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

Michael J. Paul*
Johns Hopkins University
3400 N. Charles Street
Baltimore, MD 21218
mpaul@cs.jhu.edu

Ryen W. White
Microsoft Research
One Microsoft Way
Redmond, WA 98052
ryenw@microsoft.com

Eric Horvitz
Microsoft Research
One Microsoft Way
Redmond, WA 98052
horvitz@microsoft.com

ABSTRACT

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 disruptive shifts in focus of attention and long-term patterns of search behavior consistent with the diagnosis and treatment of breast cancer. We build and apply probabilistic classifiers to detect these searchers from multiple sessions and to detect the timing of diagnosis, using a variety of 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 likely time of diagnosis. We 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 from searchers for whom diagnosis was predicted, using a detailed ontology of cancer-related search terms. Our 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 dominant stage expressed by searchers.

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

Keywords: medical search, cancer, behavior analysis

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 future 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 [6, 15, 21, 27].

Disruptive changes, such as being diagnosed with a life-threatening illness, may lead to characteristic patterns of search over extended timelines. Aligning and aggregating patterns in longitudinal search behavior across many searchers can serve as lens for understanding the interests and intentions of searchers over time. We present such an analysis in this paper, focusing on breast cancer as a sensor of human behavior and attention, understanding the information needs that searchers traverse over time. Observed patterns of interest and concern in search logs strongly correlate with expected questions and informational needs associated with the diagnosis and treatment of someone who has been faced with the disruptive news about cancer.

To study the evolving and episodic nature of search in the context of breast cancer, we leverage anonymized search logs from a popular Web search engine 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 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 a diagnosis had been received, based on the appearance of a flood of detailed pathology and staging queries coming after queries on screening and biopsy, at a well-understood rhythm of the life history of breast cancer [29, 24]. We then 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 cor-

*Work conducted during a Microsoft Research internship.

WOODSTOCK '97 El Paso, Texas USA

relate rates of breast cancer estimated by considering the geographical (U.S. state) and demographic (gender and age) distribution of searchers 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 our 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 identified as characteristic of 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

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 [11]. Internet-based health information acquisition enables broader and faster information access. However, concerns have arisen about the information available via the Web as healthcare information on the Web varies in quality and clarity [8] and the information can be misunderstood and misused [3]. Several researchers have sought to gain an understanding of health information-seeking on the Internet, using interviews and focus groups [23], surveys [28], and more recently, large-scale analysis of search engine logs [2, 5, 32].

Information access plays an important role for cancer patients [35], most especially during treatment [26], when decisions with difficult tradeoffs and unclear answers must be made. Almost all cancer patients want access to all relevant information [13], and a majority of breast cancer patients prefer to have a role in decision making [18]. These reasons, coupled with the rise of the Web as a common information source for cancer patients [27], have led to further study of online information-seeking for cancer [6, 15]. In a recent study, resonating with motivations of this paper, Ofran et al. [21] 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.

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 [30], to predict future search actions and interests [17, 9, 10], and to detect real-world events and activities [25]. 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 [14] and dengue [7] outbreaks, the discovery of side effects of medications [34], insights into healthcare utilization [33], and measuring the effectiveness of public health awareness campaigns [1]. Studies in the information retrieval (IR) community have also examined search tasks that span multiple sessions [16], with a view to supporting task resumption over time.

3. SURVEY ON CANCER-RELATED WEB ACTIVITY

To give additional context and motivation for our log-based analysis, we first present results of a survey we conducted asking a random sample of employees at our institution about their Web activity and experience following a cancer diagnosis. We collected 867 anonymous responses using an internal Web-based survey system.

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 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 numbers suggest that a substantial number of people search the Web regarding a recent cancer diagnosis.

Quality of Information.

Although many people 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 for information about cancer and its treatment:

- *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 Eastern 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

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

Table 1: The percentage of personally diagnosed survey respondents who searched for various types of information, as well as the ratio of values between those diagnosed and all other respondents.

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. On the other side of opinions, respondents expressed clear advantages of accessing information from 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. The first 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 paper 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 affected by cancer, we asked respondents to state the types of content that were searched (from a checkbox list), e.g. information about prognosis or treatments. Table 1 shows the percentage of respondents who were personally diagnosed that searched for different categories of search.

We calculated the ratio of each categories percentage 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 2.6), 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 experiencing cancer, as described in the next section.

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

Table 2: An example of queries by a fictitious user consistent with many users in our dataset. In this example, Dec 12 would have been labeled as DDX.

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. 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 popular Web search engines (Google, Bing, Yahoo!, etc.), 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 Internet analytics company, comScore. comScore recruits millions of opt-in consumer panelists who give explicit permission to passively measure their online activities using monitoring software citeFulgoni05. Search queries in these logs are associated with the user's age and gender. We employ the comScore data only in section 6.2.2, to compare the demographic distribution 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 diag-

nosis with experiential search sessions. Given the terms of use under which the 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.

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. 138,306 users met this criteria. Then, we set out to manually tag searchers whose histories were consistent with a cancer diagnosis. We sought tags on 480 of these users drawn randomly from the larger set. The three co-authors 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 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 2 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 our term ontology described in the next section) as well as timestamp information (but not the query content) 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.

Annotator 1 labeled all 480 searchers, while Annotators 2 and 3 each labeled disjoint sets of 150 searchers. 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% ($\kappa = 0.61$). A plurality (40%) of disagreements were between the PN and PP labels, while a minority (28%) of disagreements were on N versus PP. 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. 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.

4.3 Term Ontology

We created a large ontology of health- and cancer-related words and phrases. The purpose for 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 has 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. Table 3 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.

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

Category			
Level 1	**Level 2**	**Level 3**	**Terms**
Cosmetic	Post-Surgery	Post-Surgery	{cosmetic,plastic} {surgery,surgeon}, prosthesis, prosthetic(s), implant(s), reconstruction
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, metaplastic, mucinous, inflammatory
Description	Staging/Grading	Staging/Grading	what stage, stages, staging, what grade, grades, grading, differentiated
Description	Staging/Grading	Stage	pre( )cancer, 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, protein, recipe(s), cookbook
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

Table 3: A 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.

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").

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. We also created a set of strings that mention the age of a person (the AGE category, including strings such as "at 55", "age 55", and "55 year(s) old".

We created also categories with symptom and disease words and phrases, composed by White and Horvitz [31]. 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 US.

5. DIAGNOSIS CLASSIFICATION

5.1 DX Classifier (Searcher Model)

A key subgoal 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.

We extracted the set of terms and categories in Table 3 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). We created additional such conjunctions with whether queries contain question words or first/second person personal or possessive pronouns, which could indicate that the terms are searched from an *experiential* perspective—that is, in a manner signaling that the searcher is experiencing symptoms rather than only exploring their meaning.

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 3, 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).
- Number of different ontology/cancer categories ever searched, and the distribution over categories.
- 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.
- 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.

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 2, 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.
- Average, minimum, and maximum number of days between sessions containing an ontology/cancer terms or "cancer", and the average number of 7+ or 30+ day gaps between such sessions.
- 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.
- 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. This models our conjecture that searchers with the illness tend to search for categories in a certain order; e.g., searches about diagnostics precede searches about treatment options.

5.2 DDX Classifier (Timeline Model)

A second important task is to predict the point in time when a searcher likely receives a 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.

For this model, we extracted two separate sets of lexical count features (query/session/user counts of the ontology terms and categories) for the days before d and the days after d, as well as d itself. For each of the various volume and temporal pattern features for the DX model, we created DDX a feature whose value is the *ratio* of the values of the feature in the days before d to the value after d. We created similar "before-and-after" features extracted from the only the 10 days before and after d. Additionally, we created these features for day d:

- Number of queries containing ontology/cancer terms from Table 3 searched on this day.
- Number of different ontology categories searched on this day and searched for the first time on this day, as well as binary features for each category indicating if it 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.

5.3 Training and Prediction

We used Multiple Additive Regression Trees (MART) [12] for both the classification tasks. MART uses gradient tree boosting methods for regression and classification. Advantages of employing MART include model interpretability, facility for rapid training and testing, and robustness against noisy labels and missing values.

We trained two DX classifiers, one to predict whether the searcher appears to have 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). Predictions are made using the joint probability of these two classifiers. We built DDX classifiers by applying binary classifiers to each day in a searcher timeline independently. We trained three classifiers to predict DDX at different granularities: (1) all days within 7 days of DDX are labeled as positive and all others are negative; (2) the single day of diagnosis is labeled positive, the others within 7 days are negative, and all others outside this window are excluded; and (3) all days after and including DDX are positive and all days before are negative. The predictions from these three classifiers are combined to produce a final DDX prediction.

6. EXPERIMENTAL EVALUATION

6.1 Classifier Validation

We evaluated the DX and DDX classifiers by performing 10-fold cross-validation. For the DX classifier, we measured the maximum F1 score reached at all prediction thresholds, 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, Table 4 compares 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). We see that the full DX model performs better along several metrics than the 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. Not shown in the table is the accuracy within 30 days, which is 96.3% ($\pm$ 0.7) at 100% recall in the full model.

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.

6.2 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).[1] These statistics include 49

[1] http://apps.nccd.cdc.gov/uscs/cancersbystateandregion.aspx

Features	Max F1	Rec@P90	Pre@R25
Lexical Only	74.6 ± 2.1	30.7 ± 6.3	90.6 ± 4.0
Full Model	76.5 ± 2.7	39.6 ± 4.4	94.3 ± 2.3

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
Full Model	42.5 ± 4.7	72.2 ± 0.5	88.5 ± 0.3
	25% Recall		
Lexical Only	59.0 ± 2.6	83.0 ± 0.6	90.4 ± 0.1
Full Model	72.8 ± 8.2	90.2 ± 4.2	99.0 ± 0.2

Table 4: Cross-validation performance with 95% confidence intervals for the DX (top) and DDX (bottom) classifiers.

U.S. states (data from the state of Wisconsin is suppressed by law). 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 ≥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 ten-fold 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.[2] 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 ≥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.

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

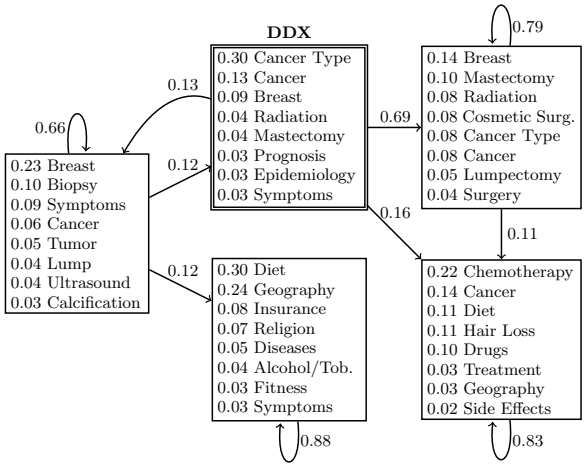

Figure 1: A five-state HMM inferred from searcher timelines. A special state was reserved for the day of diagnosis. Each box shows the highest probability ontology categories for that state, and edges between boxes indicate transition probabilities. Only edges with transition probability ≥0.1 are shown.

7. ANALYSIS OF LIFE HISTORIES

We 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 [29].

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.

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 [22], 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 5 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 [24] and mean times from diagnosis to surgery of 5.6 weeks and surgery to chemotherapy as 6.3 weeks in the U.S. [29]. 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 reweighted 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.

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 base-

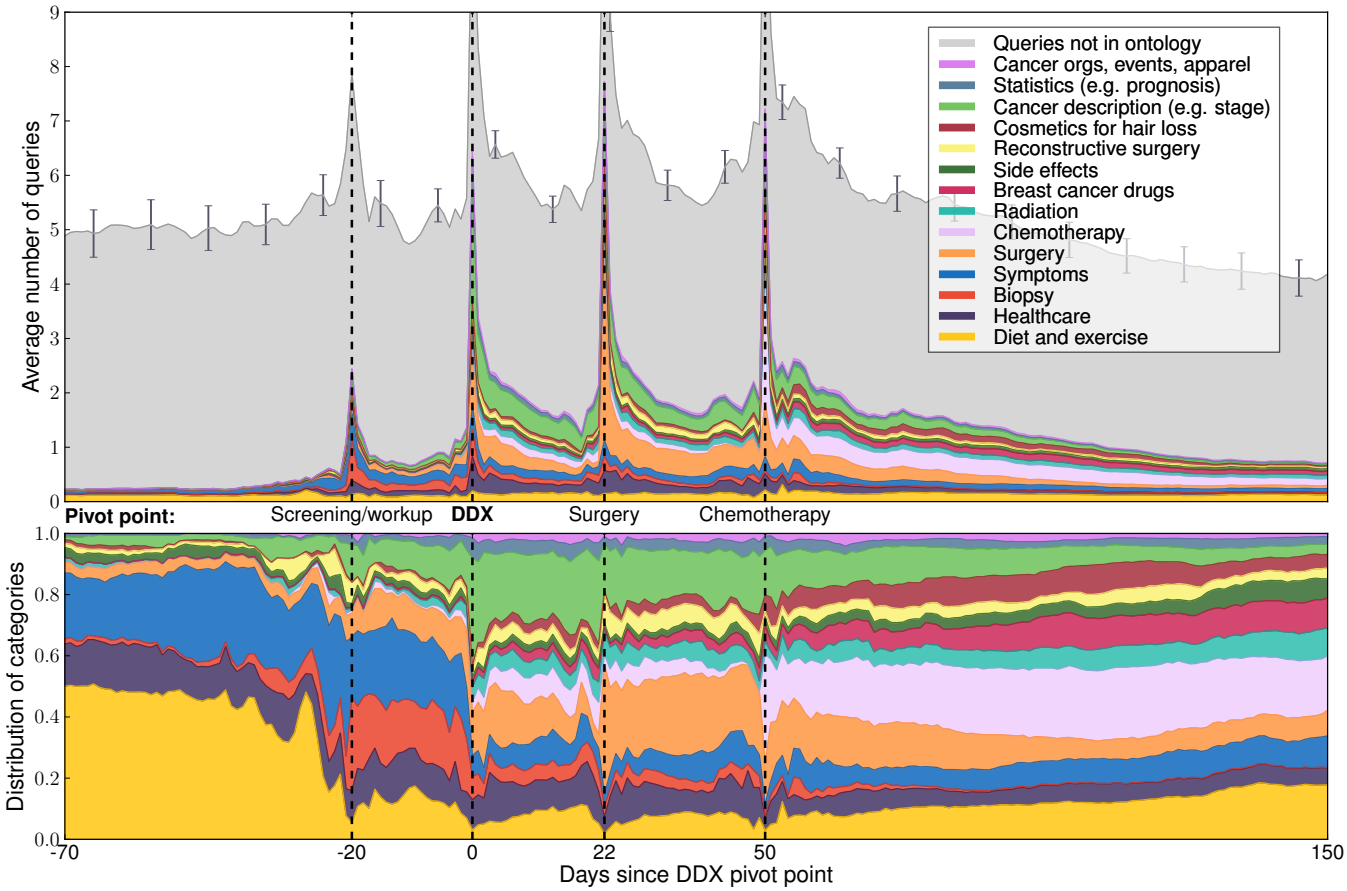

Figure 2: The raw (above) and normalized (below) average number of queries per day for different search categories. The day on the x-axis is with respect to the pivot point, while the y-axis value is averaged between the values of the two surrounding pivot points. Standard error of the unsmoothed values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

line 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 [20]. 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 number 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 dis-

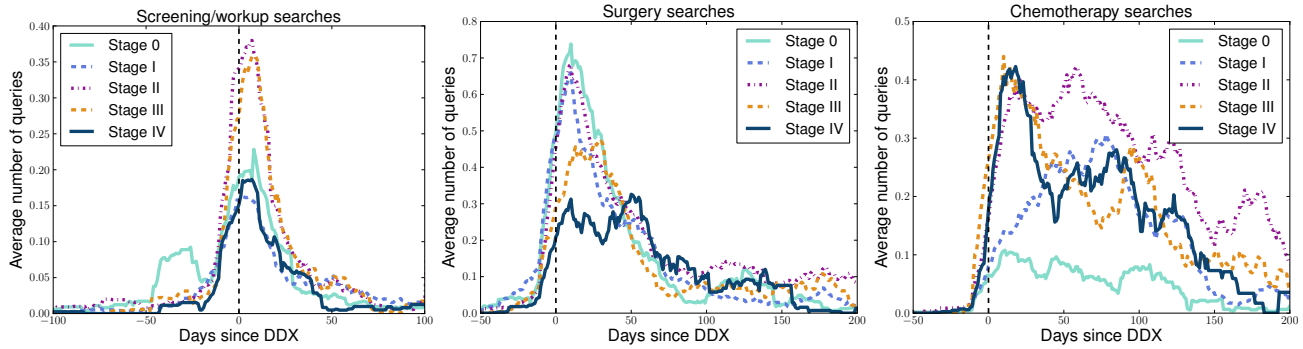

Figure 3: The average number of queries per day by users associated (based on most frequent search) for each cancer stage. Patterns are consistent with cancer patients with these stages, e.g. surgery is not standard treatment for stage IV and chemotherapy is not standard treatment for stage 0 [19].

covered through routine screening mammography [4], 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 [19], 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 [19].

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 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. 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. The patterns of activity are generally consistent with the episodic timing of cancer patients, as described in the medical literature, and correlate significantly with reported incidence rates. The analyses provide evidence that many of the searchers identified by our classifiers are experiencing a cancer diagnosis. Improving our understanding of the information needs of people facing major diseases over the episodes of illness is a first step to enhancing search and retrieval for these searchers. Given new insights about the episodic phases of information needs and retrieval, designers of search systems may then wish to tailor the content surfaced to searchers so that it is appropriate for the current episode.

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 appear similar to newly diagnosed patients. At the time of writing, we are engaging directly with oncologists, surgeons, and breast cancer patients to understand the nature of relevant search activities before and after diagnosis. We have created a survey that provides options for newly diagnoses patients to consent sharing their long-term query histories, as well as dates of diagnosis and other key milestones of the life history of their illnesses. The ability to connect long-term search data, user self-reports, and electronic health records (in IRB-approved studies) can serve as a powerful joint methodology, for learning about the links between search behavior and associated clinical situations. To date, only a small number of participants have agreed to share this data with us. While we are actively working to recruit more patients, and hope to present a small-scale study with the ground truth data in future work, the more invasive and detailed methods will not match the scale of the experiments presented in this paper. We argue that there is clear utility in the classifier-based approach of the current study, which offers broad insights on health seeking over time from large user populations.

The temporal trends illustrated in this paper are illuminating yet 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. We showed that relevant pivot points in search timelines can be identified with simple heuristics like the first day of particular search times, or identified with more sophisticated classifiers. 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.

9. REFERENCES

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

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

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

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

[5] M.-A. Cartright, R. W. White, and E. Horvitz. Intentions and attention in exploratory health search. In *SIGIR*, 2011.

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

[7] E. H. Chan, V. Sahai, C. Conrad, and J. S. Brownstein. Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. *PLoS Negl Trop Dis*, 5(5), 2011.

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

[9] D. Downey, S. Dumais, and E. Horvitz. Models of searching and browsing: languages, studies, and applications. In *IJCAI*, 2007.

[10] G. E. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In *SIGIR*, 2008.

[11] S. Fox and M. Duggan. Health online 2013. Technical report, Pew Internet and American Life Project, 2013.

[12] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. In *Technical Report, Stanford University.*, 1998.

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

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

[15] P. R. Helft. Patients with cancer, internet information, and the clinical encounter: A taxonomy of patient users. In *American Society of Clinical Ontology*, 2012.

[16] A. Kotov, P. Bennett, R. White, S. Dumais, and J. Teevan. Modeling and analysis of cross-session search tasks. In *SIGIR*, 2011.

[17] T. Lau and E. Horvitz. Patterns of search: analyzing and modeling web query refinement. In *7th international conference on user modeling*, 1999.

[18] D. LF, K. LJ, B. D, and et al. Information needs and decisional preferences in women with breast cancer. *JAMA*, 277(18), 1997.

[19] M. Morrow and J. Harris. Local management of invasive breast cancer. In J. Harris, M. Lippman, M. Morrow, and C. Osborne, editors, *Diseases of the Breast*. Lippincott, Williams & Wilkins, 2000.

[20] National Cancer Institute. Stages of breast cancer, 2013. [Online; accessed 28-January-2014].

[21] Y. Ofran, O. Paltiel, D. Pelleg, J. M. Rowe, and E. Yom-Tov. Patterns of information-seeking for cancer on the internet: An analysis of real world data. *PLOS One*, 7(9), 2012.

[22] M. J. Paul. Mixed membership markov models for unsupervised conversation modeling. In *EMNLP-CoNLL*, 2012.

[23] G. Peterson, P. Aslani, and K. A. Williams. 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.

[24] E. J. PÃľrez-Stable, A. Afable-Munsuz, C. P. Kaplan, L. Pace, C. Samayoa, and C. Somkin. Factors influencing time to diagnosis after abnormal mammography in diverse women. *J Women's Health*, 22(2), 2013.

[25] M. Richardson. Learning about the world from long-term query logs. *ACM Transactions on the Web*, 2(4), 2009.

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

[27] M. J. Satterlund, K. D. McCaul, and A. K. Sandgren. Information gathering over time by breast cancer patients. *J Med Internet Res*, 5(3), 2003.

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

[29] J. Vandergrift, J. Niland, R. Theriault, S. Edge, Y. Wong, and et al. Time to adjuvant chemotherapy for breast cancer in national comprehensive cancer network institutions. *J Natl Cancer Inst*, 105(2), 2013.

[30] R. W. White and S. M. Drucker. Investigating behavioral variability in web search. In *WWW*, 2007.

[31] R. W. White and E. Horvitz. Cyberchondria: Studies of the escalation of medical concerns in web search. *ACM TOIS*, 27(4), 2009.

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

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

[34] R. W. White, N. P. Tatonetti, N. H. Shah, R. B. Altman, and E. Horvitz. Web-scale pharmacovigilance: Listening to signals from the crowd. *J Am Med Informatics Assoc*, 20(3), 2013.

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