Diagnoses, Decisions, and Outcomes: Web Search as Decision Support for Cancer

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
People diagnosed with a serious illness often turn to the Web for their rising information needs, especially when decisions are required. We analyze the search and browsing behavior of searchers who show a surge of interest in prostate cancer. Prostate cancer is the most common serious cancer in men and is a leading cause of cancer-related death. Diagnoses of prostate cancer typically involve reflection and decision making about treatment based on assessments of preferences and outcomes. We annotated timelines of treatment-related queries from nearly 300 searchers with tags indicating different phases of treatment, including decision making, preparation, and recovery. Using this corpus, we present a variety of analyses toward the goal of understanding search and decision making about treatments. We characterize search queries and the content of accessed pages for different treatment phases, model search behavior during the decision-making phase, and create an aggregate alignment of treatment timelines illustrated with a variety of visualizations. The experiments provide insights about how people who are engaged in intensive searches about prostate cancer over an extended period of time pursue and access information from the Web.

Categories and Subject Descriptors: H.2.8 [Database management]: Database applications—data mining
Keywords: medical search; decision making; cancer

1. INTRODUCTION

Upon diagnosis of a major illness, people frequently turn to the Web for information about the course and prognosis of the disease, and to better understand treatments and outcomes [6, 19, 27, 36]. We seek to understand the use of Web search as a medical decision support system by patients who have been diagnosed with a significant disease. In particular, we study the use of Web search to support the decisions of searchers who show salient signs of having been recently diagnosed with prostate cancer. We aim to characterize and enhance the ability of Web search engines to provide decision support for different phases of the illness.

We focus on prostate cancer for several reasons. Prostate cancer is typically slow-growing, leaving patients and physicians with time to reflect on the best course of action. There is no medical consensus about the best treatment option, as the primary options all have similar mortality rates, each with their own tradeoffs and side effects [24]. The choice of treatments for prostate cancer is therefore particularly sensitive to consideration of patients’ preferences about outcomes and to the assessed likelihoods of achieving different outcomes. For these reasons, prostate cancer is an “archetypal condition” for the use of decision aids [25].

Guidance on decisions is provided primarily via consultation with a patient’s physician and larger care team. Formal decision-making materials may be available through patients’ physicians. However, Web search is becoming a common supplement to traditional decision aids [29, 4]. In a 2011 survey of prostate cancer patients, the Internet was found to be the second most common information source for making treatment decisions, after “doctor’s recommendation” [37]. An earlier survey found that more than half of the respondents who had searched the Web prior to deciding on a treatment reported that information reviewed online had influenced their decision [32]. Search for information on major illnesses includes information gathering about treatments and outcomes, including characterization of uncertainties about outcomes for different treatments. Beyond providing information about illness, search and retrieval sessions serve as opportunities for interactive assessment of preferences about outcomes and risk. Thus, we are particularly interested in how people learn about different therapeutic options, including how they progress over time in navigating trees of possibilities.

To pursue insights about search and retrieval as medical decision support, we adapt methods developed in a prior study on the use of Web search for pursuing information on breast cancer [31]. In that work, classifiers were constructed from annotated logs to infer that a searcher had likely received a diagnosis of breast cancer. An ontology of different kinds of information needs was introduced to characterize the dynamics of information-seeking for a large set of searchers aligned by an inferred date of diagnosis. We harness similar methods to seek insights about prostate cancer, but focus on decision making and other search activities aligned with different phases of treatment. Specifically, we
perform studies that extend previous work in a number of ways and make the following contributions:

• We create a hierarchy of treatments and associated search terms, as well as an annotated corpus of 272 timelines of treatment search queries.

• We present a characterization of different phases of treatment search, exemplified by n-grams from queries and the content of visited webpages associated with each phase in the annotated corpus. We further characterize the phases and their progression over time by creating a multiple sequence alignment of timelines. We create a series of visualizations illustrating how the phases and queries evolve over time, based on the alignment.

• Focusing specifically on queries tagged as pursuit of decision support, we analyze the number and specificity of treatments that are searched over time, the treatments that co-occur in comparative searches, and transitions among the treatments at the focus of attention in successive queries. We identify and visualize typical sequences of treatment queries traversed by searchers by applying a spanning-tree algorithm to a graph of query transitions.

After presenting analyses and findings on the use of the Web for decision support, we discuss implications and directions for the design of search and retrieval systems that help people to better understand diagnoses and treatment decisions moving forward.

2. RELATED WORK

The Web is an important source of health-related information for many people. According to a 2013 survey, 59% of American adults had used the Web to find health information in the year preceding the survey, 35% of those adults engaged in self-diagnosis, and over half of these self-diagnosing searchers then discussed the matter with a clinician [14]. Despite the potential benefits, concerns have been raised about the quality of online health information [9] including cancer information [17]. A survey of oncologists noted that Web use can “simultaneously make patients more hopeful, confused, anxious, and knowledgeable.” [19] In a large-scale survey of the use of search for self-diagnosis, White and Horvitz [42] found that almost 40% of participants experienced increased anxiety from searching health information online.

Such challenges highlight the criticality of understanding how patients use the Web, including the nature and dynamics of queries, and the content delivered in response to queries. To better understand how people pursue health information, studies have examined online health search using a variety of methods, including interviews [33], surveys [38], and analyses of large-scale search log data [3, 1, 20, 5, 43]. Search logs analyses can provide insights about how people use search engines [41], predict future search actions and interests [23, 10, 11], and detect real-world events and activities [34]. Applications of search log data in the health and medical domains include the detection of influenza [16] and the discovery of side effects of medications [44]. Studies of online information-seeking for cancer [6, 19] have characterized how cancer patients use Web resources and have proposed patient taxonomies describing how people employ retrieved health information. In the realm of search and retrieval on cancer, Bader et al [2] categorized cancer-related search queries from three months in 2001. More recently, Ofran et al. [27] used search log data to identify five phases of cancer search activity and showed that the phases mirror those associated with coping and grief that had been previously documented in the literature.

Information access about treatment options has been found to be important for cancer patients facing difficult therapeutic decisions. Cancer patients typically seek access to all relevant information [35, 15]. Valuable information about treatments and outcomes can come via reviewing testimonials from those afflicted with similar conditions [28]. Beyond providing information about treatment options, sharing health experiences online can help people to feel supported and to better engage with health services [46]. Formal decision aids [26] have been used to help patients decide among treatment alternatives, by providing information that helps to resolve or to clarify their uncertainties [25]. A study found that working with decision aids for prostate cancer could influence decisions about treatment strategy [39].

3. TREATMENT TIMELINES

We focus on the analysis of treatment timelines extracted from search query logs. A treatment timeline for prostate cancer is a time-stamped sequence of search queries that contain terms pertaining to prostate cancer treatment, where sequences of queries are associated with unique, anonymized user identifiers. The set of treatment timelines was created by first identifying relevant search histories via a series of filters that are described in detail in Section 4. We tag the queries in the treatment timelines with labels indicating the assessed phase of treatment, including whether the searcher appears to be seeking information for an initial or a follow-on, secondary treatment, and whether the queries appear to be aimed at seeking information decisions about a treatment, preparation for a chosen treatment, or recovery from a treatment, as described in Section 4.4.

In Section 5, we present a series of experimental analyses of the treatment timelines. We characterize the content associated with different phases of information pursuit and show how these phases evolve over time for a set of searchers who are temporally aligned by inferred date of diagnosis.

3.1 Treatment Hierarchy

In order to extract treatment timelines from search histories, we identify a set of search terms that searchers tend to use to refer to prostate cancer treatment options. As queries range from very general (e.g. “cancer treatment”) to very specific (e.g. “low dose radiation seed implants”), we organized the terms into a hierarchical ontology of known treatments, moving from broad categories down to detailed therapies. Such a treatment hierarchy enables us to analyze treatment timelines in terms of categories of treatments as well as the raw text used to describe options. We can characterize the different types of treatments that are searched and the degree of specificity of queries, based on the depth of the query terms in the hierarchy.

Table 1 shows the treatment hierarchy and the terms associated with each category. The treatment hierarchy was constructed by an extensive review of the literature on the management of prostate cancer. Categories in the treatment hierarchy reflect current standard treatment options for prostate cancer. “Observation” is a treatment option that refers to a decision to forgo treatment for the time being; the two common methods of observation are typically
referred to as “watchful waiting” and “active surveillance” by clinicians. These options may be recommended when the cancer is low grade and the risks of treatment are assessed as outweighing the risks of the disease [21]. HIFU (high-intensity focused ultrasound) and cryotherapy are newer experimental treatments that are less common, though still found to be frequently searched. Another type of treatment, immunotherapy (affecting the patient’s immune response), is rarely found in queries in our dataset, and is not included in our analysis.

4. DATASET CREATION

We now review the extraction and tagging of data for the study, including the formulation of relevant search terms, labeling of searchers as likely facing prostate cancer decisions, and annotating phases of long-term timelines.

4.1 Ontology of Relevant Terms

To identify relevant search queries for the study, we relied on a manually-curated ontology of terms of interest. Terms are organized in a four-level hierarchy and include terms related to screening methods, diagnosis (e.g. “biopsy”), cancer staging and grading information (e.g. “stage II”, “low grade”), and various treatment options. The formulation of relevant query terms is similar to efforts to construct an ontology for search and retrieval for breast cancer [31], and is distinct from the treatment ontology in Table 1.

4.2 Search and Browsing Logs

The data for this study comes from a proprietary set of anonymized logs from consenters using the Internet Explorer Web browser. The data includes time-stamped search queries issued through the browser (primarily via interactions with the Microsoft Bing search engine) and time-stamped webpage visits. Each log entry includes a unique anonymized user identifier. The data spans an 18-month period from March 2013 to August 2014.

The initial dataset consists of logs collected from users whose queries include the bigram “prostate cancer” at least three times during this time period. This policy was employed as an initial high-recall filter to identify search histories that would likely be relevant to the study. Given our focus on treatment-related search, we filtered the set for search histories including queries containing treatment-related terms listed in Table 1 (excluding the most general term, “treatment”). This extraction procedure yielded a set of 3,066 search histories.

4.3 Experiential vs. Exploratory Searchers

A key initial task with the extracted data is to identify a high-precision set of search histories from the 3,066 candidate histories. Ideally, the resulting focused set would contain only histories of those who actually experienced a diagnosis of prostate cancer—either personally, or via the intensive searching performed with regard to the diagnosis of a close family member or friend. However, we cannot conclude with certainty whether a searcher is in this situation based only on information in the logs. We consider searchers as experiential versus less-involved exploratory searchers based on an assessment of sustained and focused interest in prostate cancer. We exclude search histories that are inconsistent with a cancer diagnosis. Following an annotation of logs as being experiential versus exploratory, we train a classifier (described below) on a small set of labeled histories to identify experiential searchers among the set of candidates in an automated manner.

We annotated a sample of 100 histories with binary relevance tags using the criteria outlined above. One of the authors (EH), with formal medical and decision analysis train-
Table 2: Number of search histories and search queries labeled with each phase in the dataset.

<table>
<thead>
<tr>
<th>Phase</th>
<th># histories</th>
<th># queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial decision</td>
<td>174</td>
<td>2008</td>
</tr>
<tr>
<td>Initial post-preparation</td>
<td>126</td>
<td>882</td>
</tr>
<tr>
<td>Secondary decision</td>
<td>140</td>
<td>1232</td>
</tr>
<tr>
<td>Secondary preparation</td>
<td>84</td>
<td>769</td>
</tr>
<tr>
<td>Secondary post-treatment</td>
<td>24</td>
<td>114</td>
</tr>
<tr>
<td>Total</td>
<td>272</td>
<td>5090</td>
</tr>
</tbody>
</table>

4.4 Annotation of Treatment Timelines

Last, we annotated the treatment timelines—the experiential search histories projected down to only those queries containing treatment terms—with richer tags to allow for finer-grained analysis. The queries are tagged as belonging to different phases of the treatment process that we had observed as common patterns in the timelines. We were particularly interested in tagging queries that appeared to indicate decision making, but we also tagged other phases of treatment-related queries for queries that appeared to come before and after a treatment had occurred.

Each query was annotated with two labels. The first is one of three states of the treatment deliberation process:

- **Decision:** Queries that appear to be used to help a searcher decide between or learn more about different treatment options. These queries would sometimes contain explicit indicators that the searcher is considering different options (e.g., “best treatment options”, “which is better”, “pros and cons”). Queries for many different treatments in the same timeframe are considered decision queries.

- **Preparation:** Queries about a treatment that appears to be scheduled but before the treatment has taken place (e.g. “what to expect”). If a search history focuses on a single treatment for many days (as opposed to exploring multiple options), we consider these queries preparatory.

- **Post-treatment:** Queries that appear to take place after a treatment has commenced or completed. These may seek information about recovery, or queries regarding side effects that are experienced. Some queries include specific timing references (e.g., “3 weeks after surgery”) which can help with determining this label.

The second label captures whether the context of a search session is an initial or follow-up treatment:

- **Initial:** The first round treatment that the searcher is considering, typically surgery or radiation.

- **Secondary:** Any treatment that follows an initial treatment, typically adjuvant radiation, hormone therapy, or chemotherapy for more advanced cancer. Secondary status is often clear from queries with explicit indicators like the term “adjuvant” or including such terms as “after surgery”.

The cross-product of these tags defines a total of six different phases of treatment-related search.

The queries were categorized assuming that they were issued by patients experiencing cancer and in reference to treatment for a specific patient, who may be the searcher himself or a family member deeply involved in decisions about the illness. The goal was to group the search activity based on common characteristics that are observable in the data and consistent with a typical patient timeline.

Ambiguous queries were tagged with multiple phases. If more than one phase was included, the phases were ranked based on which phase the annotator believed was most likely. Queries that did not fit these phase labels were not annotated. The tagging of the phases was done by the first author and a second professional annotator formally trained in linguistics. The first annotator reviewed the secondary annotations to ensure consistency of tagging procedures.

Table 2 provides the number of queries labeled with each phase as well as the number of search histories with at least one query labeled with the phase. In the case of ambiguous annotations, only the most likely label was counted in this table. Additionally, the 272 histories contained 33,945 queries that were not annotated with these phase labels.
5. ANALYZING TREATMENT PATTERNS

We now discuss methods and results on characterizing the different phases of treatment as well as the dynamics of the progression of searchers through the phases over time.

5.1 Phase Characterization

We characterize different annotated phases of searcher timelines by identifying n-grams—from search queries and webpage bodies—and domain names that are most associated with retrieval in each phase. We wish to identify features that are salient—both probable and representative of the phase [8]. We achieve this with a two-component mixture model that mixes phase-specific feature distributions with a phase-independent background distribution which accounts for common features that are not representative of any particular phase [45].

With this model, the probability of a feature i (an n-gram or a domain name) in the text associated with phase k is a mixture of two parameters θ:

\[
P(\text{feature} = i | \text{phase} = k) = \lambda \theta^B_i + (1 - \lambda) \theta^k_i \quad (1)
\]

Each θ^B_i is a distribution over features specific to the phase k, while θ^k_i is a background distribution over features independent of phase, and λ is the mixing weight. Then, we examine the θ^k_i distributions to find salient feature associations with phase k, because these distributions will put the most mass on n-grams that are probable within the phase but are not better explained by the background distribution.

Experiment Details.

We created a model for each class of features (search and page n-grams, and page domain names) from the annotated queries. We modeled bigrams and trigrams. Features derived from webpage content include the pages visited during the same search session following an annotated query (i.e., on the post-query navigational trail), which is possible since we used browser logs for our analysis. To extract page content, we used the method described in [40], which extracts lines of HTML such that the ratio of text tokens to tags tokens is at least one standard deviation above the mean ratio. This is a simple heuristic for identifying the core content in the page, rather than supplementary text such as navigation menus and page footers.

The parameter posteriors for the mixture model are inferred using Gibbs sampling. We averaged the parameter values from every 100 sampling iterations, collected from 4000 iterations after a 2000-iteration burn-in. The θ parameters were given Dirichlet(0.01) priors, and λ was given a Beta(9000, 1000) prior, so that high values of λ favoring the background distribution are a priori more likely, resulting in stronger feature associations.

When encoding feature values for the mixture model (i.e., the number of times each feature is observed within each class label), we used fractional values when annotators included multiple labels for a query. Since annotators provide class labels, we used fractional values when annotators inferred using Gibbs sampling. We averaged the parameter values from every 100 sampling iterations, collected from 4000 iterations after a 2000-iteration burn-in. The θ parameters were given Dirichlet(0.01) priors, and λ was given a Beta(9000, 1000) prior, so that high values of λ favoring the background distribution are a priori more likely, resulting in stronger feature associations.

When encoding feature values for the mixture model (i.e., the number of times each feature is observed within each class label), we used fractional values when annotators included multiple labels for a query. Since annotators provide class labels, we used fractional values when annotators inferred using Gibbs sampling. We averaged the parameter values from every 100 sampling iterations, collected from 4000 iterations after a 2000-iteration burn-in. The θ parameters were given Dirichlet(0.01) priors, and λ was given a Beta(9000, 1000) prior, so that high values of λ favoring the background distribution are a priori more likely, resulting in stronger feature associations.

When encoding feature values for the mixture model (i.e., the number of times each feature is observed within each class label), we used fractional values when annotators included multiple labels for a query. Since annotators provide class labels, we used fractional values when annotators inferred using Gibbs sampling. We averaged the parameter values from every 100 sampling iterations, collected from 4000 iterations after a 2000-iteration burn-in. The θ parameters were given Dirichlet(0.01) priors, and λ was given a Beta(9000, 1000) prior, so that high values of λ favoring the background distribution are a priori more likely, resulting in stronger feature associations.

1 We used existing software for cross-collection Latent Dirichlet Allocation (ccLDA) [30], a topic model that learns topics for multiple collections of text as well as collection-independent background topics. This two-component model is a special case of ccLDA with only one topic.

Results.

Table 3 shows the top features for each phase, displaying the highest probability n-grams and domain names under each θ^k. For space and simplicity, the table only displays bigrams and not trigrams.

We see that the general query “treatment options” is associated with both the initial and secondary decision phases. The initial decision phase is also associated with queries containing the trigram “pros and cons”, as well as explicit comparative n-grams like “surgery vs radiation”. The decision phase also has a high probability of queries for “active surveillance” and “watchful waiting”—options for non-treatment that do not apply to later phases, once treatment has started. The page n-grams are similar, with general terms regarding treatment options. The bigrams “clinical trial(s)” are highly probable in the secondary decision pages.

A top query trigram for the initial preparation phase is “what to expect”, while many of the top query n-grams for initial post-treatments are variants of “after surgery” or “after prostatectomy”. Searchers look for general recovery information, as well as information about performing various activities after treatment (e.g. “sex after”) and treatment side effects (e.g. “incontinence after”). The n-grams from retrieved pages for these two phases both include a number of treatment side effects (“erectile dysfunction”, “urinary incontinence”), as well as n-grams containing the word “catheter”.

The top search n-grams for all secondary phases include medications used in hormonal therapies (e.g. “lupron”, “zytiga”), and “adjuvant radiation”, referring to a type of radiation that is given after the initial surgery. The top page n-grams for the secondary phases have terms related to drugs and their side effects.

We see that youtube.com is the top domain name for initial preparation. In general, videos are associated with the initial preparation and decision phases. 26.2% and 28.2% of users visited pages with “video(s)” in the title or URL during initial preparation and decision sessions, while only 8.3% of users visited such pages during the initial post-treatment sessions. Almost no users visited video pages during the secondary phases.

We also observe that the initial recovery phase contains many n-grams containing first person pronouns in the top page content n-grams, and cancerforums.net is the top domain name. There is an association of forums with the initial post-treatment phase. 48.8% of users visited pages with “forum(s)”, “discussion(s)”, or “community” in the title or URL during this phase. This is substantially higher than the percentage during the initial decision (35.9%) or secondary decision (37.1%) phases, which had the next highest percentages of such visits.

The top domain names associated with the background distribution in the mixture model (the distribution independent of phase) are webmd.com, cancer.org, cancer.gov, ask.com, and ehow.com. These are the most probable domains visited, even though many of these are not associated with particular phases, and so do not appear in Table 3.

5.2 Evolution of Queries on Treatments

We now focus specifically on the “initial decision” phase, with the goal of seeking an understanding of the sequential patterns of information gathering about treatments and outcomes during decision making.
### Table 3: Most probable bigrams associated with each of the six phases of treatment queries. Bigrams are estimated from the relevant search queries and text content and domain names of pages visited following those queries.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>prostate cancer</td>
<td>after prostate</td>
<td>after prostate</td>
<td>after a</td>
<td>adjunct radiation</td>
<td>seed implants</td>
</tr>
<tr>
<td>cancer treatment</td>
<td>surgery for</td>
<td>after prostate</td>
<td>a radical</td>
<td>how much</td>
<td>hbr treatment</td>
</tr>
<tr>
<td>proton therapy</td>
<td>robotic prostatectomy</td>
<td>after prostate</td>
<td>radical prostatectomy</td>
<td>taking lupon</td>
<td>pain in</td>
</tr>
<tr>
<td>best treatment</td>
<td>on the</td>
<td>after radical</td>
<td>what are</td>
<td>seed implants</td>
<td>cause pain</td>
</tr>
<tr>
<td>prostate</td>
<td>what to</td>
<td>radical prostatectomy</td>
<td>after radical</td>
<td>psa</td>
<td>radiation burns</td>
</tr>
<tr>
<td>treatments</td>
<td>home on</td>
<td>incontinence after</td>
<td>are the</td>
<td>after radical</td>
<td>treatment cause</td>
</tr>
<tr>
<td>treatment options</td>
<td>the same</td>
<td>treatment</td>
<td>include radiation</td>
<td>radiation after</td>
<td>lupon treatment</td>
</tr>
<tr>
<td>active surveillance</td>
<td>cinic</td>
<td>therapy</td>
<td>cancer treatment</td>
<td>cancer treatment</td>
<td>after seed</td>
</tr>
<tr>
<td>and cons</td>
<td>vin prostate</td>
<td>the side</td>
<td>protect</td>
<td>treatment after</td>
<td>not effective</td>
</tr>
<tr>
<td>surgery for</td>
<td>same day</td>
<td>prostate be</td>
<td>will i</td>
<td>treatment after</td>
<td>psa after</td>
</tr>
<tr>
<td>da vinci</td>
<td>go home</td>
<td>treatment after</td>
<td>do i</td>
<td>adjuvant radiation</td>
<td>flomax not</td>
</tr>
<tr>
<td>surgery vs</td>
<td>davin prostate</td>
<td>after robotic</td>
<td>to avoid</td>
<td>radiation</td>
<td>effective</td>
</tr>
<tr>
<td>watchful waiting</td>
<td>prostatectomy</td>
<td>after prostate</td>
<td>prostate seed</td>
<td>prostate</td>
<td>for high</td>
</tr>
<tr>
<td>vs radiation</td>
<td>for radical</td>
<td>for radical</td>
<td>with catheter</td>
<td>treatment</td>
<td>lupon</td>
</tr>
<tr>
<td>treatment for</td>
<td>expect</td>
<td>do i</td>
<td>zytiga cost</td>
<td>after high</td>
<td>radical</td>
</tr>
<tr>
<td>cyberknife prostate</td>
<td>day of</td>
<td>what to</td>
<td>catheter in</td>
<td>high</td>
<td>prostate</td>
</tr>
<tr>
<td>cons of prostate treatment</td>
<td>life after</td>
<td>after lupon</td>
<td>lupon treatment</td>
<td>after radiation</td>
<td>radical</td>
</tr>
<tr>
<td>prostate treatment</td>
<td>is surgery</td>
<td>levels after</td>
<td>radiation after</td>
<td>radiation</td>
<td>proctomy</td>
</tr>
<tr>
<td>the best</td>
<td>cryotherapy surgery</td>
<td>blood in</td>
<td>be next</td>
<td>taking casodex</td>
<td>enlarged</td>
</tr>
<tr>
<td></td>
<td>do legel exercises</td>
<td>long does</td>
<td>post psa</td>
<td>on lupon</td>
<td>abdomen</td>
</tr>
</tbody>
</table>

#### Number and Specificity of Treatments.

To understand the progression of treatment-related searches, we examine the average depth of the treatments searched (as defined in the treatment hierarchy displayed in Table 1) and the average cumulative number of treatments searched, as functions of the number of treatment queries conducted. These metrics can provide insights about the decision-making process, including changes over time in the specificity of treatment-related searches and the overall number of different treatments studied by the average user.

Figure 1 shows these values averaged across users for the first 11 initial decision queries, which is the average (10.9) number of queries in this phase (median 8).

We find that the first query is, on average, the broadest, with an average depth below 1, which means the bulk of initial queries contained general terms such as “treatment options”. The depth increases nearly monotonically for the first 11 queries, showing that searchers ask for progressively more specific information. Beyond 11 queries, this trend levels off and becomes noisier, as the results are averaged among fewer users.

The average cumulative number of treatments searched in the initial query is 0.65, which means that only 65% of searchers specified a treatment initially, and the remaining 35% conducted a general query such as “best treatment options”. After 11 queries, the average user has searched for 2.4 different treatments.

#### Treatment Comparisons.

We next consider treatments that are referenced together within the same query. Queries with multiple treatments are likely comparative, and we indeed observed a number of queries with explicit comparative language such as “vs”, as highlighted in our n-gram characterization of the initial decision phase in the previous subsection.

We found that 9.6% of initial decision treatment queries contained more than one treatment in the same query, and 43.6% of users searched at least one query with multiple treatments. Examining the treatments that co-occurred, we found that 75% of such queries contained surgery and radiation, 7.3% contained different types of surgery, 7.3% con-
tained surgery and observation, 6.3% contained radiation and hormone therapy, and 4.2% contained different types of radiation. Of the co-occurrences of surgery and radiation, 65.3% were for the most general terms (e.g., “surgery vs radiation”), while the others contained more specific types (e.g., “robotic surgery or seed implants”).

Transitions among Treatments.

Finally, we examined the transition structure among different treatment types by analyzing the sequences of treatments that appear in consecutive queries within the same search session. These experiments were undertaken to pursue insights about how searchers refine their queries as they explore treatment options.

We examined how successive queries progress through the treatment hierarchy in Table 1. We found that 68.8% of the time, the same treatment is searched as in the previous query. In 12.7% of the cases, the subsequent query is more specific than the previous one, going deeper down the same branch in the hierarchy (e.g., searching “robotic surgery” after searching “surgery”), while the subsequent query is coarser (higher up along the branch) in 9.5% of cases. We found that 9.0% of subsequent queries are situated in an entirely different branch of the hierarchy (e.g., searching radiation after searching surgery).

We also explored which treatments are likely to be searched after a preceding treatment query. We construct a transition graph, where treatment categories are nodes, and directed edges are weighted by the number of times that one treatment was searched after the other. The graph includes a dummy START node, whose outgoing edges to each treatment node are weighted by the number of times that each treatment was in a user’s initial query. To show a concise visualization of the typical transitions among treatments, we compute the maximum directed spanning tree of this dense graph, using the Chu-Liu-Edmonds algorithm [7, 12].

The induced tree is shown in Figure 2. The general treatment category (e.g., for non-specific queries such as “treatment options”) and the coarsest surgery category follow from START. This may be taken as intuitive as our results above showed that search histories most often begin with the general treatment category, and searches for surgery are the most common of all specific treatment types. Many of the edges are the same as those captured by the curated treatment hierarchy in Table 1: hormone therapy and cryotherapy following the general treatment category, robotic surgery following surgery, and external beam radiation following radiation. Other edges do not follow the natural hierarchy, but instead illustrate a typical order of treatments that searchers pursue via their queries. For example, searches for radiation and observation are most likely to follow surgery searches, which might be expected in light of the finding above that the most common comparative searches contain surgery and radiation, or surgery and observation.

The lowest-weight edges (weight of 3) are excluded to keep the tree concise, and because very low weight edges are likely to introduce noise. (Recall that edge weights are the number of times each treatment category was searched after the other.) One edge was from “Treatment” to “HIFU”, which fits with the hierarchy in Table 1. The other edge was from START to “Open surgery”, which is not a sensible result, but open surgery happened to be searched in the first query (3 times) more than after other queries.

Figure 1: Left: Specificity of treatments searched over time by the average user during the initial decision phase, as given by each treatment query’s depth in the treatment hierarchy in Table 1. Right: Cumulative number of different treatments searched over time by the average user. We use the mean number of queries in the decision phase (11 queries) as the range of the x axis.

Figure 2: Maximum directed spanning tree induced from the treatment query transition graph.
(a) Distribution over non-gap phases and content categories in each alignment column.

(b) The distribution over content categories in each alignment column, restricted to each particular phase.

(c) Multiple sequence alignment of 272 treatment timelines. Colored dots represent the label in each row/column, using the legend at the top of (a). White space represents gaps.

Figure 3: Different visualizations of treatment timeline alignments.
5.3 Progression of Phases and Search Content

We focused in the previous section on temporal patterns within the initial decision phase. We now seek to understand the temporal patterns across all phases. We wish to visualize how the phases progress over time, and how the content of treatment queries evolves over time within each phase and across entire timelines.

No individual tagged search history contained all six phases described in Section 4.4. However, we can align and stitch together the partial timelines to visualize an “average” complete timeline, aggregated across the hundreds of histories.

Toward this goal, we computed a multiple sequence alignment of the timelines. Multiple sequence alignment (MSA) methods were developed in computational biology and are typically used to build a molecular sequence via alignment of smaller sequence snippets. Alignments are scored based on how well symbols at each position align, penalizing gaps and mismatches. The optimization problem is then to solve for a single alignment that gives the highest score.

Solving for the best alignment between two sequences can be done efficiently with dynamic programming, using the same procedure that is used to compute string edit distance. The size of the dynamic programming table increases exponentially with the number of sequences, making this problem NP-hard for an arbitrary number of sequences [18], and impractical for more than a few. Many methods have been developed in computational biology to approximately solve for an MSA efficiently, such as the merging of pairwise alignments. For this experiment, we used ClustalW (from clustal.org), a software package for aligning protein sequences [22]. While protein alignments typically use domain-specific scoring functions, we created custom scores appropriate for our task.

Each treatment timeline was considered to be a sequence, and each phase label in the timeline was treated to be a symbol in the sequence. The most likely label was chosen in cases where annotators listed more than one possibility, using the annotators’ highest-ranked choice. We created special symbols for the first query with each phase label in the timeline, to encourage the start of each phase to align, so that they are not aligned to arbitrary positions. This means that there are 12 total symbols and 272 sequences.

Alignments are scored such that each position in the alignment is given a score of 1 if the symbols match. A score of 0.1 is given if the phases match, but one of the symbols is the special ‘first time’ indicator and the other is not. No credit is given for aligning different phases. We did not apply strong penalties for gaps, preferring alignments where different phases do not overlap. We used a penalty of 0.1 for gap creation, to discourage gaps with all other options being equal, with no penalty for introducing successive gaps.

Figure 3(c) shows the resulting MSA of the timelines. We see that the left is dominated by initial phases (blue) while the right is dominated by secondary phases (green), though the phases do not progress monotonically. We note that the initial post-treatment and secondary decision phases are often interleaved, as searchers tend to search for recovery or side effects following the initial treatment at the same time as they search for the next steps.

To more clearly see the phase progression, Figure 3(a) shows the distribution of the phases after gaps are removed for each column. To reduce noise, we excluded columns with less than ten non-gap symbols. This roughly halves the number of columns included in the visualization. The values are smoothed by averaging the values from the preceding/following three columns.

In addition to the distribution of phases, the figure also shows the distribution over various categories of search terms in each column. The categories are based on the term ontology (Section 4.1), including terms referencing treatments and side effects, as well as searches for healthcare providers, search terms referencing mental health or seeking social or emotional support, and searches that are aimed at retrieving statistics such as prognosis or success rates. We see some variation over time in the content. Searches for hormone therapy and prostate cancer medications (many of which are hormonal therapies, but categorized separately for this visualization) increase over time, while searches for observation only appear in the first half. The general term “side effects” is prominent initially and declines, and more specific terms for side effects (related to incontinence and impotence) begin to rise. This highlights a shift from a general interest in learning about side effects to more specific concerns.

Many of the differences in the category distributions over time are smoothed over due to overlapping phases at each point. Figure 3(b) shows the content distribution within each phase in isolation, clarifying the differences. For these images, we did not require a minimum number of non-gap values, other than excluding columns with only gaps. This reveals differences among the phases. For example, searches for healthcare (dark green) appear mostly in the initial decision phase, while searches describing mental health (yellow) appear mostly in the initial post-treatment phase. The treatment distribution differs between the initial and secondary phases, with fewer references to surgery and more references to hormone and chemotherapy in the latter.

6. AGE COMPOSITION OF SEARCHERS

We now present a final experiment focused on examining ages inferred for the experiential searchers in our dataset. We performed this experiment to understand the demographic composition of the dataset, as well as to provide an auxiliary form of validation, to determine whether the searchers in our dataset exhibit similar demographics as patients diagnosed with prostate cancer. Since rates of cancer are higher in older age groups than in the general population, we would expect to see a shift toward this demographic in our classified dataset, should our classifier be capturing a higher proportion of people experiencing cancer.
We associated age groups with searchers by looking for references to ages in search queries. Specifically, we matched queries against expressions of the form “at/age _ year(s) old”, and “in my/his/her _” for different numeric values. A similar self-reporting methodology was used to estimate the stage in pregnancy of expectant mothers or the age of new-borns, based only on search logs [13]. If a search history included multiple such expressions, the majority age group was chosen. We were able to associate ages with 142 user identifiers (out of the set of 1,413 classified).

We computed the distribution of age groups for the search histories identified by the classifier. For comparison, we computed the age distribution for the larger set of histories from our initial filter (those who searched “prostate cancer” three times), and the entire set of search logs from the most recent two months. From the two month sample, we estimated the expected distribution of ages in the logs among those diagnosed with prostate cancer in the United States (US). We computed this estimate using Bayes’ theorem: \( P(\text{cancer}|\text{age})P(\text{age}) \), where \( P(\text{cancer}|\text{age}) \) is defined by the age-specific prostate cancer incidence rates from the US National Cancer Institute (NCI),\(^2\) and \( P(\text{age}) \) is the distribution in the sample of logs (first column of Table 4).

Table 4 shows the distribution of age groups from 20s to 80s (the NCI data does not list rates for specific age groups beyond 80s) for the three sets of log data as well as the expected distribution from incidence rates.

The age distribution among positively classified searchers is strikingly similar to the expected distribution, particularly for the ages of 60s and 70s, which are each within 1 percent of the expected rate. The Pearson correlation between these two distributions is highly significant (\( r = .959, p < .001 \)). The distribution among users passing through our initial filter (three queries for prostate cancer) skews toward the older ages as might be expected. However, after applying the experiential classifier, the percentages further increase for the age groups with the highest incidence rates (people in their 60s and 70s) and decrease for the younger age groups. We believe this provides additional evidence that many of the classified searchers included in our study were likely to be experiencing a diagnosis of prostate cancer.

7. DISCUSSION AND IMPLICATIONS

A limitation of using search log analysis to learn about Web-based decision support is that we lack the larger context that frames the information-seeking activity. To address this shortcoming, we are developing methodologies to connect long-term log behaviors with self-reported data from consenting search participants. This approach would provide details on the context behind the search and retrieval activity appearing in logs of online activity. This information would allow us to understand the influence of Web content on a patient’s decision-making and details such as the resources that were helpful in deliberating about care decisions, whether a decision aligned with a doctor’s recommendation, and the outcome following a decision. Many of these details are impossible to infer from the observed activities in logs alone. Engaging directly with patients, and aligning online activity with patients’ clinical situations, would enable us to perform rich analyses grounded in detailed patient reports. While we hope to conduct such a study as future work, we also believe that there is complementary value in the lightweight, large-scale analyses presented in this paper.

We believe that there are opportunities to leverage the reported findings to inform the design of search and retrieval systems for supporting healthcare decision making. Given the focus of this paper, we are especially interested in enhancing the ability of Web search engines to serve as decision support systems, per the expectations that people appear to have when they turn to the Web for critical assistance with challenging treatment decisions under uncertainty. The different phases of treatment appear to be identifiable, with distinctive n-gram associations and timing characteristics. This suggests that search output can be tailored to the user’s current phase. For example, a searcher during a decision-making phase may find it helpful to gain access to results that include comparisons of treatments. We found that comparative queries are common, with nearly half of users conducting an explicit comparison, and from our experiments, we know which comparisons are most common. These insights could be used to return results, or interface treatments such as direct answers using data pulled from external sources, that include relevant comparisons even if the query did not explicitly include a comparison. We also now have data on the progression of queries on treatments, including the depth over time and transitions between treatments, which can be used to model treatment search behavior. Such models could be used by search providers for important tasks such as appropriately suggesting or expanding queries.

Beyond tailoring results to a current phase of search aligned with a phase of care, it may be valuable to provide searchers with content that is typically viewed in later phases of an illness. For example, searchers making a decision may find it useful to read the content that is commonly viewed by those in a post-treatment phase, in order to understand the expected recovery process and side effects from particular treatments. Our analyses showed that many people may seek out discussions of the personal experiences of patients (e.g., through forums) during the post-treatment phase, and surfacing the concerns and issues of others could provide searchers facing decisions with more context than traditional decision-making materials. Such content might not be discovered in the course of normal searching in an early phase without designing a system for such proactive retrieval, as queries in the initial decision phase tend to be broad.

Promising future directions include adapting the annotations, classifiers, and overall methodology to understand the information needs and to guide decision support for treatment decisions for other illnesses. Beyond the pursuit of enhancing search for decisions about treatments, we can employ the methods to enhance search and retrieval for other healthcare needs, such as selecting a care provider.

8. REFERENCES


[47] [841]