# Detecting Devastating Diseases in Search Logs

John Paparrizos*
Columbia University
jopa@cs.columbia.edu

Ryen W. White
Microsoft Research
ryenw@microsoft.com

Eric Horvitz
Microsoft Research
horvitz@microsoft.com

## ABSTRACT

Web search queries can offer a unique population-scale window onto streams of evidence that are useful for detecting the emergence of health conditions. We explore the promise of harnessing behavioral signals in search logs to provide advance warning about the presence of devastating diseases such as pancreatic cancer. Pancreatic cancer is often diagnosed too late to be treated effectively as the cancer has usually metastasized by the time of diagnosis. Symptoms of the early stages of the illness are often subtle and nonspecific. We identify searchers who issue credible, first-person diagnostic queries for pancreatic cancer and we learn models from prior search histories that predict which searchers will later input such queries. We show that we can infer the likelihood of seeing the rise of diagnostic queries months before they appear and characterize the tradeoff between predictivity and false positive rate. The findings highlight the potential of harnessing search logs for the early detection of pancreatic cancer and more generally for harnessing search systems to reduce health risks for individuals.

## Keywords

Health screening; Search logs; Behavioral data; Digital disease detection; Pancreatic cancer; Temporal analysis

## 1. INTRODUCTION

Web search is a primary resource for people concerned about the significance of health-related symptoms [16]. Researchers have studied symptom and illness-related searches in pursuit of insights about how people search about health concerns, including patterns of querying and review of information in pursuit of diagnoses [54], healthcare utilization signals [55], traces of therapeutic decision making for challenging illnesses [40], and identification of new adverse effects of medications [56, 57]. Prior studies have examined how population-level signals in social media can be used to detect the emergence of diseases [10, 19].

*Work conducted during a Microsoft Research internship. John Paparrizos is an Alexander S. Onassis Foundation Scholar.



*KDD '16, August 13-17, 2016, San Francisco, CA, USA*
 ISBN 978-1-4503-4232-2/16/08…$15.00
DOI: http://dx.doi.org/10.1145/2939672.2939722

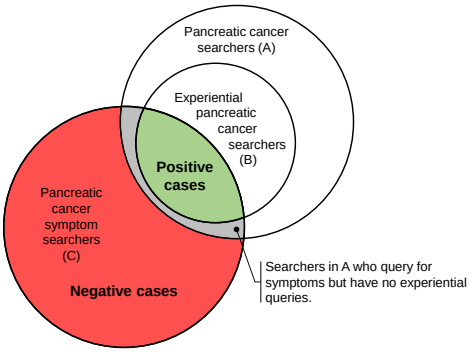


**Figure 1:** Venn diagram depicting sets of users employed in analyses: pancreatic cancer searchers ($A$), pancreatic cancer searchers exhibiting experiential diagnostic queries ($B$), and those who search for the symptoms of pancreatic cancer ($C$). $|A \cup C|$ (i.e., number of users in original, pre-filtered dataset) is 9.2 million. Positives and negatives are sourced from $B \cap C$ and $C \setminus A$, respectively. Relative set sizes are not to scale.

We explore the prospect of harnessing anonymized long-term sequences of health-related search queries to yield information that could provide valuable signals for detection of illness in advance of traditional diagnosis. Leveraging online behavioral data to provide earlier detection of a disease or of the raised risk of illness on a large scale can make significant contributions to healthcare. Better outcomes can be achieved by earlier confirmation of illnesses and risks via gaining access to more timely diagnoses, treatments, and other proactive interventions. As an example, such capabilities might help to identify those at significant risk of suffering the onset of advance of chronic disease processes such as diabetes or heart disease or the rise of acute processes such as atrial fibrillation or more severe cardiac arrhythmias. Interventional programs ranging from changes in diet or exercise to taking (or avoiding) certain medications can yield significant health benefits.

The diagnosis for certain medical conditions can be particularly devastating if the chances of survival are typically low at the time of diagnosis. Survival rates may be improved significantly via earlier detection and treatment. With many cancers, screening methods can be effective for early detection and therapy [41], but they involve explicit testing prompted by policies around risk factors such as family history [33] or medical history [14, 32]. Ideally, health screening systems would be able to observe people passively as they engage in their normal activities and alert them (per their preferences on vigilance) to potential health risks with-

out requiring the investment of time and effort in special screening activities. Individual-level postings on social media have been mined for this purpose [10, 46], but may not have representative coverage of symptoms associated with social stigma [11].

We present a study of the feasibility of doing early detection of devastating diseases based on large-scale logs of health-related Web search activity. We consider the content of queries over time, and the prospect that temporal relationships and patterns among queries over multiple sessions over several months provide subtle fingerprints of lurking illness. We focus on the early detection of the presence of pancreatic cancer, a devastating diagnosis given the typical progression of the disease to an inoperable situation by the time it is found. Our work showcases the potential value of innovative approaches for speeding up the time to diagnosis of this deadly disease.

Pancreatic cancer is the fourth leading cause of cancer death in men and women in the United States and the sixth leading cause in Europe [36]. The illness is widely known as being difficult to detect and is frequently diagnosed too late to be treated effectively [22, 30]. A recent study found that the progression of pancreatic cancer from stage I to stage IV happens in just over one year [59]. Approximately 75% of pancreatic cancer patients die within a year of diagnosis, and only about 4% survive for five years post diagnosis. Exploration of the possibility that a patient has pancreatic cancer involves a careful and costly consideration of history, labwork, and imaging studies (in contrast to the passive screening methods described in this paper) [45]. Screening is largely performed to detect the disease at an early phase (pre-invasive or early invasive) when it is still curable by surgical intervention and chemotherapy. Earlier diagnosis of pancreatic cancer improves the feasibility of discovering the illness at an earlier stage [7]. For patients diagnosed early who undergo curative surgery (e.g., a Whipple procedure), five-year survival rate is higher, but it remains less than 25% [58].

We take as a proxy for ground truth of a diagnosis of pancreatic cancer the detection of *experiential diagnostic* queries issued by searchers. Experiential queries show strong evidence of being linked to the actual presence of symptomatology or conditions versus less directly involved, more distant *exploratory* queries seeking information about symptoms or diseases [40]. Experiential diagnostic queries for pancreatic cancer are identified via consideration of the structure of queries and of patterns of information gathering over multiple users in search logs. Experiential queries often include first-person assertions such as [*i was just diagnosed with pancreatic cancer*], which when associated with prior queries about symptoms identifies the positive cases.

We construct models to predict the future rise of experiential queries from longitudinal search data. Figure 1 shows the different subsets of users in our analysis, including people who search for pancreatic cancer ($A$), the subset of these searchers who issue experiential diagnostic queries ($B$), and those who search for a set of symptoms linked to pancreatic cancer ($C$). Those who *only* search for one or more related symptoms with no evidence of pancreatic cancer searching constitute the negative cases. We find that our methods can detect cases where people show evidence of being diagnosed with pancreatic cancer many months in advance of their experiential diagnostic queries.

We make the following contributions with this research:

- Introduce early detection of diseases as a promising new application of search log mining and machine learning that scales to millions of searchers.
- Present a case study on the early detection of pancreatic cancer from longitudinal individual search activity.
- Forecast with significant lead times that users will later input experiential queries for pancreatic cancer.
- Explore the influence of different factors, such as the lead time or the presence of specific symptoms in the search activity, on the predictive performance of our learned models, including true positive rates when false positive rates are strictly controlled. Controlling false positives is especially important to reduce unnecessary costs and concerns given potential future applications such as providing early warnings and suggestions to searchers about undertaking more formal screenings.

We now describe related research in this important area.

## 2. RELATED WORK

Related research in a number of areas is relevant to our work. These include (i) health searching; (ii) large-scale analysis of search behavior; and (iii) methods for the early detection of disease, with a focus on pancreatic cancer.

The Web is an important source of health-related information for many people. To better understand how people pursue health information, studies have examined online health search using a variety of methods, including interviews [42], surveys [49], and analyses of large-scale search log data [2, 5]. 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 following the search [16]. Despite the potential benefits, concerns have been raised about the quality of online health information [8]. In a large-scale survey of the use of search for self-diagnosis, White and Horvitz [53] found that almost 40% of participants experienced increased anxiety from searching health information online. Studies have characterized problems with symptom search, including the influence of poor accounting for base rates of diseases and people's bias to focus on results covering serious illnesses versus more likely benign explanations. Such biases can lead to inappropriate anxiety [28, 53] and highlight the criticality of studying how patients use the Web, including the nature and dynamics of queries, and content delivered in response.

There has been a large amount of research on the analysis of search behavior from search engine logs. Log analysis provides insights to understand how people engage in information seeking in online settings [51], while also having applications for tasks such as result ranking [1, 24], query suggestion [25], prediction of future search actions and interests [13, 27], and detection of real-world events and activities [44]. Given access to population-scale data on how people search for health information, this can be applied for important tasks such as the detection of influenza [19], the detection of adverse drug reactions [56], population-scale studies of nutrition [50], epidemiology [19], and studies of chronic medical conditions such as pregnancy [15]. Related to this research, but focused on activity post-diagnosis, are studies of cancer-related searching [3, 6, 21], some of which have revealed strong similarities between temporal patterns

in search logs and those in practice [38, 40]. Studies have leveraged online behavioral signals for early disease detection at the population level [19], and individually [11, 46].

Screening high-risk individuals for pancreatic cancer is the only practical approach to detect precancerous or cancerous changes in the pancreas at the phase in which surgical intervention will have a high chance of cure [26]. Risk level can be determined by factors such as race [9], family history [33], and a history of pancreatitis [32]. Imaging studies via methods such as endoscopic ultrasound, computer tomography scans, and magnetic resonance imaging [35, 37] have been useful to diagnose pancreatic cancer once the tumor is large enough to cause unusual, salient symptoms that induce people to seek medical attention (e.g., yellow eyes, changes in stool), but at this point the disease is more likely to be at an advanced and unresectable stage (i.e., locally advanced or metastatic, when it cannot be removed by surgery) [29]. Common, seemingly innocuous symptoms such as back pain, abdominal pain, itchy skin, unexplained weight loss and nausea (and combinations and temporal patterns of these and other symptoms) may also be observed in the query stream. Such symptom searches can provide patterns of symptoms that might one day be employed in new kinds of health surveillance systems. Such systems could be used to alert people who would otherwise not feel moved to see a healthcare professional.

Active, explicit screening for early signs of pancreatic cancer is not cost effective unless there is a reasonable probability of detecting invasive or pre-invasive disease (at least 16% according to one study [45]). A log-based methodology provides scale that is not achievable with more traditional epidemiological studies, which tend to be on the order of tens or hundreds of participants, e.g., [23, 43].

# 3. DATASET CREATION

We now describe the data used, starting with a description of the logs (Section 3.1). We then discuss the creation of an ontology with symptoms commonly experienced by people with pancreatic cancer (Section 3.2) and provide details on extracting pancreatic cancer and symptom searchers (Section 3.3). We review the augmentation, tagging, and filtering steps for our dataset (Section 3.4). Finally, we summarize the creation of query timelines for the positive cases (i.e., experiential diagnostic searchers who also search for pancreatic cancer symptoms) and the negative cases (i.e., those who only search for the symptoms) (Section 3.5). Since reliable labels cannot be determined for the non-experiential pancreatic cancer searchers, we exclude them to create a cleaner dataset for training and testing. We show later (see Section 5.6) that predictive performance is largely unchanged if these searchers are included as negative examples during the application of the model in a realistic scenario.

## 3.1 Anonymized Web Search Engine Logs

Search engines track various characteristics during their interaction with users so as to better capture information needs, improve their responses, and personalize the content. Every such interaction corresponds to a log entry that includes a unique, anonymized user identifier based on a Web browser cookie. This enables the extraction of the search history comprising queries and clicks from an identifier for up to 18 months. Note that the identifier may comprise the search activity of multiple users on shared machines and does not consolidate activity from a user across multiple machines. We use the logs of a randomly-selected subset of Bing search engine users in the English-speaking United States locale from October 2013 to May 2015 inclusive.

## 3.2 Symptoms and Risk Factors

Warning signs and symptoms for pancreatic cancer usually include generic, subtle signs and symptoms, such as abdominal and back pain, loss of appetite, and unexplained weight loss. We performed an extensive review of possible signs, symptoms, and risk factors associated with pancreatic cancer and developed an ontology with 21 categories of symptoms. This manually-curated ontology consists of two levels. The first level includes the names of the symptoms and the second level includes multiple names, synonyms, and expressions with which the corresponding symptom in the first level may appear in our data. We performed multiple iterations of refinements of this ontology to remove noise and to minimize erroneous query matches. Table 1 presents the 21 symptom categories with some representative examples of associated query expressions. Also shown are 12 risk factors and associated synonyms, derived from the literature (e.g., [31]), describing attributes, characteristics, or exposures that may increase the likelihood of pancreatic cancer. The symptoms and the risk factors are featurized in predictive models, and they are also used in policies to determine when predictive models should be applied (see Section 5.5).

## 3.3 Extraction of Searchers

In order to identify positive and negative cases for the generation of our learned model, we built a dataset comprising two groups of users (Figure 1). The *pancreatic cancer searchers* group, denoted as $A$ in the figure, includes all searchers with at least one query explicitly on pancreatic cancer (i.e., a query matches this expression [('pancreas' *OR* 'pancreatic') *AND* 'cancer']). The *symptom searchers* group, denoted as $C$, includes all users with at least one query related to symptoms linked to pancreatic cancer, as captured by the symptoms and synonyms described in Section 3.2.

Having unique identifiers for each user in the union of $A$ and $C$ (i.e., $A \cup C$) permits the extraction of the full query histories of 9.2 million searchers. We first sought to remove searchers who are likely healthcare professionals (HCPs). To do this, we employed a proprietary Bing classifier that identifies health-related queries to remove users from the study for whom 20% or more of queries are health related. This threshold was based on a prior analysis of identifying health professionals in search logs [52].

## 3.4 Dataset Augmentation

Age and gender are important factors associated with developing pancreatic cancer [31, 36]. As such, we augmented the dataset with demographic information from proprietary search engine classifiers that estimate age (discretized as $<$ 18, 18–24, 25–34, 35–50, or 50–85) and the gender for each user. The classifiers are trained on data where ground truth of demographic details are provided explicitly by users. The predictions are based on signals derived from searchers' long-term search activity, including their search queries and Web domains of their clicked results. Since pancreatic cancer incidence rates vary by geographic location, we also annotated searchers with the U.S. state from which they searched most (based on reverse Internet provider (IP) lookup data).

**Table 1:** Ontology with symptoms, risk factors, and examples of associated synonyms for pancreatic cancer.

| Type | Name | Example synonyms |
|---|---|---|
| Symptom | back pain | pain in lower back, lowback pain |
| | blood clot | blood clots, thrombosis |
| | dark or tarry stool | dark poop, tarry feces |
| | dark urine | orange pee, brown urine |
| | enlarged gall bladder | swollen gallbladder, inflamed gallbladder |
| | floating stool | floating stool, floaters |
| | greasy stool | greasy poop, oily feces |
| | high blood sugar | frequent urination, sudden diabetes |
| | itchy skin | skin itching, hands itchy |
| | yellow skin or eyes | jaundice, yellow eyes |
| | light stool | pale stool, white crap |
| | loss of appetite | poor appetite, decreased appetite, not hungry |
| | nausea or vomiting | throwing up, nauseous |
| | smelly stool | stinky feces, smelly poop |
| | sudden weight loss | unexplained weight loss, weight loss sudden |
| | taste changes | changes in taste, dysgeusia |
| | loose stool | loose feces, loose stool, diarrhea |
| | constipation | constipated, backed up |
| | indigestion | acid reflux, heartburn |
| | abdominal swelling or pressure | swollen stomach, pressure abdomen |
| | abdominal pain | belly pain, stomach ache |
| Risk factor | alcoholism | heavy drinking, alcoholics anonymous, alcoholic |
| | hepatitis | hep b, hep c |
| | pancreatitis | – |
| | ulcers | ulcer |
| | obesity | obese, very fat, extremely fat |
| | smoking | smoker, cigarette, cigar |
| | chills or fever | chills, fever |
| | multiple endocrine neoplasia | men1 |
| | hereditary nonpolyposis colorectal cancer | lynch syndrome, hnpcc |
| | von hippel-lindau syndrome | hippel-lindau syndrome |
| | hereditary intestinal polyposis syndrome | peutz-jeghers syndrome |
| | familial atypical multiple mole melanoma syndrome | fammm, b-k mole syndrome |

Beyond the demographic information, we are also interested in the subject matter of the queries and results that were visited over searchers' timelines. We augmented each query and corresponding clicked websites with their estimated Open Directory Project (ODP, dmoz.org) category. We used a text-based classifier, similar to [4], that uses logistic regression to predict the ODP categories. When optimized for the score in each category, this classifier has a micro-averaged F1 score of 0.60. For queries, the ODP category is that of the top-ranked search result. The remaining users after the augmentation and filtering steps total 7.4 million, from which 479,787 are pancreatic cancer searchers.

## 3.5 Positive and Negative Cases

We create *query timelines* for experiential pancreatic cancer searchers and experiential symptom searchers which we then featurize for the early detection task. Figure 2 summarizes the strategies for identifying positives and negatives. To avoid including users with very short histories, we filter out all users with less than five search sessions[1] spanning five different days. This reduced the population to 6.4 million users, with a mean total duration (time from first to last query for a user) of 210.32 days, standard deviation (SD) of 182.93 days, and interquartile range of 120 days.

**Positive Cases:** To identify experiential pancreatic cancer users, we created a set of first-person diagnostic queries for pancreatic cancer (denoted $Exp_0$). Some examples of such diagnostic queries are [*just diagnosed with pancreatic cancer*], [*why did i get cancer in pancreas*], and [*i was told i have pancreatic cancer what to expect*].

From the set of 479,787 pancreatic cancer searchers, 3,203 match the pattern of diagnostic queries. In order to consider them as experiential users, we require them to have searched at least for one symptom *prior to* the diagnosis query. This step generates a set of 1,072 query timelines of experiential searchers that contain periods of *symptom lookup* followed by the diagnostic query. The symptom lookup period starts when the first symptom is detected as matching terms represented in the symptom ontology. For positive cases, the symptom lookup period completes at least one week before diagnosis (i.e., we consider the week before the diagnostic query as the period of diagnosis in real life and do not count queries in that time since they may be polluted with ground truth signals). For reliability, especially given the need to compute *Temporal* features (Section 4.2), the minimum time period for which our features are computed is four weeks.

**Negative Cases:** To generate the negative set, we sample from the users who searched for pancreatic cancer symptoms but did not search for pancreatic cancer anywhere in their timeline (i.e., $C \setminus A$), either before or after the symptom lookup period. Performing this additional check on data

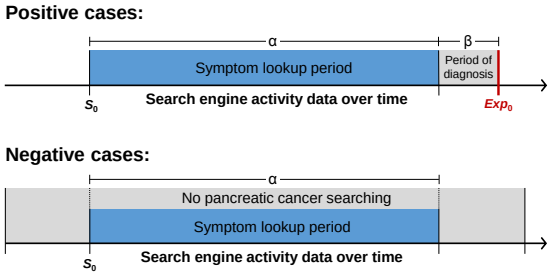


**Figure 2:** Schematic illustration of the query timelines used in the selection of positive and negative cases. $S_0$ refers to the first symptom query and $Exp_0$ is the first experiential diagnostic query. $\alpha$ is the duration of the symptom lookup period, which was approximately equal in the aggregate for the positives and negatives. $\beta$ is the duration of the period of diagnosis, set to one week in this study.

[1] *Session* is a query sequence with $\leq$ 30 minutes between queries [51].

**Table 2:** Summary statistics, namely, mean ($M$), standard deviation ($SD$), and number of cases ($N$), of durations and numbers of queries in positive and negative datasets.

| Class | Duration (days) | | Total # queries | | $N$ |
|---|---|---|---|---|---|
| | $M$ | $SD$ | $M$ | $SD$ | |
| Positives | 109.34 | 49.66 | 380.66 | 150.83 | 1,072 |
| Negatives | 108.04 | 48.35 | 378.09 | 151.01 | 3,025,046 |

from outside the lookup period is required to increase the likelihood that the negative cases are indeed negative. Users who searched for pancreatic cancer and its symptoms, but did not issue an experiential query (the gray subset in Figure 1 representing $(A \cap C) \setminus B$), were excluded since a label could not be reliably determined. In Section 5.6 we describe an additional experiment where pancreatic cancer searchers were included during model testing.

We were concerned that rudimentary behavioral differences that may reflect artifacts in the data could invalidate the learning task. For example, if our experiential users were just more active generally, then a feature that computed the total number of queries would have strong predictive value, yet would be uninteresting scientifically. We sought to address this by downsampling the negative cases to attain a similar distribution of symptom lookup periods in terms of the temporal duration and query volume as observed for the positive cases.[2] We did this by selecting users with a symptom lookup period duration within three standard deviations of the mean of the positive cases. This reduces the number of negative cases to 3,025,046. Table 2 presents the summary statistics on the symptom lookup periods in terms of the number of days and the number of queries in the two datasets. The table shows that the distributions for positive and negative cases (in terms of number of days and number of queries) are similar. The distributions are statistically indistinguishable using two-sample Kolmogorov-Smirnov tests for temporal duration ($D = 0.005; p = 0.7017$) and number of queries ($D = 0.003; p = 0.7681$), even though the latter was not a filtering criterion. We note that query timelines are not aligned: the absolute point in time where people issue the experiential diagnostic query, and the accompanying symptom lookup period can differ between searchers.

# 4. EARLY DETECTION

We now present the problem, summarize features extracted from query timelines, and review the prediction model.

## 4.1 Problem Description

We address the problem of early detection of experiential searchers for pancreatic cancer via anonymized Web search engine logs. We cast this as a binary classification task, where the model is trained on features extracted from search log query timelines of experiential pancreatic cancer searchers and symptom-only searchers. We focus on maintaining very low false-positive rates (i.e., 1 misprediction in 100k correctly identified cases) while keeping high the imbalance ratio of positive and negative cases (i.e., one thousand positives vs. millions of negative cases); these properties are important for potential future large-scale real-world applications such as an alerting mechanism in search engines.

[2]We could also have addressed this by making all features relative percentages. Sampling gave us more flexibility in feature construction. As an additional check, we included features such as the number of queries in the symptom lookup period; those were found to carry little evidential weight in the learned model.

## 4.2 Features

We now describe the features extracted from query timelines. We group our features into five different categories: (i) demographic information about the user; (ii) characteristics about user sessions, query classes, and URL classes; (iii) characteristics about symptoms; (iv) features that capture the temporal dynamics, and (v) risk factors.

**Demographics:** Cancer statistics from the U.S. National Cancer Institute[3] show that pancreatic cancer is more common with increasing age, is slightly more common in men than in women, and varies by geographic location. As such, we develop features related to the demographics of the users. In particular, we use the estimated age bucket and gender (see Section 3.4) along with the classifier's probabilities as confidence values. The dominant location (U.S. state) of a searcher is also included as a feature.

**Search Characteristics:** People express their information needs and preferences through queries and click behavior (i.e., the website visits). We extract various features to capture these search and retrieval activities. As we discussed previously, the queries, as well as the visited websites, were tagged with their ODP category in an attempt to identify domains of interest (see Section 3.4). A first set of features *SearchHistory* contains several generic statistics, such as counts, ratios, and percentages, which are characteristics of the global behavior of the user. For example, we compute the number of queries, sessions, and clicks, as well as ratios of clicks per query for each user. Then, we compute a large number of features with respect to the ODP categories of queries *QueryTopic*, clicked search results *URLTopic*, and the combination *QueryURLTopic*. These include compute counts and percentage of queries and sessions in each ODP category, the average time until queries appear in the same category, as well as the time of the day that queries appear in each category. Similar features are also computed for each category of the visited websites (e.g., counts, ratios, and percentages of visited websites that belong to each category). We additionally compute features to characterize the user sessions, including features that capture the click behavior of users associated with queries. For example, we compute counts and percentages of all the combinations of query categories that led to visits in website categories.

**Symptoms:** Features described above attempt to capture generic characteristics from user sessions. However, for the problem of interest, we seek to also leverage features from queries containing terms captured in the symptom ontology for pancreatic cancer (Section 3.2). The symptom features are divided into two classes: (i) *SymptomGeneric* and (ii) *SymptomSpecific*. Generic symptom features contain counts and percentages for the queries and sessions matching symptoms in our ontology, the average time between symptom queries, as well as the average number of symptom queries that are issued daily. Specific symptom features are generated per symptom category. For example, for each symptom, we compute counts and percentages of appearance, the time between distinct symptoms, and the time of day such symptom queries are issued. As with the user session features, we combine symptoms to capture the click behavior and, hence, we compute counts and percentages of each symptom query leading to a visit on a website belonging to particular ODP categories. Finally, we define features that capture the *se-*

[3]http://seer.cancer.gov/statfacts/html/pancreas.html

**Table 3:** Performance at four-week intervals for users where features can be computed from $Exp_0$ – 1 week to $Exp_0$ – 21 weeks. Values averaged across the ten folds of cross-validation. The significance of differences in AUROC and TPR using paired $t$-tests for each week versus $Exp_0$ – 1 indicated * $p < 0.01$, ** $p < 0.001$, and *** $p < 0.0001$. Weeks denote lead time before $Exp_0$ ($\beta$ in Figure 2).

| Weeks before $Exp_0$ | TPR (as %) at FPRs ranging from 0.00001–0.1 | | | | | AUROC |
|---|---|---|---|---|---|---|
| | 0.00001 | 0.0001 | 0.001 | 0.01 | 0.1 | |
| 1 week | 7.122 | 10.386 | 20.772 | 36.202 | 71.810 | 0.9112 |
| 5 weeks | 7.122 | 10.979 | 20.178 | 34.421 | 70.620 | 0.9047 |
| 9 weeks | 7.122 | 10.683 | 18.991* | 33.234* | 70.023 | 0.8854* |
| 13 weeks | 7.122 | 9.792 | 17.804* | 32.937* | 67.359* | 0.8700* |
| 17 weeks | 6.825 | 9.199* | 17.209* | 32.640** | 64.688** | 0.8539** |
| 21 weeks | 6.528* | 9.199* | 16.319** | 32.345** | 61.424*** | 0.8315** |

*quence* in which symptoms appear in query timelines.

**Temporal:** All previous features produce aggregated statistics over the full time window under consideration. Following [34], we include a set of features to capture the temporal variation of these statistics over misaligned query timelines with noise and missing values. For every feature, we generate a time series with points that represent aggregated values for intervals of the time window. For example, each feature can be computed per month, per week, or per day, depending on the level of granularity we seek to capture. Since the occurrence of specific features can be sparse, we set the time window to four weeks for temporal features. For features that are not percentages or ratios, we also compute the cumulative time series. For each time series, we use the first coefficient of the linear least-square estimates to devise features that capture the trend (i.e., increasing, decreasing, and unchanged) and the rate of change (i.e., slope).

**Risk Factors:** This class contains features related to the presence of terms representing risk factors in the symptom lookup period. For each risk factor, we note its presence or absence, and also the number of queries containing that risk factor and the fraction of all queries from that user that these risk factor queries represents. Total number of distinct risk factors in the symptom lookup period is also a feature.

## 4.3 Prediction Model

The prediction model uses the features outlined in the previous section, computed for each searcher, to make predictions about the future occurrence of experiential diagnostic searches in each searcher's query timeline. We use gradient boosted trees [17], which employ an ensemble of decision trees to construct a better learned model. Advantages include the ability to capture non-linear relationships, model interpretability (e.g., a ranked list of important features is generated), facility for rapid training and testing, and robustness against noisy labels and missing values. We experimented with different learning algorithms, but gradient boosted trees yielded superior accuracy.

# 5. EXPERIMENTAL RESULTS

We now present the findings of our experiments. We report the overall performance in Section 5.1 and the performance as we increase the lead time before the first experiential diagnostic query (Section 5.2). We then inspect the model to understand the contributions that each feature makes towards early detection (Section 5.3) and the performance of different feature *classes* (Section 5.4). We examine the effect on model performance of conditioning on symptoms and risk factors (Section 5.5) and consider a realistic deployment scenario (Section 5.6). We use the area under the receiver operating characteristic curve (AUROC) and recall (TPR, true positive rate) at fixed, extremely low false positive rates (FPRs) as our primary evaluation metrics. We applied 10-fold cross validation, stratified by user to evaluate the generalizability of the model when applied to new users. Significance level is $p < 0.05$ unless stated.

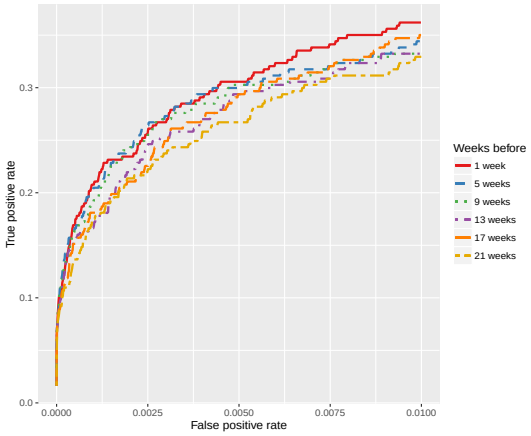


**Figure 3:** Average partial ROC curves in the FPR range 0–0.01, for models learned using data up to 21 weeks before the first experiential diagnostic query (error bars excluded for clarity). Variance in FPR and TPR is minor.

## 5.1 Overall

The overall performance of the classifier in making predictions based on data up to the beginning of the period of diagnosis (i.e., $Exp_0$ – 1 week) in AUROC is 0.9003. Given that low error rates would be vital in practice to avoid unnecessary patient alarm, we focus on the true positive rate (fraction of all positives that are recalled by the model) at low false positive rates (FPR). Focusing on FPR in the range 0.00001–0.01, the model is able to recall 5–30% of the positive cases, depending on the specific FPR. We see this performance as promising given the limited information (primarily search-related activity) available to the model.

## 5.2 Performance by Week

A key part of early detection is being able to predict the emergence of the disease well in advance. To understand how prediction performance changed as we move further back in time before the first experiential diagnostic query we selected the set of 337 positive searchers and 945,394 negative searchers who were still observed in the logs many weeks prior to the experiential diagnostic query. We report results from one week before the experiential diagnostic query, all the way up to 21 weeks before the diagnostic query. To count as being present at $Exp_0$ – 21 weeks, a searcher needs to have symptom queries extending back at least four weeks before that point (i.e., to $Exp_0$ – 25 weeks, or approximately six months before the first experiential diagnostic query).

We trained a model for these users in the same way as we did for Section 5.1. The ratio between positive and negative searchers remains similar to that for all users (i.e., approximately 1:3000). Table 3 reports the TPR at different false positive rates for this same set of users at different four-week increments, as well as the AUROC. The general trend is that the performance drops fairly consistently as we increase the lead time, but even 20 or so weeks before the first experiential diagnostic query the predictive perfor-

**Table 4:** Top 10 features by evidential weight relative to the top feature. "Positive" or "Negative" direction means that a feature correlates positively or negatively, respectively, with the rise of experiential queries.

| Feature | Weight | Direction | Class |
|---|---|---|---|
| NumOfDistinctSymptoms | 1.0000 | Positive | SymptomGeneral |
| NumOfQueriesInHealthCategory | 0.8253 | Positive | QueryTopic |
| NumOfDistinctSymptomsVariants | 0.6899 | Positive | SymptomGeneral |
| AgeClassProbability_5085 | 0.6889 | Positive | Demographic |
| HasBackPain | 0.6622 | Negative | SymptomSpecific |
| HasIndigestion | 0.6432 | Negative | SymptomSpecific |
| HasIndigestionThenAbdominalPain | 0.6349 | Positive | Temporal |
| SlopeNumOfDistinctSymptoms | 0.6154 | Positive | Temporal |
| HasBackPainThenYellowSkinOrEyes | 0.6004 | Positive | Temporal |
| AgeClassProbability_LT18 | 0.5869 | Negative | Demographic |

mance is still quite strong (AUROC=0.8315, TPR=6.528% at FPR=0.00001). Assuming that pancreatic cancer progresses steadily from stage I to stage IV in just over one year (as has been previously reported [59]), accurate predictions 20 weeks in advance of the diagnostic period could lead to a sizable increase in the five-year survival rate (e.g., moving the point of diagnosis from Stage III to Stage II could increase the survival rate from 3% to 5–7% [47]).

Focusing on the FPR region from 0 to 0.01 (i.e., false positives occur less than 1 in 100 times) and visualizing that part of the ROC curve (Figure 3) we observe some clear differences in the performance of the models in this important region. The average normalized partial AUROC ranges from 0.292 ($Exp_0$ – 1 week) to 0.231 ($Exp_0$ – 21 weeks). All differences in AUROC for $Exp_0$ – 5 or more weeks versus $Exp_0$ – 1 week are significant ($p < 0.01$ using paired $t$-tests).

## 5.3 Feature Contributions

In addition to understanding the overall performance, we are also interested in understanding the features that are most important in the learned model for predicting the future issuance of experiential queries. Table 4 shows the top 10 features with the highest weight, along with their weight relative to the top-ranked feature (*NumOfDistinctSymptoms*) and the feature class. The direction is based on the correlation between the feature value and the labels in the training data, using Pearson biserial correlation or the phi coefficient, depending on whether or not the feature data is binary. Table 4 shows that there is a broad range of features. The number of distinct pancreatic cancer symptoms was the most important feature. Temporal features representing changes over time and sequence ordering of symptom pairs are also important. Age is important, and it is positively correlated if the searcher is older and is negatively correlated if they are younger. Individual symptom features related to back pain and indigestion are important but have a negative influence on predicting future experiential queries, likely because (i) there are many explanations for why these symptoms appear in a query timeline, and (ii) they are positive for many negative cases (16.7% of negatives search for back pain, 7.4% search for indigestion).

## 5.4 Feature Classes

Beyond the individual features, we can also consider the accuracy of the models based on feature classes. This can be particularly important when some classes of features are easy to obtain in practice, e.g., the demographic features may be available for all searchers without the need to perform temporal modeling of query patterns. Table 5 presents the AUROC for models trained on each of the feature classes. The findings show that the *Temporal* class is particularly important, signifying the key role of temporal dynamics for this prediction task. The model is still accurate solely with access to demographics and basic features about general searching. However, performance improves considerably if we consider the specifics of the symptoms searched (*SymptomSpecific*) or the topics of the queries and results clicked (*QueryTopic*).

**Table 5:** Performance of individual feature classes (as AUROC and TPR at a FPR of 0.00001) averaged across the 10 experimental folds. The performance differences against the *Overall* performance are statistically significant using paired $t$-tests at * $p < 0.01$, ** $p < 0.001$, and *** $p < 0.0001$.

| Feature Class | AUROC | TPR (as a %) at FPR=0.00001 |
|---|---|---|
| Demographic | 0.6565*** | 0.280%*** |
| RiskFactor | 0.6988*** | 0.653%** |
| SearchHistory | 0.7202*** | 1.399%** |
| QueryURLTopic | 0.7597** | 2.052%** |
| SymptomGeneral | 0.7672** | 2.146%** |
| URLTopic | 0.7753** | 2.332%* |
| SymptomSpecific | 0.8176* | 2.800%* |
| QueryTopic | 0.8137* | 2.892%* |
| Temporal | 0.8391* | 2.985%* |
| Overall | 0.9003 | 4.851% |

## 5.5 Symptoms and Risk Factors

We also considered the impact of the presence of symptoms and risk factors on the performance of the model.

- *Symptoms*: We filtered the positive and negative cases to those where a symptom was present in query timelines.
- *Risk factors*: These are risk factors corresponding to the presence of factors such as pancreatitis, smoking, and obesity, as well as cancer syndromes such as hereditary intestinal polyposis syndrome or familial atypical multiple mole melanoma syndrome (i.e., genetic disorders that predispose individuals to develop pancreatic cancer), all of which have been shown to lead to increased likelihood of developing pancreatic cancer [18, 20, 32, 48].

Recall that our cross-validation was stratified by user. During cross validation, we learned a model on the users in the training folds and then for testing we limited to users with evidence of the specific symptoms or risk factors in their search history prior to the experiential diagnostic query. In each case the number of positives and negatives is less than the full set.[4] Table 6 presents statistics on the performance for each model where the number of positive examples was at least 10 (to help ensure that AUROC calculations were meaningful). The table also presents TPRs at different false positive rates, as well as the percentage of positive or negative cases that have the symptom or risk factor searches. Finally, the last three columns shows the estimated number of true positives (capture) and false positives (cost) that would be observed, assuming a FPR of 0.00001, and the associated capture-cost ratio. Ideal targets for rates of capture versus cost in a deployed service can be derived via a decision analysis that considers the net expected value of the early detection and the expected costs of unnecessary anxiety. Such an optimization would leverage a careful characterization of the value of early intervention and details of designs of methods for engaging people.

Table 6 shows that focusing on users who search for risk factors such as smoking, hepatitis, and obesity leads to better overall performance. There were fewer than ten users searching for each of the cancer syndromes (e.g., hereditary nonpolyposis colorectal cancer) and, hence, they were excluded from Table 6. Focusing on the percentage of posi-

[4] An alternative would be to train a separate model for symptom or risk factor. An issue with doing that is there are insufficient positive examples about each dataset with which to train a robust model.

**Table 6:** Performance of the models conditioned on a variety of symptom (S) and risk factors (RF). Values below the dashed line have a higher AUROC than *Overall*. Capture represents the number of TP cases in the cohort of positives ∪ negatives at FPR=0.00001. Cost is computed as the target FPR (.00001) multiplied by the size of the negative set in each subgroup. Since this exact FPR may not be attainable in each subgroup, cost may not be an integer.. A capture-cost ratio of $> 1.0$ means that more people would benefit from an alert than would be mistakenly alerted. Statistically significant differences with *Overall* model (DeLong's test [12]) are marked ** $p < 0.001$ and *** $p < 0.0001$ (where $\alpha$ following a Bonferroni correction is 0.002).

| Symptom or Risk Factor | TPR at FPRs ranging from 0.00001–0.1 | | | | | AUROC | # pos (%) | # neg (%) | FPR = 0.00001 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.00001 | 0.0001 | 0.001 | 0.01 | 0.1 | | | | Capture | Cost | Capture/Cost |
| Dark or tarry stool (S) | 7.692 | 7.692 | 23.077 | 38.462 | 46.154 | 0.7173*** | 13 (1.2%) | 58,597 (1.9%) | 1 | 0.5860 | 1.7066 |
| Abdominal swelling (S) | 4.167 | 8.333 | 16.667 | 20.833 | 45.833 | 0.7735*** | 24 (2.2%) | 45083 (1.5%) | 1 | 0.4508 | 2.2183 |
| Ulcers (RF) | 0.000 | 0.000 | 0.000 | 7.895 | 50.000 | 0.7894*** | 38 (3.5%) | 16,081 (0.5%) | 0 | 0.1608 | 0.0000 |
| Dark urine (S) | 0.000 | 5.556 | 16.667 | 27.778 | 50.000 | 0.8129** | 18 (1.7%) | 51,236 (1.7%) | 0 | 0.5124 | 0.0000 |
| Pancreatitis (RF) | 6.061 | 9.091 | 12.121 | 24.242 | 54.546 | 0.8220** | 33 (3.1%) | 34,184 (1.1%) | 2 | 0.3418 | 5.8514 |
| Abdominal pain (S) | 5.385 | 10.000 | 16.923 | 32.308 | 60.000 | 0.8343** | 130 (12.1%) | 311,266 (10.3%) | 7 | 3.1127 | 2.2489 |
| Enlarged gallbladder (S) | 0.885 | 2.655 | 9.735 | 25.664 | 53.982 | 0.8358** | 113 (10.5%) | 98,454 (3.3%) | 1 | 0.9845 | 1.0157 |
| Constipation (S) | 3.529 | 7.059 | 9.412 | 22.353 | 57.647 | 0.8469** | 85 (7.9%) | 317,300 (10.5%) | 3 | 3.1730 | 0.9455 |
| Smoking (RF) | 3.846 | 3.846 | 7.692 | 15.385 | 53.846 | 0.8585 | 26 (2.4%) | 27,817 (0.9%) | 1 | 0.2782 | 3.5945 |
| Blood clot (S) | 4.494 | 10.112 | 14.607 | 31.461 | 61.798 | 0.8589 | 89 (8.3%) | 351,385 (11.6%) | 4 | 3.5139 | 1.1383 |
| High blood sugar (S) | 6.135 | 8.896 | 16.564 | 31.595 | 60.429 | 0.8611 | 326 (30.4%) | 429,543 (14.2%) | 20 | 4.2954 | 4.6561 |
| Nausea or vomiting (S) | 3.200 | 8.800 | 17.600 | 30.400 | 63.200 | 0.8706 | 125 (11.7%) | 639,502 (21.1%) | 4 | 6.3950 | 0.6255 |
| Chills or fever (RF) | 3.636 | 7.273 | 20.909 | 30.909 | 65.455 | 0.8727 | 110 (10.3%) | 357,536 (11.8%) | 4 | 3.5754 | 1.1188 |
| Loose stool (S) | 4.615 | 7.692 | 18.462 | 35.385 | 72.308 | 0.8756 | 65 (6%) | 74,720 (2.5%) | 3 | 0.7472 | 4.0150 |
| Indigestion (S) | 7.547 | 12.264 | 20.755 | 38.679 | 68.868 | 0.8932 | 106 (9.9%) | 504,462 (16.7%) | 8 | 5.0446 | 1.5859 |
| Itchy skin (S) | 18.750 | 25.000 | 25.000 | 25.000 | 75.000 | 0.8982 | 16 (1.5%) | 79,448 (2.6%) | 3 | 0.7945 | 3.7760 |
| Back pain (S) | 7.801 | 14.184 | 19.858 | 34.752 | 69.504 | 0.9047 | 141 (13.2%) | 223,586 (7.4%) | 11 | 2.2359 | 4.9197 |
| Yellow skin or eyes (S) | 2.174 | 5.439 | 19.565 | 38.044 | 73.913 | 0.9217 | 92 (8.6%) | 85,805 (2.8%) | 2 | 0.8581 | 2.3307 |
| Hepatitis (RF) | 7.692 | 10.256 | 20.513 | 38.462 | 71.795 | 0.9275 | 39 (3.6%) | 25,158 (0.8%) | 3 | 0.2516 | 11.9237 |
| Alcoholism (RF) | 12.500 | 16.667 | 27.083 | 41.667 | 89.583 | 0.9494** | 48 (4.5%) | 32,333 (1.1%) | 6 | 0.3233 | 18.5586 |
| Obesity (RF) | 20.690 | 20.690 | 37.931 | 62.069 | 82.7590 | 0.9572** | 29 (2.7%) | 22,153 (0.7%) | 6 | 0.2215 | 27.0880 |
| Overall | 4.851 | 8.302 | 17.258 | 36.474 | 72.015 | 0.9003 | 1,072 (100%) | 3,025,046 (100%) | 52 | 30.2505 | 1.7190 |

tives and negatives that contain each of the symptoms or risk factors, we observe that there are some that are much more likely to occur in positives (e.g., pancreatitis and smoking are 5.9 and 3.6 times as likely, respectively). Focusing on the utility, we find that if we set the FPR to 0.00001, overall we would find 52 positives in the union of positives and negatives at the expense of 30 negatives, who would be altered mistakenly. There are some symptoms and risk factors for which the capture-cost is more favorable. For example, in the case of alcoholism or obesity, we would find 20–30 times as many TPs as FPs. There are others symptoms such as nausea or vomiting, or chills or fever, where the costs in mistakenly alerting users equal or outweigh the benefits. Presence of symptoms or risk factors could help decide whether to apply early detection models for a searcher.

## 5.6 Applying Learned Model in Practice

Up to now, our model considers experiential diagnostic users as positives and symptom-only users as negatives. This is a clean dataset for algorithm training and testing but it ignores the symptom searchers who issue non-experiential pancreatic cancer searches (gray region in Figure 1). These users may have been diagnosed or may simply be exploring. Regardless, they should be considered in practice.

We perform an additional experiment on a separate set of symptom searchers that included non-experiential pancreatic cancer searchers as negatives. We trained a model on all data described thus far and applied it to identify (i) experiential and (ii) experiential+treatment users in a new held out dataset in advance of their first experiential diagnostic query. We generated the test set from logs of a separate randomly selected subset of Bing users, over an 18-month period from August 2014 to January 2016 inclusive. There was no overlap in users with the set used for training. We identified positive cases as earlier and expanded the definition of negatives to include pancreatic cancer searchers. This resulted in 2.9 million negatives, including 48,221 non-experiential pancreatic cancer searchers, and 945 experiential searchers with preceding symptom searches. To help target the identification of cases where experiential queries are issued, we created a subset of the positives who issued treatment-related queries following $Exp_0$ (e.g., whipple procedure, 5-fu); in total, 494 users (52%) met this requirement.

**Table 7:** Average AUROC and average TPR (as %) at FPR=0.00001 for identifying experiential users and experiential+treatment users from held out dataset. Differences in AUROC and TPR between $Exp_0$ – 1 week and other weeks noted (*** $p < 0.001$, ** $p < 0.001$, and * $p < 0.01$).

| Weeks before $Exp_0$ | Experiential | | Experiential+Treatment | |
|---|---|---|---|---|
| | AUROC | TPR (%) | AUROC | TPR (%) |
| 1 week | 0.9012 | 8.677 | 0.9225 | 12.145 |
| 5 weeks | 0.8892 | 8.571 | 0.9089* | 11.943 |
| 9 weeks | 0.8754** | 8.278* | 0.8902** | 11.343* |
| 13 weeks | 0.8611** | 8.018** | 0.8795** | 10.738** |
| 17 weeks | 0.8456*** | 7.666** | 0.8683*** | 10.400*** |
| 21 weeks | 0.8330*** | 7.438*** | 0.8508*** | 9.645*** |

The symptom lookup durations for positives and negatives were similar to Section 3.5. We randomly split the test data into ten equally-sized subsets for significance testing. Table 7 reports the predictive performance at different lead times.

Table 7 shows that the performance of the model remains strong on this held-out set and is comparable to that reported in the earlier sections. The performance decreases with increased lead time as noted previously (see Table 3). Interestingly, the performance in identifying the subset of experiential diagnostic users who subsequently searched for treatments is higher than for the experiential-only set. This is promising confirmatory evidence as these users are assumed to have experienced a cancer diagnosis, per definitions of experiential queries [40].[5]

# 6. DISCUSSION AND CONCLUSIONS

We studied the potential feasibility of learning from search engine logs to predict future issuance of experiential queries about pancreatic cancer at a low error rate. The success of these methods has implications for online methods that would provide passive screening of searchers to provide early warning about potential signs of pancreatic cancer and other devastating diseases. We discovered that conditionalization on different symptoms and risk factors can enhance predictive power. We found capture-cost tradeoffs associated with

[5]For completeness, we also trained a model on all data from earlier in the paper, including the non-experiential pancreatic cancer searchers as negative cases, and tested it on this held-out set. The performance is around 5% lower than reported in Table 7, for both AUROC and TPR, across all weeks. Including the non-experiential pancreatic cancer searchers may add noise to model training.

different symptoms and risk factors in terms of the total number of truly positive cases identified versus the number of searchers who would be mistakenly alerted. We characterized model performance as we increase lead time and we found that we can attain a TPR in region of 5–30%, while controlling the FPR to 0.00001–0.01 months before a diagnostic query is observed. Looking forward, we seek to understand the costs and clinical significance of these methods, including how they might offer early warning of devastating disease onset to enhance outcomes (e.g., quality of life).

Despite the promising findings, we note several important limitations. First, we lack explicit ground truth about diagnoses per the anonymity of our logs. We rely on models of self-reporting in queries. We have found that streams of queries following the experiential queries can provide confirmatory evidence of pancreatic cancer diagnoses. Indeed, in the weeks immediately following the experiential diagnostic query, over 40% of searchers queried for treatment options, with many using sophisticated terminology (e.g., Whipple procedure, pancreaticoduodenectomy, neoadjuvant therapy) and over 20% of searchers searched for pancreatic cancer medications (e.g., gemcitabine, 5-fu). In contrast, only 0.5% and 0.02% of searchers in our negative set searched for treatments and medications, respectively, at any point in their query timeline. We need to work with diagnosed patients to understand (i) the relationship between experiential searching and diagnosis; and (ii) the model performance with the use of traditional data as ground truth about diagnoses (e.g., medical records). We also need to understand the role of factors such as race [9], family history [33], medical histories [32], diabetes [14], and other factors (e.g., smoking [18]). Some of these can be crudely estimated from geographic and census data (race), whereas others (family and medical histories) are best sought from searchers directly. To reflect anticipated performance in a natural setting, we focused on our imbalanced dataset. We re-ran the analysis with a balanced set, with highly similar results. Finally, we note that this is a retrospective analysis for model training/testing and we need to consider its representativeness for real-time screening, e.g., identifying negative cases in the retrospective study relies on symptom lookup durations. Potential additional analyses include explorations of predicting on fixed dates versus at the end of the observation period.

We are interested in several research directions. We believe it would be valuable to collect ground truth data, e.g., via targeted surveys, where responses and electronic health records could be linked (given consent) to long-term search activity. There is opportunity to develop more sophisticated time-series models and with applying our methods to other diseases. We leave to future reflection and efforts the design of methods for fielding the methods. We seek to engage with the medical community on directions for deploying the technology. In a recent sister publication, we shared these findings with practicing oncologists [39]. For real-world deployment, we need to consider whether online services would wish to provide individuals with early warnings about undiagnosed diseases given false positive rates, anxiety and associated costs of ruling out illness, privacy implications, and liability concerns. Beyond alerting searchers, a system could provide summaries of symptom searches as talking points for dialog with a medical professional, or contact a physician on the individual's behalf. One could imagine services enabling users to opt-in to such screening programs with appropriate education and caveats about false-positive rates and their associated costs. In another approach, models could be trained from anonymized data yet fielded in a private manner, e.g., as an application on a searcher's smartphone. Future work should consider the value of the search-centric analyses in the context of more traditional screening methods such as direct (active) cancer screening. Larger designs would consider how the search-based methods could be integrated with traditional screening to develop a more cost-effective screening program. Such work would require a careful consideration of the accuracies and costs of pre-screening and screening and the expected benefits of increased rates of survival associated with different policies.

# 7. REFERENCES


[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. *SIGIR*, 19–26, 2006.

[2] S. L. Ayers and J. J. Kronenfeld. Chronic illness and health-seeking information on the internet. *Health*, 11(3): 327–347, 2007.

[3] J. L. Bader and M. F. Theofanos. Searching for cancer information on the internet: Analyzing natural language search queries. *JMIR*, 5(4), 2003.

[4] P. Bennett, K. Svore, and S. Dumais. Classification -enhanced ranking. *WWW*, 111–120, 2010.

[5] E. V. Bernstam, J. R. Herskovic, and W. R. Hersh. Query log analysis in biomedicine. *Handbook of Research on Web Log Analysis*, 359–377, 2009.

[6] K. Castleton et al. A survey of internet utilization among patients with cancer. *Supportive Care in Cancer*, 19(8): 1183–1190, 2011.

[7] S. T. Chari et al. Early detection of sporadic pancreatic cancer: Summative review. *Pancreas*, 44(5): 693, 2015.

[8] R. J. Cline and K. M. Haynes. Consumer health information seeking on the internet: The state of the art. *Health Education Research*, 16(6): 671–692, 2001.

[9] S. S. Coughlin et al. Predictors of pancreatic cancer mortality among a large cohort of united states adults. *Cancer Causes & Control*, 11(10): 915–923, 2000.

[10] M. De Choudhury, S. Counts, and E. Horvitz. Predicting postpartum changes in emotion and behavior via social media. *SIGCHI*, 3267–3276, 2013.

[11] M. De Choudhury, M. R. Morris, and R. W. White. Seeking and sharing health information online: Comparing search engines and social media. *SIGCHI*, 1365–1376, 2014.

[12] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 837–845, 1988.

[13] D. Downey, S. T. Dumais, and E. Horvitz. Models of searching and browsing: Languages, studies, and application. *IJCAI*, 2740–2747, 2007.

[14] J. Everhart and D. Wright. Diabetes mellitus as a risk factor for pancreatic cancer: A meta-analysis. *JAMA*, 273(20): 1605–1609, 1995.

[15] A. Fourney, R. W. White, and E. Horvitz. Exploring time-dependent concerns about pregnancy and childbirth from search logs. *SIGCHI*, 737–746, 2015.

[16] S. Fox and M. Duggan. Health online 2013, 2013.

[17] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232, 2001.

[18] C. S. Fuchs et al. A prospective study of cigarette smoking and the risk of pancreatic cancer. *Archives of Int. Med.*, 156(19): 2255–2260, 1996.

[19] J. Ginsberg et al. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232): 1012–1014, 2009.

[20] A. M. Goldstein et al. Increased risk of pancreatic cancer in melanoma-prone kindreds with p16 ink4 mutations. *NEJM*, 333(15): 970–975, 1995.

[21] P. R. Helft. Patients with cancer, internet information, and the clinical encounter: a taxonomy of patient users. *American Society of Clinical Oncology Educational Book*, pages e89–92, 2011.

[22] R. H. Hruban et al. Progression model for pancreatic cancer. *Clin. Cancer Res.*, 6(8): 2969–2972, 2000.

[23] R. Huxley et al. Type-II diabetes and pancreatic cancer: a meta-analysis of 36 studies. *British J. of Cancer*, 92(11): 2076–2083, 2005.

[24] T. Joachims. Optimizing search engines using clickthrough data. *SIGKDD*, 133–142, 2002.

[25] R. Jones et al. Generating query substitutions. *WWW*, 387–396, 2006.

[26] J. Klapman and M. P. Malafa. Early detection of pancreatic cancer: Why, who, and how to screen. *Cancer Control*, 15(4): 280–287, 2008.

[27] T. Lau and E. Horvitz. Patterns of search: Analyzing and modeling web query refinement. *UMAP*, 1999.

[28] C. Lauckner and G. Hsieh. The presentation of health-related search results and its impact on negative emotional outcomes. *SIGCHI*, 333–342, 2013.

[29] P. Legmann et al. Pancreatic tumors: comparison of dual-phase helical CT and endoscopic sonography. *American J. Roentgenology*, 170(5): 1315–1322, 1998.

[30] D. Li et al. Pancreatic cancer. *The Lancet*, 363(9414): 1049–1057, 2004.

[31] A. B. Lowenfels and P. Maisonneuve. Epidemiology and risk factors for pancreatic cancer. *Best Practice & Research Clinical Gastro.*, 20(2):197–209, 2006.

[32] A. B. Lowenfels et al. Pancreatitis and the risk of pancreatic cancer. *NEJM*, 328(20): 1433–1437, 1993.

[33] H. T. Lynch et al. Familial pancreatic cancer: A review. *Seminars in Oncology*, 23: 251–275, 1996.

[34] K. McKeown et al. Predicting the impact of scientific concepts using full-text features. *JASIST*, 2016.

[35] H. R. Mertz et al. EUS, PET, and CT scanning for evaluation of pancreatic adenocarcinoma. *Gastrointestinal Endoscopy*, 52(3): 367–371, 2000.

[36] D. Michaud. Epidemiology of pancreatic cancer. *Minerva Chirurgica*, 59(2): 99–111, 2004.

[37] M. Müller et al. Pancreatic tumors: Evaluation with endoscopic US, CT, and MR imaging. *Radiology*, 190(3):745–751, 1994.

[38] Y. Ofran et al. Patterns of information-seeking for cancer on the internet: An analysis of real world data. *PLoS One*, 2012.

[39] J. Paparrizos, R. W. White, and E. Horvitz. Screening for pancreatic adenocarcinoma using signals from web search logs: Feasibility study and results. *JOP*, 2016.

[40] M. J. Paul, R. W. White, and E. Horvitz. Search and breast cancer: On disruptive shifts of attention over life histories of an illness. *TWEB*, 10(2): 13, 2016.

[41] M. S. Pepe et al. Phases of biomarker development for early detection of cancer. *J. Nat. Cancer Inst.*, 93(14): 1054–1061, 2001.

[42] 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. *JMIR*, 5(4), 2003.

[43] A. G. Renehan et al. Body-mass index and incidence of cancer: A systematic review and meta-analysis of prospective observational studies. *The Lancet*, 371(9612): 569–578, 2008.

[44] M. Richardson. Learning about the world from long-term query logs. *TWEB*, 2(4): 21, 2009.

[45] S. J. Rulyak et al. Cost-effectiveness of pancreatic cancer screening in familial pancreatic cancer kindreds. *Gastrointestinal Endoscopy*, 57(1): 23–29, 2003.

[46] A. Sadilek, H. Kautz, and V. Silenzio. Modeling spread of disease from social interactions. *ICWSM*, 2012.

[47] American Cancer Society. Staging pancreatic cancer. http://www.cancer.org/cancer/pancreaticcancer/detailedguide/pancreatic-cancer-pdf, 2014.

[48] G. Talamin et al. Alcohol and smoking as risk factors in chronic pancreatitis and pancreatic cancer. *Digestive Diseases and Sci.*, 44(7): 1303–1311, 1999.

[49] M. I. Trotter and D. W. Morgan. Patients' use of the internet for health related matters: A study of internet usage in 2000 & 2006. *Health Inf.*, 14(3):175–181, 2008.

[50] R. West, R. W. White, and E. Horvitz. From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. *WWW*, 1399–1410, 2013.

[51] R. White and S. Drucker. Investigating behavioral variability in web search. *WWW*, 21–30, 2007.

[52] R. W. White et al. Toward enhanced pharmacovigilance using patient-generated data on the internet. *Nature CPT*, 96(2): 239–246, 2014.

[53] R. W. White and E. Horvitz. Cyberchondria: Studies of the escalation of medical concerns in web search. *TOIS*, 27(4): 23, 2009.

[54] R. W. White and E. Horvitz. Studies of the onset and persistence of medical concerns in search logs. *SIGIR*, 265–274, 2012.

[55] R. W. White and E. Horvitz. From health search to healthcare: Explorations of intention and utilization via query logs and user surveys. *JAMIA*, 21(1): 49–55, 2014.

[56] R. W. White et al. Web-scale pharmacovigilance: Listening to signals from the crowd. *JAMIA*, 20(3): 404–408, 2013.

[57] R. W. White et al. Early identification of adverse drug reactions from search log data. *JBI*, 59: 42–48, 2016.

[58] C. J. Yeo et al. Pancreaticoduodenectomy for pancreatic adenocarcinoma: Postoperative adjuvant chemoradiation improves survival. *Annals of Surgery*, 225(5): 621, 1997.

[59] J. Yu et al. Time to progression of pancreatic ductal adenocarcinoma from low-to-high tumour stages. *Gut*, 2015.