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 insights and methodologies that could lead to advance warning about the presence of devastating diseases such as pancreatic adenocarcinoma. Pancreatic adenocarcinoma is often diagnosed too late to be treated effectively as the cancer has usually metastasized by the time of diagnosis. There are few symptoms in the early stages of the illness; specific constellations of symptoms that raise concerns about pancreatic cancer typically appear only after the disease is already at an advanced stage. We identify experiential searchers who issue credible, first-person diagnostic queries for pancreatic cancer and we learn models from prior search histories that predict the later appearance of experiential queries. We show that we can infer the likelihood of seeing the rise of experiential queries months before they appear and characterize the tradeoff between positive predictivity and false positive rate. Our findings have implications 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 diagnosis, pancreatic cancer

1. INTRODUCTION
Web search is a primary resource of information 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 [52], healthcare utilization signals [53], traces of therapeutic decision making for challenging illnesses [38], and identification of new adverse effects of medications [55, 54]. Prior studies have examined how population-level signals in social media can be used to detect the emergence of diseases [19, 10]. 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 low; in many cases, survival rates may be improved significantly via earlier detection and treatment. With many cancers, screening methods can be effective for early detection and therapy...
Pancreatic cancer is widely known as difficult to detect and yet remains less than 25% [56]. 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) [43]. 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% [56].

We take as a proxy for ground truth for a diagnosis of pancreatic cancer among searchers the detection of experiential diagnostic queries. 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 [38]. 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. They 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 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 [40], surveys [47], and analyses of large-scale search log data [5, 2]. 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 [51] 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 bins to focus on results covering serious illnesses versus more likely benign explanations. Such biases can lead to inappropriate anxiety [51, 28] 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 [49], while also having applications for tasks such as result ranking [24, 1], query suggestion [25], prediction of future search actions and interests [27, 13], and detection of real-world events and activities [42]. 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 [54], population-scale
studies of nutrition [48], 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 [37, 38]. Studies have leveraged online behavioral signals for early disease detection at the population level [19], and individually [44, 11].

Screening high-risk individuals for pancreatic adenocarcinoma 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 [34, 36] 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, itchy skin, and nausea (and combinations and temporal patterns of these and other symptoms) may also be observed in the query stream. Symptom searches can also provide useful information to support early detection by health surveillance systems before patients would typically consider seeking professional medical attention.

Active, explicit screening for early signs of pancreatic adenocarcinoma is not cost effective unless there is a reasonable probability of detecting invasive or pre-invasive disease (at least 16% according to one study [43]). 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, 41].

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 common symptoms that appear in pancreatic cancer patients (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 (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 the information need, 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 featured in predictive models, and they can also be 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 ∪ 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 [50]. The remaining users after the augmentation and filtering of our dataset total 7.4 million, from which 479,787 are pancreatic cancer searchers.

3.4 Dataset Augmentation

Age and gender are important factors associated with developing pancreatic cancer [31, 35]. 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).

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.

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 average 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].

\(^1\) A session is a query sequence with at most 30-minutes between queries [49].

![Figure 2](image-url)
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 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 ($D = 0.003; p = 0.7017$) 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.

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

<table>
<thead>
<tr>
<th>Class</th>
<th>Duration (days)</th>
<th>Total # queries</th>
<th>$N$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$ $SD$</td>
<td>$M$ $SD$</td>
<td></td>
</tr>
<tr>
<td>Positives</td>
<td>109.34 49.66</td>
<td>380.66 150.83</td>
<td>1072</td>
</tr>
<tr>
<td>Negatives</td>
<td>108.04 48.35</td>
<td>378.09 151.01</td>
<td>3,025,046</td>
</tr>
</tbody>
</table>

\(^2\)We could also have addressed this by making all of our features relative percentages, but we preferred the sampling because it gave us more flexibility in feature construction. As an additional check, we included features such as the raw count of the number of queries in the symptom lookup period, and found them 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 four 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:** Users 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** contain 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 symp-

Table 3: Performance at four-week intervals for users where features can be computed from $\text{Exp}_0$ – 1 week to $\text{Exp}_0$ – 21 weeks. Values are averaged across the ten folds of cross-validation. The significance of differences in AUROC and TPR using paired t-tests for each week versus $\text{Exp}_0$ – 1 is indicated * $p < 0.01$, ** $p < 0.001$, and *** $p < 0.0001$. Weeks denote the lead time before $\text{Exp}_0$ ($\beta$ in Figure 2) (e.g., “5 weeks” predict five weeks before the first experiential diagnostic query ($\text{Exp}_0$) using prior data).

<table>
<thead>
<tr>
<th>Weeks before $\text{Exp}_0$</th>
<th>TPR at FPRs ranging from 0.00001–0.1</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Week 1</td>
<td>2 weeks</td>
</tr>
<tr>
<td>1 week</td>
<td>0.0102</td>
<td>0.0388</td>
</tr>
<tr>
<td>5 weeks</td>
<td>0.0102</td>
<td>0.0388</td>
</tr>
<tr>
<td>9 weeks</td>
<td>0.0102</td>
<td>0.0388</td>
</tr>
<tr>
<td>13 weeks</td>
<td>0.0102</td>
<td>0.0388</td>
</tr>
<tr>
<td>17 weeks</td>
<td>0.0102</td>
<td>0.0388</td>
</tr>
<tr>
<td>21 weeks</td>
<td>0.0102</td>
<td>0.0388</td>
</tr>
</tbody>
</table>

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

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 query timelines.

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., $\text{Exp}_0$ – 1 week) in AUROC is 0.9003. Given that low error rates would be vital in practice to avoid unnecessary patient alarm, we care about 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 $\text{Exp}_0$, a searcher needs to have symptom queries extending back at least four weeks before that point (i.e., to $\text{Exp}_0$ – 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 users 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 regress in time, but even 20 or so weeks before the first experimental diagnostic query the predictive performance 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 [57]), 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% [45]).

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 (Exp0 – 1 week) to 0.231 (Exp0 – 21 weeks). All differences in AUROC for Exp0 – 5 or more weeks versus Exp0 – 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 bivariate 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 a negatively correlated if younger. Individual symptom features related to back pain and indigestion are important but have a negative influence on predicting future experiential queries, likely because there are many explanations for why these symptoms appear in a query timeline, and 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).

### 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.

- **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 (genetic disorders which predispose individuals to develop pancreatic cancer), all of which have been shown to lead to increased likelihood of developing pancreatic cancer [32, 18, 46, 20].

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 experimental diagnostic query. In each case the number of positives and negatives is less than the full set. 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/false positives that would be observed, assuming an FPR of 0.00001, and the associated benefit-cost ratio.

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
nonpolypsis colorectal cancer), meaning that they were excluded from Table 6. Focusing on the percentage of positives and negatives that contain each of the symptoms / risk factors, we observe that there are some that are much more likely to occur in positives (e.g., pancreaticitis and smoking are 6.7 and 3 times as likely, respectively). Focusing on the cost-benefit, we find that if we set the FPR to 0.00001, overall we would find 52 positives in the union of positives/negatives at the expense of 30 negatives, who would be altered mistakenly. There are some symptoms and risk factors for which the benefit-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 other 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

Our model has used 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 searches for pancreatic cancer (gray subset in Figure 1). These users may have been diagnosed or may simply be exploring the topic. Regardless, they need to be considered in practice.

We ran 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. 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 (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 conformational evidence as these users are assumed to have experienced a cancer diagnosis, per definitions of experiential query [38].

### 6. DISCUSSION AND CONCLUSIONS

We have shown that we can predict evidence of the future issuance of experiential queries about pancreatic cancer well in advance of their appearance in individuals’ query streams at low error rate. The success of these methods has implications for online methods that would provide passive treatment-related queries following \( E_{x0} \) (e.g., whipple procedure, 5-fu); in total, 494 users (52%) met this requirement. 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 (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 conformational evidence as these users are assumed to have experienced a cancer diagnosis, per definitions of experiential query [38].

| Symptom or Risk Factor | TPR at FPR ranging from 0.00001-0.1 | AUROC | \( \# \text{pos} \% \) | \( \# \text{neg} \% \) | FPR = 0.00001
|------------------------|-----------------------------------|------|-------------------|-------------------|---------------------|
| Dark or tarry stool (S) | 0.9081 124,573 114,805 6,768 29,947 | 0.7575** | 12 (1.2%) | 18,420 (1.9%) | 1 0.4508 2.2181
| Abdominal swelling (S) | 0.9147 22,186 16,105 6,081 1,000 | 0.7675** | 24 (2.2%) | 16,081 (1.6%) | 0 0.1608 0.0000
| Pancreatitis (RF) | 0.9000 0.0000 6,795 50,000 78,945 | 0.7949** | 38 (3.5%) | 80,000 (8.0%) | 0 0.5143 11.343
| Diabetes (S) | 0.9056 27,775 24,000 18,126 7,226 | 0.8126** | 55 (5.5%) | 75,000 (7.5%) | 0 0.0126 0.0126
| Ulcers (RF) | 0.9061 9,091 12,242 24,000 54,546 | 0.8220** | 33 (3.3%) | 54,546 (5.4%) | 0 0.3418 5.8507
| Abdominal pain (S) | 0.8383 10,000 16,123 32,368 60,000 | 0.8383** | 120 (12.0%) | 60,000 (6.0%) | 0 0.1127 2.2489
| Enlarged gallbladder (S) | 0.8855 2,655 25,664 53,982 0.8383** | 113 (10.5%) | 53,982 (5.3%) | 0 1.0945 1.0517
| Constipation (S) | 0.7059 4,412 22,353 57,647 0.8469** | 85 (7.5%) | 57,647 (5.7%) | 0 0.0126 0.0126
| Yellow skin or eyes (S) | 0.8466 3,862 7,198 15,385 0.8585 | 52 (5.2%) | 15,385 (1.5%) | 0 0.2782 3.5949
| Blood clot (S) | 0.8946 10,112 14,607 31,461 0.8589 | 88 (8.5%) | 31,461 (3.1%) | 0 0.3539 1.1383
| High blood sugar (S) | 0.6135 8,896 16,564 31,595 0.8611 | 326 (30.4%) | 31,595 (3.1%) | 0 0.4254 1.3584
| Nausea or vomiting (S) | 0.7200 8,800 17,600 30,400 0.8703 | 126 (11.7%) | 30,400 (3.0%) | 0 0.6350 4.6561
| Loose stool (S) | 0.6366 2,766 20,909 30,690 65,455 | 0.8727 | 110 (10.3%) | 65,455 (6.5%) | 4 0.3574 6.2255
| Chills or fever (RF) | 0.8146 7,929 18,422 35,385 72,308 | 0.8756 | 65 (6.5%) | 72,308 (7.1%) | 0 0.7422 1.1888
| Indigestion (S) | 0.7547 12,264 20,755 34,675 68,488 | 0.8932 | 108 (10.8%) | 68,488 (6.8%) | 0 5.0446 4.0150
| Itchy skin (S) | 0.1675 25,000 25,000 25,000 75,000 | 0.8982 | 16 (1.6%) | 75,000 (7.5%) | 0 0.7045 1.5858
| Leucocytes (RF) | 0.7029 7,765 17,756 38,344 73,913 | 0.8917 | 92 (9.2%) | 73,913 (7.3%) | 0 0.3851 2.3309
| Smoking (RF) | 0.2174 5,439 19,565 38,044 73,913 | 0.9217 | 92 (9.2%) | 73,913 (7.3%) | 0 0.3851 2.3309
| Hepatitis (RF) | 0.7692 10,256 20,513 38,462 71,795 | 0.9275 | 39 (3.9%) | 71,795 (7.1%) | 0 0.2516 11.9246
| Alcoholism (RF) | 12,500 16,687 27,083 41,697 89,583 | 0.9494** | 48 (4.5%) | 89,583 (8.9%) | 0 0.3233 18.5569
| Obesity (RF) | 0.2069 20,690 37,031 62,069 82,769 | 0.9572** | 29 (2.7%) | 82,769 (8.2%) | 0 0.2215 27.0444
| Overall | 0.7851 4,304 17,258 34,174 72,015 | 0.9003 | 1,072 (100%) | 72,015 (72.0%) | 52 38.2658 7.1160

** indicate \( p < 0.001 \), *** indicate \( p < 0.01 \).
screening of searchers with a view to providing early warning about potential signs of pancreatic adenocarcinoma and other devastating diseases. We discovered that conditionalization on different symptoms and risk factors can enhance predictive power. We found cost-benefit tradeoffs associated with 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 the performance of the predictive models as we moved further back in time from the appearance of experiential diagnostic queries. As an example, we found that we can attain a true positive rate in region of 6–32%, while controlling the false positive rate to 0.00001–0.01 months before the appearance of the diagnostic query. Looking forward, we seek to understand the costs and clinical significance of using these methods, including how they might be used to provide early warning of the onset of devastating diseases to enhance outcomes such as quality of life and survival rates.

Despite the promise of our findings, we note important limitations. First and foremost, we lack explicit ground truth about diagnoses per the anonymity of our logs and must rely on models of self-reporting in queries. We note that streams of queries following the experiential queries can provide confirmatory evidence of a diagnosis of pancreatic cancer. Indeed, in the weeks immediately following the experiential diagnostic query, over 40% of users queried for treatment options, with many using sophisticated terminology (e.g., whipple procedure, pancreaticoduodenectomy, neoadjuvant therapy) and over 20% of users searched for pancreatic cancer medications (e.g., gemcitabine, 5-fu). In contrast, only 0.5% and 0.02% of users 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 diagnostic searching and diagnosis, and (ii) model performance with the use of traditional data on ground truth about diagnoses (e.g., from medical records). We also need to understand the role of additional factors, including race [9], family history [33], medical histories [32] or diabetes [14] and other factors (e.g., cigarette 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 (many more negatives). We re-ran the analysis with a balanced set, with highly similar results.

There are important directions for future study. There is opportunity to develop more sophisticated models to handle the time series of search interactions for early detection of pancreatic cancer and other diseases. Moving onto real-world applications and considerations of these methods, we need to examine whether search engines or other services would wish to provide individuals with early warnings about undiagnosed diseases given prediction uncertainty, the limited information available to the classifiers, privacy implications, and liability concerns. Beyond alerting the searcher, a system could provide them with 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. A real-world system would need to appropriately communicate the predictions to searchers to convey the uncertainty, and also balance other issues such potential alarm and anxiety for searchers and liability for search providers. Future work would also need to be focus on analyses of the performance of classifiers based on search log analysis versus diagnoses using more traditional methods such as direct (active) cancer screening to understand their accuracy, their survival rates, and changes in screening costs. Comparative analyses on performance would pave the way to better understanding the opportunity to employ combinations of search-based and traditional screening methods.

7. REFERENCES


