives a likely cancer diagnosis. Let $D_u = \{1, 2, \ldots, m_u\}$ be the set of days of searcher u's search history, where the first day of search activity is indexed as day 1, and m_u is the number of days spanned by the searcher's history. We create a feature vector for each day $d \in D_u$ and use these features to predict whether d is the day the searcher appears to have first learned of the cancer diagnosis.

Many of the features in this model compare searchers' histories before and after a day d. We will refer to $\{1, 2, \ldots, d-1\}$ as the before-d set, and $\{d+1, d+2, \ldots, m_u\}$ as the after-d set. We also consider the 10 days before and 10 days after d ($\{d-10, d-9, \ldots, d-1\}$ and $\{d+1, d+2, \ldots, d+10\}$), referred to as the 10-before-d and 10-after-d sets. Finally, we created a set of features that compare only day d to the rest of the searcher history, $D_u - \{d\}$.

5.2.1. Before-and-After Features. For this model, we created two separate sets of the lexical count features (query/session/user counts of the ontology terms and categories; Section 5.1.1) extracted from the two time intervals, before-d and after-d, as well as d itself.

For each volume and temporal feature described in Sections 5.1.2–5.1.3, we created a feature whose value is the ratio of the values of the feature in the two time intervals. For example, if 10% of a searcher's sessions contain ontology terms in the before-d set, and 40% of the searcher's sessions contain ontology terms in the after-d set, then we will have a feature for the ratio of these two percentages with value 0.25. We exclude the following features, which were only designed for the DX classifier: the percentage of EXPERT sessions (Section 5.1.2), the number of disease and symptom terms (Section 5.1.2), the number of days to/from the beginning/end the searcher history to the first/last ontology queries (Section 5.1.3), and the differences in time between pairs of ontology queries (Section 5.1.3).

We created similar before-and-after features extracted from the 10-before-d and 10-after-d time intervals. Since these time intervals are short, we do not include any of the temporal features of Section 5.1.3 for these 10-day intervals.

5.2.2. Day-Of Features. We also include the same query features (Section 5.1.1) described in the previous subsection extracted only from day d, as well as the percentage features from the first two items in the list of volume features in Section 5.1.2. We normalize the values of these features by the values of these features extracted from the rest of the searcher history, $D_u - \{d\}$. As with the 10-day intervals, we do not use temporal features.

We also created the following additional features:

— Number of queries containing ontology terms from Table III searched on this day, normalized by the average number of ontology queries searched per day (averaged across the other days of the searcher's history). A similar feature is included for experiential terms and the CANCER category. Similar features are included which are normalized by the maximum rather than average number of queries per day.
— Number of different ontology categories and experiential categories searched on this day, normalized by the average and maximum number of different ontology categories searched per day.
— Number of different ontology categories and experiential categories that were searched for the first time on this day.
— For each ontology category, a binary feature indicating whether this category was searched for the first time on this day.
— For each ontology category, the number of days since the previous occurrence of this category, and the number of days to the next occurrence. For classes which do not appear before or after day d, we set the value to the number of days in the searcher's history, so that this value is larger than the largest possible value the feature could take if the class were present.

5.2.3. Boundary Indicators. We included binary features to indicate whether $d \leq 10$ and whether $d > m_u - 10$. The estimates of the before-and-after features are noisy when they are close to the beginning or end of the searcher's query timeline, and these boundary features will fire to inform the classifier that this is the situation.

5.3. Training and Prediction

We used Multiple Additive Regression Trees (MART) [Friedman et al. 2000] for both the classification tasks, which was the best-performing classification method during development (compared to support vector machines and logistic regression). MART uses gradient tree boosting methods for regression and classification. Advantages of employing MART include model interpretability and robustness against noisy labels and missing values.

5.3.1. Searcher Prediction. We trained two DX classifiers, one to predict whether the searcher appears to have breast cancer or not (N vs P*), and another trained on the subset of searchers with cancer to distinguish whether it was recent (PN vs PP). The joint probability used for prediction is the product of the two probabilities produced by these classifiers. While this performed nearly identically to a single classifier trained on PP vs *N, this two-classifier approach gives us flexibility if we want to also identify all searchers with breast cancer.

5.3.2. Timeline Prediction. For simplicity, we train the DDX classifiers to assume each day is independent, and simply use the day of diagnosis (or others; see below) as a positive label and others as negative. In early experiments, we found that training a timeline classifier with a "single day" labeling scheme (with one positive instance per searcher) did not perform very well, perhaps due to the sparsity of positive instances, so we experimented with other labeling schemes:

— **Window around diagnosis:** The 7 days before and after the day of diagnosis are also labeled with 1. Thus, this classifier identifies the 15-day window surrounding the day of diagnosis, which we hypothesize to be an easier task than identifying the exact day.

— **Day within window:** This classifier tries to identify the exact day, but only within the 15-day window containing the correct day. Days further away are not used as training data for this classifier.
— **Has been diagnosed:** All days before the day of diagnosis are labeled with 0 and all days including and after the day of diagnosis are labeled with 1. That is, this classifier tries to predict at any point in time whether the searcher has already been diagnosed.

We can combine these three classifiers as follows. Let Θ_{AD}, Θ_{WW}, Θ_{HB} respectively denote the parameters of the three classifiers. The first two classifiers can be combined in a "coarse-to-fine" fashion in which we consider the probability that a day d is in the 15-day window modeled by the AD classifier, along with the WW probability that d is the day, conditioned on d being within the 15-day window. The probability that d is within the 15-day window is the sum of the probabilities of all 15-day windows that contain d, where the probability of each window is proportional to the product of the probability of each individual day in the window; i.e.:

$$P(d \text{ is in window}|\Theta_{AD}) = \sum_{i=d-14}^{0} P(win = \{i, \ldots, i+14\}) \tag{1}$$

where $P(win = \{d, \ldots, d+14\}) \propto \prod_{i=d}^{d+14} P(i = 1|\Theta_{AD})$.

The joint coarse-to-fine probability is:

$$P(d = 1|d \text{ is in window}, \Theta_{WW})P(d \text{ is in window}|\Theta_{AD}). \tag{2}$$

Finally, we multiply this by the probability that the previous day has label 0 under the HB classifier and the following day has label 1: $P(d-1 = 0|\Theta_{HB})P(d+1 = 1|\Theta_{HB})$.

This combined prediction scheme gave the best performance in early-stage development, so we used this as the prediction rule that we use in our final experiments.

5.4. Baseline Models

We also experimented with simple heuristic methods for DX and DDX classification. These simple methods will be used for baseline comparison when evaluating our classifiers. The goal is not to create highly-performant baselines, but rather to compare the classifier performance with simple heuristics that one could implement with little effort, in order to contextualize the results of our classifiers which require significant effort to construct.

5.4.1. Searcher Prediction. We experimented with two heuristics for identifying a searcher as positive for DX. First, we labeled a searcher as positive if a high percentage (above some threshold, tuned for performance in the experiments) of the searcher's queries fall into one of the experiential categories described in Section 4.4. We attained better performance when focusing on experiential categories rather than all ontology categories, so we used this constraint. Second, we labeled a searcher as positive if the number of different experiential categories searched was above a threshold. These two thresholds are prediction parameters that affect precision and recall, similar to classifier confidence. The motivation behind both of these heuristics is that they indicate that a searcher has a stronger than average interest in topics related to experiencing breast cancer.

5.4.2. Timeline Prediction. As a heuristic for identifying the DDX day, we selected the first day such that the number of queries containing experiential categories (Section 4.4) was above some threshold. The threshold can be adjusted to trade off precision and recall. As with the DX baseline, we found that we achieved better performance

Table IV. Cross-validation performance with 95% confidence intervals for the DX classifiers.

Features	Max F1	Rec@P90	Pre@R25
Lexical Only	74.6 ± 2.1	30.7 ± 6.3	90.6 ± 4.0
Lex. + Conj.	75.8 ± 1.9	35.8 ± 3.6	91.3 ± 5.1
Lex. + Query	75.1 ± 1.3	29.2 ± 6.0	91.5 ± 4.2
Lex. + Temporal	75.9 ± 2.6	39.6 ± 6.3	90.6 ± 2.2
Lex. + Volume	75.4 ± 0.4	25.0 ± 3.3	89.1 ± 4.6
Full Model	$\mathbf{76.5 \pm 2.7}$	$\mathbf{39.6 \pm 4.4}$	$\mathbf{94.3 \pm 2.3}$
Baseline-Queries	53.5 ± 8.2	0.0 ± 0.0	48.2 ± 11.5
Baseline-Classes	54.7 ± 8.1	0.0 ± 0.0	56.1 ± 12.4

Table V. Cross-validation performance with 95% confidence intervals for the DDX classifiers.

Features	0-Day Acc.	7-Day Acc.	15-Day Acc.
	100% Recall		
Lexical Only	38.2 ± 4.5	73.5 ± 1.3	85.8 ± 0.9
Lex. + Conj.	44.1 ± 3.3	75.4 ± 1.0	86.6 ± 1.4
Lex. + Day-Of	$\mathbf{45.8 \pm 0.9}$	74.4 ± 1.2	85.8 ± 0.9
Lex. + Query	41.1 ± 6.0	$\mathbf{76.3 \pm 2.8}$	$\mathbf{89.5 \pm 2.2}$
Lex. + Temp.	38.9 ± 6.4	71.5 ± 3.8	84.8 ± 1.1
Lex. + Volume	38.5 ± 4.5	71.5 ± 3.8	84.8 ± 1.1
Full Model	42.5 ± 4.7	72.2 ± 0.5	88.5 ± 0.3
Baseline-First	2.9 ± 2.8	7.9 ± 3.7	17.7 ± 6.1
Baseline-Max	2.0 ± 2.5	6.9 ± 4.8	12.9 ± 5.6
	25% Recall		
Lexical Only	59.0 ± 2.6	83.0 ± 0.6	90.4 ± 0.1
Lex. + Conj.	70.2 ± 6.1	92.5 ± 1.6	95.1 ± 1.0
Lex. + Day-Of	$\mathbf{76.7 \pm 4.8}$	93.8 ± 1.3	94.9 ± 1.0
Lex. + Query	64.4 ± 3.9	86.1 ± 2.9	94.2 ± 1.2
Lex. + Temp.	62.3 ± 7.8	86.5 ± 2.8	95.8 ± 0.9
Lex. + Volume	59.2 ± 1.5	$\mathbf{92.5 \pm 1.5}$	96.8 ± 0.7
Full Model	72.8 ± 8.2	90.2 ± 4.2	$\mathbf{99.0 \pm 0.2}$
Baseline-First	3.2 ± 3.1	16.2 ± 10.1	31.4 ± 10.4

when using only the experiential categories rather than all ontology categories. In addition to selecting the first day with a high number of experiential categories, we also experimented with selecting the day with the highest number of experiential queries. To break ties, the earliest day with the highest volume is selected. This heuristic has 100% recall, since it can be applied to all users. The motivation for these heuristics is that we observed during annotation that the DDX point coincides with a spike in related search activity.

6. EXPERIMENTAL EVALUATION

6.1. Classifier Validation

We evaluated the DX and DDX classifiers by performing 10-fold cross-validation across the 480 annotated searchers (for DX) and the 105 PP searchers (for DDX). For the DX classifier, we measured the maximum F1 score reached at all prediction thresholds (i.e., the threshold at which the classification probability is considered positive for an instance), the recall at 90% precision (Rec@P90), and the precision at 25% recall (Pre@R25). For the DDX classifier, we measured the accuracy at x days: the percentage of searchers whose predicted day was $\leq x$ of the correct day, for $x \in \{0, 7, 15\}$, at both 100% recall and 25% recall. The reported metrics are the average result across 10 folds.

To measure the improvement provided with inclusion of a variety of features, Tables IV and V compare the performance of the full models described in Section 5 with the

performance of baseline models that use only lexical features (the terms and ontology categories in the search history). To evaluate the gain when including different sets of features, we showed results for the lexical baseline feature set augmented with conjunctions of features (Section 5.1.1), query features (Section 5.1.1), volume features (Section 5.1.2), temporal features (Section 5.1.3), and day-of features for the DDX classifier (Section 5.2.2).

We see that the full DX model performs better along several metrics than the lexical baseline, though not by a significant amount. The biggest difference is in Rec@90, which is significant at the 85% level. The full DDX model is significantly better than the baseline at 15-day accuracy and all three metrics at 25% recall. (We found the DDX prediction errors to be largely symmetric, with 52% of the prediction errors being too early rather than too late.) Comparing the individual feature sets, the addition of conjunctions and temporal features yield the biggest gains for DX classification, while the day-of features seem to most improve performance for DDX classification.

In addition to the classifier models, Tables IV–V show results for the simple heuristic methods described in Section 5.4. We found that these simple heuristics—selecting users who searched for experiential terms (DX), and identifying days with high volumes of experiential search (DDX)—gave poor performance. The poor performance of the DDX heuristic was surprising based on our observations during annotation, but this result shows that the context and timing of search activity is important. Comparing the different baseline heuristics, we see that the number of different experiential categories (Baseline-Classes) is a slightly better heuristic than the number of queries (Baseline-Queries) for DX classification. For DDX classification, the first day with a high volume of experiential queries (Baseline-First) gives better performance than picking the day with the highest volume (Baseline-Max).

6.1.1. Ontology Category Comparison. We also performed ablation experiments in which we measured the performance of classifiers trained after removing one high-level category from the term ontology before computing features, so as to gauge the importance of each of the feature categories. We found that the DIAGNOSIS category is the most informative for the DX classifier, which upon removal resulted in the lowest scores in two metrics, F1 (73.5%) and Rec@90 (31.2%, significant at the 90% level). The DESCRIPTION category appears to be the most important for the DDX classifier, whose removal resulted in the lowest score in all six metrics (five significant at the 95% level), with exact-day accuracies at 32% (down from 43%) and 52% (down from 73%) at 25% recall.

6.1.2. Prediction Rule Comparison. We compared the performance using alternative prediction rules, described in Section 5.3. The differences in performance were not as large as during development, though we found that the joint DDX model performed significantly better than a classifier trained using a "single-day" labeling at both 15-day accuracy metrics (86.6% and 97.2%). We also found that removing the "has been diagnosed" predictions also significantly worsens accuracies at the 15-day level (86.5% and 95.6%). The single-day classifier seems more brittle and made larger errors than the final classifier, with the three largest errors of 260, 123, and 95 days (average 9.5 days), compared to 123, 85, and 48 (average 7.3). There were no significant differences between the two DX classifiers.

6.2. Alignment with Incidence Rates

To further evaluate our classification of newly diagnosed searchers, we compared the geographic and demographic attributes of searchers to breast cancer incidence rates available from the U.S. government.

6.2.1. Geography. We compared the geographic attributes of searchers in our data set to 2009 (the most recent year available) age-adjusted per-state breast cancer incidence rates from the National Cancer Institute (NCI) and Centers for Disease Control and Prevention (CDC).[6] These statistics include 49 U.S. states (data from the state of Wisconsin is suppressed by law). Incidence rates are given per 100,000 persons, with a mean of 123.1 and standard deviation of 8.0 across the 49 states. We measured the Pearson correlation between these incidence rates and the number of searchers in our data set from each state normalized by the total number of searchers from each state in the query log data.

We found that counts on the set of 138K searchers who searched "breast cancer" at least three times is uncorrelated with the state incidence rates (r=0.036). However, we found that counts generated for the subset of searchers assigned a high probability of recent diagnosis with breast cancer (via the DX classifier) are significantly correlated. The 5625 searchers with probability $\geq$0.5 have a positive correlation of r=0.348 (p=0.014). In contrast, the 5700 lowest-probability searchers have a correlation of -0.052.

That the highest probability searchers are much more strongly correlated with ground truth incidence rates (a tenfold increase at the maximum) provides evidence that our DX classifier may indeed be identifying recently diagnosed patients more accurately than our large baseline set of 138K searchers. This demonstrates the value of the additional features and modeling performed by our DX classifier.

6.2.2. Age and Gender. We also compared the incidence rates among demographic groups to our comScore data set (Section 4.1), which acquires and provides each searcher's age and gender if known. The NCI data we compared to are age-adjusted U.S. incidence rates from 2006–2010.[7] The comScore data set is much smaller than our primary dataset: there are 804 users who searched "breast cancer" three times, and 15 searchers with classifier probability $\geq$0.5 of being recently diagnosed.

In 2006–2010, breast cancer incidence was 103.2 times more likely in women than men, so we would expect the bulk of newly diagnosed searchers in our data to be female. Indeed, 70.0% of the 790 searchers are female, compared to 49.7% in the entire comScore data. This percentage increases within subsets of high-probability searchers: if we consider the top k searchers, we find a high point of 88.9% female among the top k=18 searchers.

Breast cancer incidence was 5.7 times higher for people aged 65+, so similarly we expect to see a higher proportion of elderly searchers in the set of DX searchers. Only 3.6% of comScore searchers are aged 65+. This increases to 5.4% in the set of 762 searchers, and this increases even further when considering high-probability searchers: a high point of 22.2% of searchers within the top k=18 are aged 65+.

With both age and gender, the demographic distribution within high-probability searchers is significantly closer to the ground truth (female and elderly) than baseline levels. We note that a breast cancer diagnosis will trigger searches from multiple people apart from the patient, such as family members, so we would not expect the demographic distribution of DX users in our data to match incidence rates exactly.

7. ANALYSIS OF LIFE HISTORIES

In the previous section, we demonstrated the feasibility and reliability of predicting which searchers have recently experienced a breast cancer diagnosis. An important part of our research is to understand the dynamics of such searchers' interests. We

[6] https://nccd.cdc.gov/uscs/cancersbystateandregion.aspx

[7] http://seer.cancer.gov/csr/1975_2010/browse_csr.php?section=4&page=sect_04_table.12.html

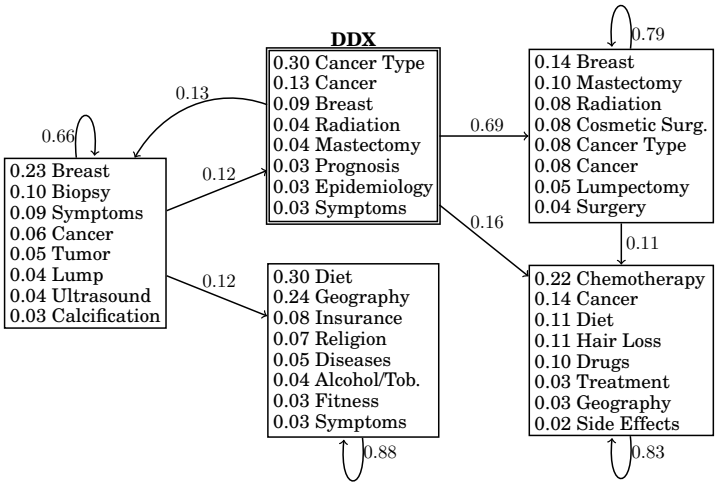

Fig. 1. Five-state HMM inferred from searcher timelines. A special state was reserved for the inferred day of diagnosis (DDX). Each box shows the highest probability ontology categories for that state, and edges between boxes indicate transition probabilities. Only edges with transition probability $\geq$0.1 are shown.

thus seek to analyze and characterize search queries across time, in terms of multiple episodes marked by shifts on focus of attention that have canonical timing characteristics. For example, breast cancer searchers often go through distinct information-seeking episodes, such as searches about suspicious symptoms followed by searches about cancer diagnosis later followed by searches about cancer treatment.

Our analysis centers around the identification and use of **pivot points**: points in time (at the granularity of one day) at which the searcher exhibits a particular shift in focus of attention with queries. We can understand general patterns of episodic search by *aligning* thousands of search histories around various pivot points, and analyzing the aggregate query volume at points in time with respect to each pivot.

The key pivot point is DDX, introduced in Section 4.2, at which a searcher's focus of attention shifts heavily toward breast cancer. While there may be some breast cancer search prior to DDX, this point marks a major shift in focus that is with the characteristics of a searcher who had just learned of a breast cancer diagnosis.

We define other pivot points using the following policies:

— **Screening and diagnostic workup:** The first point in time that a user searches for terms related to diagnostic screening technologies (i.e., mammography, ultrasound, CT scans) *prior to* DDX.
— **Surgery:** The first point in time that a user searches for terms related to surgery (including lumpectomy and mastectomy) *after* DDX.
— **Chemotherapy:** The first point in time that a user searches for terms related to chemotherapy *after* the surgery pivot point. This ordering is chosen because in cancer treatment, chemotherapy most often occurs after surgery [Vandergrift et al. 2013].

We base the determination of whether a pivot point is reached on the appearance of search term in the corresponding ontology entry, as described in Section 4.3.

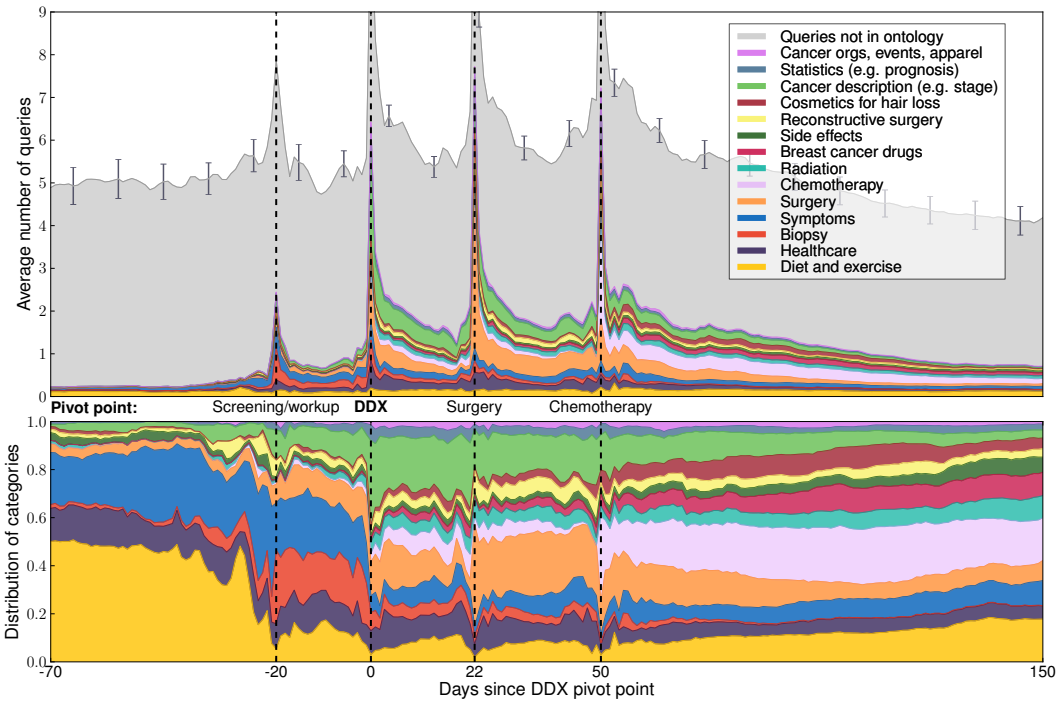

Fig. 2. Raw (above) and normalized (below) average number of queries per day for different search categories. The day on the x-axis is with respect to the pivot point, while the y-axis value is averaged between the values of the two surrounding pivot points. Standard error of the unsmoothed values is shown for the topmost curve. Details are given in Section 7.2.

7.1. Inferring Episodes with an HMM

To help visualize and understand how cancer-related search goals evolve over time, we applied a hidden Markov model (HMM) to the ontology categories within the searcher timelines. We treat each day of a searcher's timeline as a time step associated with one state of the HMM, and all of the low-level ontology categories that are searched on that day are considered emissions at that time step. To do this, we used the *block HMM* described in [Paul 2012], a type of HMM that models multiple emissions at each time step. This model also includes a separate distribution for background noise, to help filter emissions which are prevalent across all states.

We modified the HMM to include a special state for DDX. This day was constrained to this state, and all other days were constrained to the remaining states. We modeled the timelines of 558 searchers that were classified at thresholds estimated to have 90% diagnosis precision and timeline accuracies of 74% and 88% within 0 and 7 days.

Figure 1 shows the parameters estimated from an HMM with five states. The figure shows the eight most probable ontology categories in each state, with edges indicating transition probabilities between states. The most probable category associated with the day of diagnosis is CANCERTYPE, which contains terms describing specific types of breast cancer. Other high probability terms in this state include treatment options and searches about prognosis and other statistics. The state most likely to transition into the DDX state is shown on the left and appears to be associated with the diagnostic process, with terms about biopsy, screenings, and symptoms. This appears to correspond well to a search episode between the screening/workup pivot point and

DDX. The DDX state is most likely to transition to the two states shown on the right which are both associated with various treatments. The top right state is more likely to follow DDX and contains terms related to more immediate treatment solutions, including surgical procedures, while the latter is terms related to longer-term treatment like chemotherapy and side effects. These two states appear to represent episodes of search that are expected to surround the surgery and chemotherapy pivot points.

7.2. Aggregate Timeline

We aligned the searcher timelines around each inferred DDX point and the three pivot points described above by computing the average query volume at various points in time since the pivot point. For a pivot point p, d_p is the number of days since the pivot point, with $d_p = 0$ on the day and $d_p < 0$ for days before that point.

Figure 2 shows the query volume by ontology category over time, aligned around DDX and the three other pivot points. The query volume (top) is reported for various ontology categories as well as other queries outside the ontology. The lower visualization shows the same volume normalized to sum to 1 among the ontology categories.

The pivot points are positioned based on their average distance from each other for all of the searchers. The first workup searches occur an average of 20 days before DDX, the first surgery searches (excluding those on DDX) occur 22 days after DDX on average, and the first chemotherapy searches occur 28 days after the first surgery searches. For comparison to the timing of true cancer patients, recent studies have found a median time of 29 days between suspicious mammograms and diagnosis [Pérez-Stable et al. 2013] and mean times from diagnosis to surgery of 5.6 weeks and surgery to chemotherapy as 6.3 weeks in the U.S. [Vandergrift et al. 2013]. The average times in our data are likely shorter because, for example, people will search for treatment before the treatment actually begins.

When considering only a single pivot point, we simply plot the volume at each point d_p. When visualizing volume across multiple pivots, there are regions that include volume measurements from two different pivot points: for example, between the screening and DDX points, $d_{\text{screening}} = 3$ and $d_{\text{DDX}} = -17$ correspond to the same point on the x-axis. The volume at such points is measured as an average of the volume at that point from the two surrounding pivots, weighted by the distance from the pivots. For example, the volume 17 days before DDX is given as $\frac{3}{20}$ the volume at $d_{\text{DDX}} = -17$ and $\frac{17}{20}$ the volume at $d_{\text{screening}} = 3$. The weighting is uniform at the halfway point between two pivots. The motivation for this weighted scheme is so that points most immediately before and after a pivot are more heavily represented by the volume around the nearby pivot.

We gathered statistics from a larger number of searchers for the studies with alignments; we used the DDX classifier with 100% recall, which is highly accurate within two weeks. These figures are generated using the set of 1700 searchers estimated with 90% precision to be recently diagnosed. Not all search histories span all points in time, and fewer than a hundred searchers are represented at 365 days before and after the day of diagnosis. These plots (and all others in this subsection) are smoothed by taking a uniform average with days $d_p \pm |d_p|/5$ for each pivot p with a maximum of ± 10 days; this results in stronger smoothing further from the pivot points where there are fewer data points. We also re-weighted the volume at each pivot point by the percentage of users who performed any searches on the days before or after the pivot. This was done to adjust for the fact that by construction of the pivot points, all users performed searches on these days, which leads to a misleadingly high estimate of volume on these days compared to others.

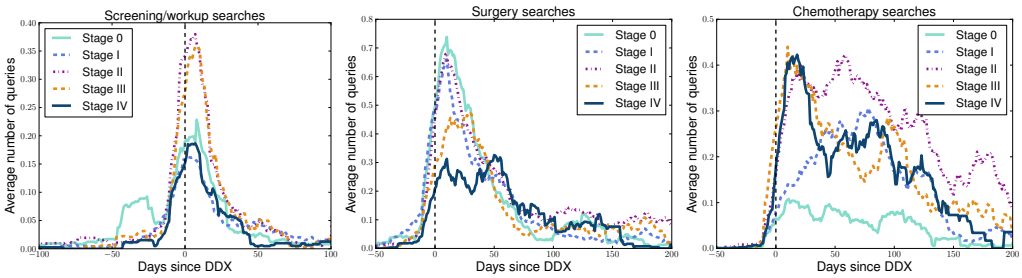

Fig. 3. Average number of queries per day by users associated (based on most frequent search) for each cancer stage. Patterns are consistent with cancer patients with these stages, e.g., surgery is not standard treatment for stage IV and chemotherapy is not standard treatment for stage 0 [Morrow and Harris 2000].

7.2.1. Multi-Pivot Histories. Figure 2 highlights a number of interesting patterns about search behavior over multiple episodes of breast cancer. We note that the overall search volume, including all other queries (beyond those captured by the ontology), remains relatively flat outside of the spikes at the pivots: this suggests that cancer-related search cuts into other search activity, and is indeed "disruptive," from the standpoint of search and retrieval performed prior to the illness. Cancer-related search is largely non-existent prior to the initial workup, but very heavy after the DDX. Cancer-related searches are 3.89 times more frequent in the 60-90 days *after* DDX than the 60-90 days *before* DDX, during which cancer searches are at baseline levels. Comparing the 30-day window beginning at 300 days after DDX, cancer-related searches are still 1.25 times greater than baseline levels. The non-ontology queries drop in frequency in these two periods after, at only .88 and .79 the baseline levels after 60-90 and 300-330 days.

We see that there is very little cancer-related search prior to the first screening/workup search, then a jump for a period of time in between the first workup searches and the DDX. The searches in this time include searches for biopsies (which take place after a suspicious screening) as well as searches about symptoms and other information. This is consistent with our observations from the logs, wherein searchers try to discern whether a suspicious mass is cancerous by searching their symptoms and related information.

There is a sudden surge of cancer-related activity at the DDX, especially in searches for more specific cancer information such as the type, stage, and grade. This aggregate behavior is consistent with searchers who have just learned of a cancer diagnosis and are searching about the specific cancer. The surgery pivot point shows an increase in searches for breast reconstruction, which appears to be a significant focus of attention for searchers during this time. Finally, the chemotherapy pivot point shows an increase in the search term "side effects" as well as searches for wigs and headscarves, which are suggestive of searchers who have experienced hair loss, a common side effect of chemotherapy. We also see an increase in searches for breast cancer drugs (many of which are chemotherapeutic) after this point.

7.2.2. Stage of Illness and Timing. To further investigate the evolution of information needs over time, we analyzed user timelines by dominant queried stage of cancer. Cancer staging describes the extent of the spread of the cancer at time of discovery: Stage 0 describes non-invasive cancer that has not spread to neighboring tissue, Stages I–III describe invasive cancers of varying size that may have spread locally, and Stage IV cancer has spread to other organs of the body [National Cancer Institute 2013]. For users who searched for specific stages of cancer at least five times, we associated each user with the coarse-grained Stage (0 through IV) searched most frequently. The num-

ber of searchers associated with Stages 0–IV are respectively 94, 217, 189, 109 and 45. For each of the five user groups, we aligned the timelines around DDX. The volume over time for the three categories associated with the change points in the previous section (screening/workup terms, surgery, and chemotherapy terms) are shown over time for each stage in Figure 3.

We find notable differences between the five searcher groups which align with clinical practices for treating patients diagnosed with each stage. Stage 0 cancer is most often discovered through routine screening mammography [Burstein et al. 2004], which may explain the notable rise in searches related to screening for this set of searchers much sooner than others. Surgery is standard care for Stages 0-III but not for Stage IV [Morrow and Harris 2000], which may explain why searchers in the stage IV group have less surgery-related search volume than others. The chemotherapy curves are consistent with the fact that higher stage patients are more likely to undergo treatment. Stage I-II patients are sometimes upstaged after surgical exploration reveals additional findings about the extent of metastasis that lead to consideration of chemotherapy, which may explain the gradual rise of stage I searchers contrasted with the sharp rise from Stage III-IV searchers [Morrow and Harris 2000].

Early stage cancer is often detected through routine screening, rather than from patients who receive a workup to explore whether such symptoms as self-detected lumps are a cause for concern. We investigated whether the logs would show findings consistent with this. Of the users who searched for a screening term (e.g., "mammogram"), we separated users based on whether they had searched any symptoms that are associated with breast cancer (such as lumps, discomfort, and pain) for the first time within 30 days prior to DDX, and users who did not. Users who searched symptoms prior to DDX had an average stage of 1.83, compared to 1.54 for users who did not (which are different with $p = 0.154$). Users searching in a pattern consistent with a prompted diagnostic workup rather than routine screening search for information about a higher stage of cancer on average.

8. LIMITS AND FUTURE DIRECTIONS

We now discuss limitations and implications of our study, as well as ongoing research on patient-centered studies of online search behavior.

8.1. Limitations of Log-Based Inferences

A limitation of this study is the lack of verified ground truth and the limitations associated with log-based inferences about whether a particular searcher has indeed been diagnosed with breast cancer. Given the terms of use under which the data were collected, we could not identify or contact any of the searchers directly to confirm a diagnosis. We can only identify searchers with new and strong shifts of attention to breast cancer, whose search characteristics over time appear similar to newly diagnosed patients. Thus, a limitation of this study is that we cannot make strong claims about breast cancer patients, but only about searchers who show strong sudden interest in breast cancer amidst a larger temporal structure of querying consistent with information needs of someone seeking information over the episodes of screening, diagnosis, and treatment of breast cancer. The patterns of activity are generally consistent with the episodic timing of cancer patients, as described in the medical literature (see Section 7.2), and correlate significantly with reported incidence rates (see Section 6.2). The analyses thus provide strong evidence that many of the searchers identified by our classifiers are indeed experiencing a breast cancer diagnosis.

Even if not all searchers in our dataset are experiencing cancer, our analyses show consistent patterns among searchers with strong interest in breast cancer, which is still noteworthy given the importance of the subject matter. Cancer is one of the most

prevalent serious medical conditions searched online. Improving our understanding of cancer-related information needs is an important step toward enhancing search and retrieval for searchers with such needs.

As future work, we envision an approach for obtaining ground truth in which consenting patients can share both their patient records and their search histories. This would provide important context such as a patient's demographics, Internet literacy, and whether they have had cancer in the past (which is a distinction that is difficult to make with logs alone). However, such a study will be slow, costly, and small in scale. We therefore argue that large-scale log analysis outside of a clinical setting is an important starting point, and we hope that the promising preliminary results reported in this paper will provide motivation for undertaking more costly patient-centered studies in the future.

8.2. Implications of Findings

Equipped with insights about the episodic phases of information needs and retrieval, designers of search systems may wish to tailor the content surfaced to searchers so that it is appropriate for the current episode. For example, if it is possible to infer that a patient is preparing for a treatment rather than recovering from it, then more appropriate results can be returned in response to otherwise ambiguous search terms. Another possibility, suggested in Paul et al. [2015], is to surface information that is typically searched in later episodes, which could help patients with deliberation about actions that can have influence on future decisions and outcomes. Such searching-ahead behavior has also been observed in other investigations of online health search behavior, for example in explorations of time-dependent concerns about pregnancy and childbirth [Fourney et al. 2015].

It might also be desirable to tailor the search engine design and interface to meet the sensitivities of patients. Such an interface might be presented to searchers exhibiting behavior that we have associated with experiencing cancer. One's information needs will vary depending on whether a searcher is a patient, survivor, loved one, or healthcare professional, and inferring different searcher types can lead to more personalized search experiences that are appropriate for the current searcher's role.

In addition to using our findings to improve search, insights from studies such as this could also be used by healthcare providers to better anticipate and address the information needs of patients. While the timing of various clinical milestones is known, this research gives new insights into the timing of *personal* milestones, concerns, and information needs that are not directly known to healthcare providers, but can be inferred from the personal Web activity of searchers. Our findings illustrate the kind of information searched by plausible patients, which we hope will motivate more targeted future research of illness in search logs.

9. CONCLUSION

We have sought to understand search behavior surrounding breast cancer-related shifts in attention, with emphasis on multiple episodes over the history of the illness. We analyzed histories of 1700 searchers with a sustained focus of attention on breast cancer. We demonstrated the reliability of the classifiers intrinsically by showing high accuracy on held-out data, and we additionally evaluated the reliability of our classifiers against external resources, showing a significant correlation between the proportions of breast cancer searchers identified in the logs and U.S. incidence rates by state and demographic group.

Our visualizations show how information-seeking patterns evolve over time with respect to clinically relevant episodes of breast cancer, including the periods of time before and after searches for diagnostic screenings, surgery, and chemotherapy, as well

as patterns for more specific groups of users by stage of cancer. We additionally examined search timelines considering the patterns demonstrated by different groups of searchers, defined by the stage of cancer they had mostly queried for. These analyses revealed interesting behavioral patterns that are often consistent with episodic timing of cancer patients, as described in medical literature.

The temporal trends illustrated in this paper are illuminating and intuitive, providing empirical evidence of the disruption caused by a serious illness. Beyond breast cancer, we believe the types of analyses and visualizations presented in this paper could be applied to other search activity surrounding events that can be described as multiple episodes. The approach offers a direction, methods, and proof-of-concept. We hope our exploration and experiments will serve as a source of ideas and directions on the prospect of making additional discoveries about information seeking around diagnosis of breast cancer and other illnesses. By improving our understanding of the evolving information needs of cancer patients and people facing other serious health problems, we ultimately hope to improve information access, decision making, and, in the end, the quality and length of life of those afflicted.

REFERENCES

J. W. Ayers, B. M. Althouse, J.-P. Allem, D. E. Ford, K. M. Ribisl, and J. E. Cohen. 2012. A Novel Evaluation of World No Tobacco Day in Latin America. *Journal of Medical Internet Research* 14, 3 (2012).

S. L. Ayers and J. J. Kronenfeld. 2007. Chronic illness and health-seeking information on the Internet. *Health* 11, 3 (2007).

M. Benigeri and P. Pluye. 2003. Shortcomings of health information on the Internet. *Health Promotion International* 18, 4 (2003).

A. Z. Broder, M. Fontoura, E. Gabrilovich, A. Joshi, V. Josifovski, and T. Zhang. 2007. Robust Classification of Rare Queries Using Web Knowledge. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)*. 231–238.

H. Burstein, K. Polyak, J. Wong, S. Lester, and C. Kaelin. 2004. Ductal carcinoma in situ of the breast. *N Engl J Med* 350, 14 (2004).

M.-A. Cartright, R. W. White, and E. Horvitz. 2011. Intentions and attention in exploratory health search. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

K. Castleton, T. Fong, A. Wang-Gillam, M. Waqar, D. Jeffe, L. Kehlenbrink, F. Gao, and R. Govindan. 2011. A survey of Internet utilization among patients with cancer. *Support Care Cancer* 19, 8 (2011).

E. H. Chan, V. Sahai, C. Conrad, and J. S. Brownstein. 2011. Using Web Search Query Data to Monitor Dengue Epidemics: A New Model for Neglected Tropical Disease Surveillance. *PLoS Negl Trop Dis* 5, 5 (2011).

R. J. W. Cline and K. M. Haynes. 2001. Consumer health information seeking on the Internet: the state of the art. *Health Education Research* 16, 6 (2001).

L. Degner, L. Kristjanson, D. Bowman, and et al. 1997. Information needs and decisional preferences in women with breast cancer. *JAMA* 277, 18 (1997).

R. Desai, A. J. Hall, B. A. Lopman, Y. Shimshoni, M. Rennick, N. Efron, Y. Matias, M. M. Patel, and U. D. Parashar. 2012. Norovirus disease surveillance using Google Internet query share data. *Clin. Infect. Dis.* 55, 8 (Oct 2012), e75–78.

D. Downey, S. Dumais, and E. Horvitz. 2007. Models of searching and browsing: languages, studies, and applications. In *International Joint Conferences on Artificial Intelligence (IJCAI)*.

G. E. Dupret and B. Piwowarski. 2008. A user browsing model to predict search engine click data from past observations.. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

G. Eysenbach. 2006. Infodemiology: Tracking Flu-Related Searches on the Web for Syndromic Surveillance.. In *AMIA Annual Symposium*.

G. Eysenbach and C. Kohler. 2002. How do consumers search for and appraise health information on the world wide web? qualitative studies using focus groups, usability test, and in-depth interviews. *BMJ* 324 (2002).

L. Fallowfield. 2001. Participation of patients in decisions about treatment for cancer. *BMJ* 323, 7322 (2001).

J. Fleiss. 1981. *Statistical Methods for Rates and Proportions. Second Edition*. Wiley, John and Sons, Incorporated, New York, N.Y.

A. Fourney, R. W. White, and E. Horvitz. 2015. Exploring Time-Dependent Concerns About Pregnancy and Childbirth from Search Logs. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*.

S. Fox and M. Duggan. 2013. *Health Online 2013*. Technical Report. Pew Internet and American Life Project. http://pewinternet.org/Commentary/2011/November/Pew-Internet-Health.aspx

S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. 2005. Evaluating Implicit Measures to Improve Web Search. *ACM Trans Inf Syst* 23, 2 (2005), 147–168.

J. Friedman, T. Hastie, and R. Tibshirani. 2000. Additive Logistic Regression: a Statistical View of Boosting. *Annals of Statistics* 28, 2 (2000), 337–407.

G. Fulgoni. 2005. The "Professional Respondent" Problem in Online Survey Panels Today. In *Market Research Association Annual Conference*.

C. M. Gaston and G. Mitchell. 2005. Information giving and decision-making in patients with advanced cancer: A systematic review. *Soc Sci Med* 61, 10 (2005).

J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant. 2008. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (2008).

R. W. Glynn, J. C. Kelly, N. Coffey, K. J. Sweeney, and M. J. Kerin. 2011. The effect of breast cancer awareness month on internet search activity - a comparison with awareness campaigns for lung and prostate cancer. *BMC Cancer* 11, 442 (2011).

Q. Guo and E. Agichtein. 2010. Ready to Buy or Just Browsing?: Detecting Web Searcher Goals from Interaction Data. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10)*. 130–137.

T. F. Hack, L. F. Degner, P. Watson, and L. Sinha. 2006. Do patients benefit from participating in medical decision making? Longitudinal follow-up of women with breast cancer. *Psycho-Oncology* 15, 1 (2006).

A. Hassan, Y. Song, and L.-w. He. 2011. A Task Level Metric for Measuring Web Search Satisfaction and Its Application on Improving Relevance Estimation. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11)*.

A. Hassan, R. W. White, S. T. Dumais, and Y.-M. Wang. 2014. Struggling or Exploring?: Disambiguating Long Search Sessions. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM '14)*. 53–62.

P. R. Helft. 2012. Patients with Cancer, Internet Information, and the Clinical Encounter: A Taxonomy of Patient Users. In *American Society of Clinical Ontology*.

A. Kotov, P. Bennett, R. White, S. Dumais, and J. Teevan. 2011. Modeling and Analysis of Cross-Session Search Tasks. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

T. Kusmierczyk, C. Trattner, and K. Norvag. 2015. Temporality in Online Food Recipe Consumption and Production. In *International Conference on World Wide Web (WWW)*.

T. Lau and E. Horvitz. 1999. Patterns of search: Analyzing and modeling Web query refinement. In *7th International Conference on User Modeling*.

M. Morrow and J. Harris. 2000. Local management of invasive breast cancer. In *Diseases of the Breast*, J.R. Harris, M.E. Lippman, M. Morrow, and C.K. Osborne (Eds.). Lippincott, Williams & Wilkins.

National Cancer Institute. 2013. Stages of Breast Cancer. (2013). http://www.cancer.gov/cancertopics/pdq/treatment/breast/Patient/page2 Online; accessed 28-January-2014.

Y. Ofran, O. Paltiel, D. Pelleg, J. M. Rowe, and E. Yom-Tov. 2012. Patterns of Information-Seeking for Cancer on the Internet: An Analysis of Real World Data. *PLOS One* 7, 9 (2012).

M. J. Paul. 2012. Mixed Membership Markov Models for Unsupervised Conversation Modeling. In *Empirical Methods in Natural Language Processing (EMNLP)*.

M. J. Paul, B. C. Wallace, and M. Dredze. 2013. What Affects Patient (Dis)satisfaction? Analyzing Online Doctor Ratings with a Joint Topic-Sentiment Model. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI*.

M. J. Paul, R. W. White, and E. Horvitz. 2015. Diagnoses, Decisions, and Outcomes: Web Search as Decision Support for Cancer. In *International Conference on World Wide Web (WWW)*.

E. J. Pérez-Stable, A. Afable-Munsuz, C. P. Kaplan, L. Pace, C. Samayoa, and C. Somkin. 2013. Factors Influencing Time to Diagnosis After Abnormal Mammography in Diverse Women. *J Women's Health* 22, 2 (2013).

G. Peterson, P. Aslani, and K. A. Williams. 2003. How do Consumers Search for and Appraise Information on Medicines on the Internet? A Qualitative Study Using Focus Groups. *J Med Internet Res* 5, 4 (2003).

K. Raman, P. N. Bennett, and K. Collins-Thompson. 2014. Understanding Intrinsic Diversity in Web Search: Improving Whole-Session Relevance. *ACM Trans Inf Syst* 32, 4 (2014), 20:1–20:45.

M. Richardson. 2009. Learning about the world from long-term query logs. *ACM Trans Web* 2, 4 (2009).

D. E. Rose and D. Levinson. 2004. Understanding User Goals in Web Search. In *Proceedings of the 13th International Conference on World Wide Web (WWW '04)*. 13–19.

L. J. F. Rutten, N. K. Arora, A. D. Bakos, N. Aziz, and J. Rowland. 2005. Information needs and sources of information among cancer patients: a systematic review of research (1980–2003). *Patient Education and Counseling* 57, 3 (2005).

M. Santillana, D. W. Zhang, B. M. Althouse, and J. W. Ayers. 2014. What can digital disease detection learn from (an external revision to) Google Flu Trends? *Am J Prev Med* 47, 3 (Sept. 2014), 341–347.

M. J. Satterlund, K. D. McCaul, and A. K. Sandgren. 2003. Information Gathering Over Time by Breast Cancer Patients. *J Med Internet Res* 5, 3 (2003).

M. I. Trotter and D. W. Morgan. 2008. Patients' use of the Internet for health related matters: a study of Internet usage in 2000 and 2006. *Health Informatics* 14, 3 (2008).

J. Vandergrift, J. Niland, R. Theriault, S. Edge, Y. Wong, and et al. 2013. Time to Adjuvant chemotherapy for Breast cancer in National comprehensive cancer Network institutions. *J Natl Cancer Inst* 105, 2 (2013).

R. West, R. W. White, and E. Horvitz. 2013. From cookies to cooks: insights on dietary patterns via analysis of web usage logs. In *International Conference on World Wide Web (WWW)*.

R. W. White, P. N. Bennett, and S. T. Dumais. 2010. Predicting Short-term Interests Using Activity-based Search Context. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10)*. 1009–1018.

R. W. White and S. M. Drucker. 2007. Investigating behavioral variability in web search. In *International Conference on World Wide Web (WWW)*.

R. W. White and E. Horvitz. 2009. Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search. *ACM Trans Inf Syst* 27, 4 (2009).

R. W. White and E. Horvitz. 2010. Web to world: predicting transitions from self-diagnosis to the pursuit of local medical assistance in web search. *AMIA Annu Symp Proc* 2010 (2010), 882–886.

R. W. White and E. Horvitz. 2012. Studies of the onset and persistence of medical concerns in search logs. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

R. W. White and E. Horvitz. 2013a. From health search to healthcare: explorations of intention and utilization via query logs and user surveys. *J Am Med Inform Assoc* epub (2013).

R. W. White and E. Horvitz. 2013b. From web search to healthcare utilization: privacy-sensitive studies from mobile data. *J Am Med Inform Assoc* 20 (2013).

R. W. White, N. P. Tatonetti, N. H. Shah, R. B. Altman, and E. Horvitz. 2013. Web-Scale Pharmacovigilance: Listening to Signals from the Crowd. *J Am Med Informatics Assoc* 20, 3 (2013).

E. Yom-Tov and E. Gabrilovich. 2013. Postmarket Drug Surveillance Without Trial Costs: Discovery of Adverse Drug Reactions Through Large-Scale Analysis of Web Search Queries. *J Med Internet Res* 15, 6 (2013), e124.

S. Ziebland, A. Chapple, C. Dumelow, J. Evans, S. Prinjha, and L. Rozmovits. 2004. How the internet affects patients' experience of cancer: a qualitative study. *BMJ* 328, 7439 (2004).

Received February 2007; revised March 2009; accepted June 2009