Screening for Pancreatic Adenocarcinoma Using Signals From Web Search Logs: Feasibility Study and Results

John Paparrizos, MSc, Ryen W. White, PhD, and Eric Horvitz, MD, PhD

Columbia University, New York, NY; and Microsoft Research, Redmond, WA

Corresponding author: Ryen W. White, PhD, Microsoft Research, One Microsoft Way, Redmond, WA 98052; e-mail: ryenw@microsoft.com.

Disclosures provided by the authors are available with this article at jop.ascopubs.org.

QUESTION ASKED: Can signals mined from large-scale anonymized Web search logs about symptom queries over time be harnessed to build a valuable screening methodology for pancreatic adenocarcinoma?

SUMMARY ANSWER: Search logs can provide valuable signals to predict the later appearance of first-person queries on disease management that are strongly suggestive of a professional diagnosis of pancreatic carcinoma. Performance of the risk stratification holds many weeks in advance and improves when conditioned on the presence of specific symptoms or risk factors found in people's search histories.

WHAT WE DID: We performed a statistical analysis of the web queries of millions of anonymized searchers. We identified experiential searchers who issued a first-person diagnostic query for pancreatic cancer (eg, "I was just diagnosed with pancreatic cancer"; Fig.) and we constructed statistical models that can be applied to predict in advance the appearance of such experiential queries from signals derived from the search activity of individuals.

WHAT WE FOUND: Early detection from log data can recall 5% to 15% of the positive cases at extremely low false-positive rates (0.00001 to 0.0001). We identified specific query terms and inferred demographic factors that provide significant boosts in predicting the rise of experiential queries.

BIAS, CONFOUNDING FACTOR(S), DRAWBACKS: Results are based on retrospective analysis of search logs, where experiential queries are used as a proxy for pancreatic cancer diagnoses in the absence of direct reporting from patients. We do not directly consider false negatives associated with missed diagnoses.

REAL-LIFE IMPLICATIONS: The results highlight the promise of using Web search logs as a new direction for screening for pancreatic carcinoma. The methods suggest that low-cost, high-coverage surveillance systems can be deployed to passively observe search behavior and to provide early warning for pancreatic carcinoma, and with extension of the methodology, for other challenging cancers. Surveillance systems could also provide for automated capture and summarization of data and landmarks over time so as to provide patients with talking points in their discussion with medical professionals. Real-world deployment of the methods would need to carefully convey the uncertainties associated with detection outcomes based on consideration of the evidential findings and prevalence rates, while also balancing such issues as searcher anxiety and cost of potentially unnecessary consultation and screening. JOP

See the figure on the following page.

DOI: 10.1200/JOP.2015.010504; published online ahead of print at jop.ascopubs.org on June 7, 2016.

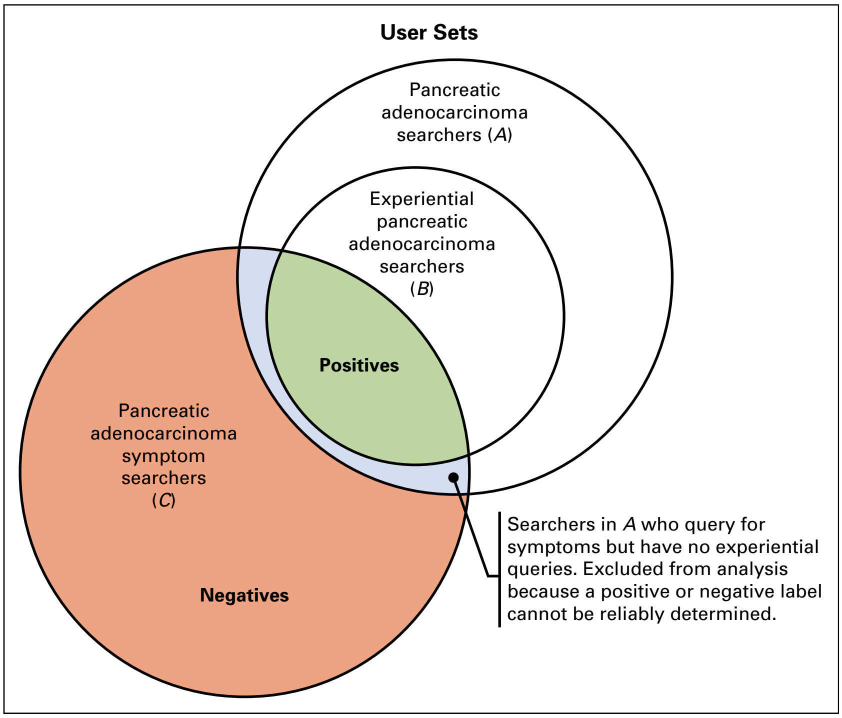
User Sets
Pancreatic adenocarcinoma searchers (*A*)
Experiential pancreatic adenocarcinoma searchers (*B*)
Positives
Pancreatic adenocarcinoma symptom searchers (*C*)
Negatives
Searchers in *A* who query for symptoms but have no experiential queries. Excluded from analysis because a positive or negative label cannot be reliably determined.

FIG. Venn diagram depicting the sets of searchers used in the search log analysis: pancreatic adenocarcinoma searchers (A), pancreatic adenocarcinoma searchers with experiential diagnostic queries (B), and those who searched for pancreatic adenocarcinoma symptoms (C). $|A \cup C|$ (ie, the total number of searchers in our original, prefiltered data set) was 9.2 million. Positives are sourced from $B \cap C$ and negatives are sourced from $C \setminus A$. Relative set sizes in the diagram are not to scale.

Screening for Pancreatic Adenocarcinoma Using Signals From Web Search Logs: Feasibility Study and Results

John Paparrizos, MSc, Ryen W. White, PhD, and Eric Horvitz, MD, PhD

Columbia University, New York, NY; and Microsoft Research, Redmond, WA

Abstract

Introduction
People's online activities can yield clues about their emerging health conditions. We performed an intensive study to explore the feasibility of using anonymized Web query logs to screen for the emergence of pancreatic adenocarcinoma. The methods used statistical analyses of large-scale anonymized search logs considering the symptom queries from millions of people, with the potential application of warning individual searchers about the value of seeking attention from health care professionals.

Methods
We identified searchers in logs of online search activity who issued special queries that are suggestive of a recent diagnosis of pancreatic adenocarcinoma. We then went back many months before these landmark queries were made, to examine patterns of symptoms, which were expressed as searches about concerning symptoms. We built statistical classifiers that predicted the future appearance of the landmark queries based on patterns of signals seen in search logs.

Results
We found that signals about patterns of queries in search logs can predict the future appearance of queries that are highly suggestive of a diagnosis of pancreatic adenocarcinoma. We showed specifically that we can identify 5% to 15% of cases, while preserving extremely low false-positive rates (0.00001 to 0.0001).

Conclusion
Signals in search logs show the possibilities of predicting a forthcoming diagnosis of pancreatic adenocarcinoma from combinations of subtle temporal signals revealed in the queries of searchers.

ASSOCIATED CONTENT

See accompanying editorial on page 699

Appendix DOI: 10.1200/JOP.2015.010504

DOI: 10.1200/JOP.2015.010504; published online ahead of print at jop.ascopubs.org on June 7, 2016.

INTRODUCTION

Pancreatic adenocarcinoma poses a difficult and resistant challenge in oncology. It is the fourth leading cause of cancer death in the United States and is the sixth leading cause of cancer death in Europe.[1] The illness is frequently diagnosed too late to be treated effectively[2,3] and can progress from stage I to stage IV in just over 1 year.[4] Approximately 75% of patients with pancreatic adenocarcinoma who are not candidates for surgery will die within 1 year of diagnosis, and only 4% will survive for 5 years postdiagnosis.[5]

Early signs and symptoms of pancreatic adenocarcinoma are subtle and often present as nonspecific symptoms that appear and evolve over time. The symptoms often do not become salient until the disease has metastasized. We studied a nontraditional, yet promising direction for the early detection of pancreatic adenocarcinoma. The approach centers on the analysis of signals from Web search logs. Specifically, we examined the feasibility of detecting "fingerprints" of the early rise of pancreatic adenocarcinoma via population-scale statistical analyses of the activity logs of millions of people performing searches on sets of relevant symptoms.

People frequently turn to Web searches to locate health-related information.[6] For example, searchers concerned about the appearance of new symptoms often input terms to search engines describing their observations and retrieve results on related medical conditions. Web searching is common among patients with cancer,[7-9] and there are strong similarities between temporal patterns in logs and behaviors observed in practice.[10,11] Analyses of logged symptom- and illness-related searches over time yields insights about medical concerns and anxieties,[12,13] and can provide evidence of health care utilization.[14] More generally, search logs enable search providers and researchers to better understand search behavior,[15] to predict future actions and interests,[16-18] to improve search engines,[19,20] and to understand in-world activities.[21]

Screening for pancreatic adenocarcinoma aims to detect the disease at a preinvasive or early invasive stage when it is still curable by surgical intervention and chemotherapy. Screening high-risk individuals for pancreatic adenocarcinoma can detect precancerous or cancerous changes in the pancreas when surgical intervention will have an increased chance of cure.[22] Risk level can be determined by factors such as race,[23] family history,[24,25] and a history of pancreatitis.[26] Imaging studies via methods such as endoscopic ultrasound, computed tomography scans, and magnetic resonance imaging[27,28] are useful to diagnose pancreatic adenocarcinoma once the tumor is large enough to cause symptoms that prompt people to seek medical attention; however, at this point, the disease is more likely to be advanced and unresectable.[29] Earlier diagnosis of pancreatic adenocarcinoma leads to earlier-stage disease[30,31] and improved chance of survival.[32,33] Although patients who are diagnosed early enough to undergo a curative resection have a higher 5-year survival rate, that survival rate is still $< 25\%$.[32]

Surveillance and screening programs for pancreatic adenocarcinoma face the challenges of engagement and coverage, especially for detecting and addressing subtle, yet important symptoms. We believe that search logs can serve as a new kind of large-scale, widely distributed sensor for capturing concerning temporal patterns of the onset and persistence of queries about symptoms. The sequences of terms that searchers input to search engines over time can capture symptoms as the illness progresses from its early stages to increasingly salient and frank symptoms.

Patterns of onset and persistence of symptoms for pancreatic adenocarcinoma include back pain, abdominal discomfort, unexplained loss of weight and appetite, light-colored stools, generalized pruritus, darkening urine, and yellowing sclera and skin. From the perspective of traditional screening, there are few salient symptoms in early stages of the disease, and the symptoms are not sufficiently specific to raise a suspicion of pancreatic adenocarcinoma. Symptoms may not even concern patients enough to schedule an appointment with their physician.

We present a feasibility study of the early identification of pancreatic adenocarcinoma based on symptom-centric search queries over time, and the temporal relationships and patterns among queries from multiple sessions over several months. Our experiments center on the early prediction of the future appearance in search logs of special queries that we term experiential diagnostic queries. Experiential diagnostic queries are terms inputted into search engines that provide evidence of searchers having recently been professionally diagnosed. These are distinct from exploratory queries, including searches on symptoms or diseases, which appear to be less intensive, more casual searches for information.[11] Experiential queries for pancreatic adenocarcinoma are identified via consideration of the query structure and patterns of information gathering over many searchers in search logs. We specifically sought evidence of credible, first-person assertions such as the query, "I was just diagnosed with pancreatic adenocarcinoma," which, when associated with prior queries about symptoms, identifies searchers who we label as positive for pancreatic adenocarcinoma. Searchers who inquire about one or more related symptoms of interest, but show no evidence over time of searches for pancreatic adenocarcinoma, constitute the negatives.

METHODS

Search services track characteristics of people's searching and clicking activities to capture intentions, improve their responses, and personalize content. Searching activities provide streams of data to construct a statistical model that can be used to risk-stratify searchers for screening. Every interaction corresponds to a log entry containing the query, the results selected, and a timestamp. A unique, anonymized identifier linked to the Web browser is also included, enabling the extraction of search log histories for up to 18 months. The anonymous identifier is tied to a single machine. On shared machines, it may represent the search activity of multiple searchers. The identifier does not enable the consolidation of activity from a single searcher across multiple machines. We used proprietary logs from Bing.com for searchers in the English-speaking United States locale, from October 2013 to May 2015 (inclusive).

Symptoms and Risk Factors

We reviewed the signs, symptoms, and risk factors associated with pancreatic adenocarcinoma. We developed a symptom set covering the following concerns: yellowing sclera or skin, blood clot, light stool, loose stool, enlarged gall bladder, dark urine, floating stool, greasy stool, dark or tarry stool, high blood sugar, sudden weight loss, taste changes, smelly stool, itchy skin, nausea or vomiting, indigestion, abdominal swelling or pressure, abdominal pain, constipation, and loss of appetite. Synonyms for each symptom were identified (eg, symptom: yellowing sclera or skin, synonym: jaundice; symptom: abdominal pain, synonyms: belly pain, stomach ache). We also identified risk factors (eg, pancreatitis, alcoholism) and their associated synonyms (see Lowenfels and Maisonneuve[34]), describing attributes or characteristics that may increase the likelihood of developing pancreatic adenocarcinoma. The symptoms and the risk factors were mapped to terms in search queries.

Extracting Pancreatic Adenocarcinoma Searchers and Symptom Searchers

To identify positive and negative cases in generating a learned model, we built a data set of searchers from two groups (Fig 1A). Pancreatic adenocarcinoma searchers (A) includes all searchers who inputted one or more queries matching the expression [(pancreas OR pancreatic) AND cancer]. We considered searchers with a diagnosis of pancreatic adenocarcinoma (B) as the subset of searchers (A) who issued one or more experiential diagnostic queries. Symptom searchers (C) includes all searchers with one or more queries related to pancreatic adenocarcinoma symptoms or synonyms (see Symptoms and Risk Factors).

The full search histories of 9.2 million searchers comprise the union of (A) and (C) in Figure 1A. We used a statistical topic classifier developed for use by the Bing search service to identify all health-related queries. We also applied statistical classifiers developed by Bing to make inferences about searchers' ages and gender. Using these statistical models as filters, we identified searchers for whom $>$ 20% of their queries were health related. We excluded those searchers, given the high likelihood that they were health care professionals.[35] A total of 7.4 million searchers remained, among whom 479,787 were pancreatic adenocarcinoma searchers. As additional features for statistical analysis, we used a classifier that provides distributions of topics for queries and clicked results.[36] We also considered the dominant geolocation for each searcher using a table that links their Internet provider address to locations.

Positive and Negative Cases

We created query timelines for those labeled as experiential diagnostic searchers and exploratory symptom searchers, and drew sets of observations from these timelines to construct a risk-stratification model. Figure 1B summarizes the strategies for identifying positives and negatives. Query timelines are aligned across searchers based on the point when people issued the first experiential diagnostic query. To ensure sufficient data about each searcher, we removed from the study those with fewer than five search sessions (comprising a sequence of search actions with no more than 30 minutes between actions)[15,17] spanning five different days. This reduced the population to 6.4 million searchers, with a mean total duration (time between first and last queries) of 210.32 days (standard deviation of 182.93 days and interquartile range of 120 days).

Positive cases

To identify experiential pancreatic adenocarcinoma searchers, we defined first-person diagnostic queries for pancreatic adenocarcinoma (Exp_0) based on an exploration of logs. Queries admitted as experiential diagnostic queries included such phrases as "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 pancreatic adenocarcinoma searchers, 3,203 matched the diagnostic query patterns. Experiential searchers must have searched for at least one symptom before Exp_0. This generated 1,072 query timelines of experiential searchers containing periods of symptom lookup followed by the diagnostic query (33.5% of all experiential diagnostic searchers). The symptom lookup period starts when the first symptom is detected in our symptom set (mean duration [α] = 109.34 days, standard deviation = 49.66 days). For positives, the symptom lookup period terminates at least 1 week before diagnosis (β = 1 week) to reduce the likelihood of overlap between them (which could

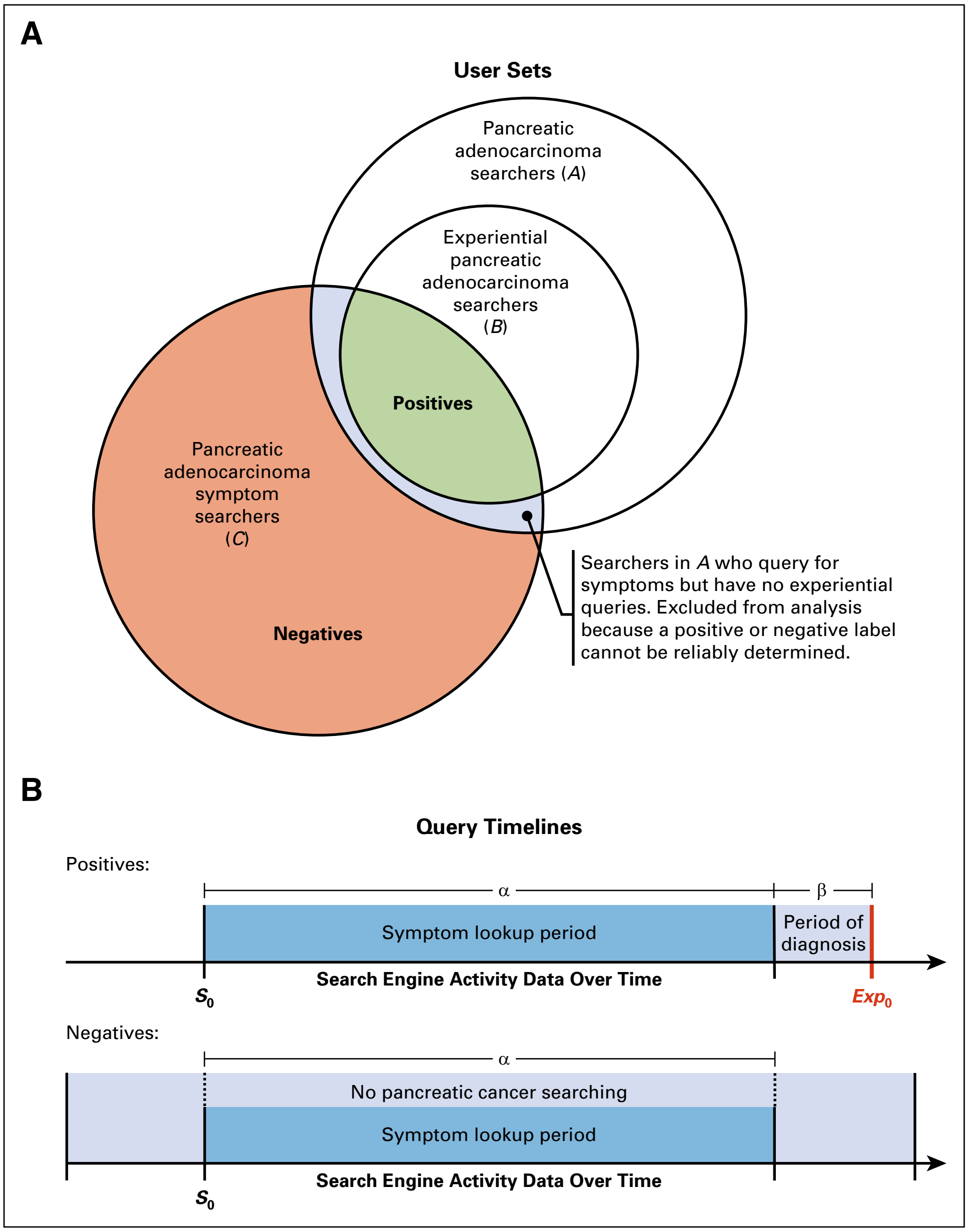
A
User Sets
Pancreatic adenocarcinoma searchers (A)
Experiential pancreatic adenocarcinoma searchers (B)
Positives
Pancreatic adenocarcinoma symptom searchers (C)
Negatives
Searchers in A who query for symptoms but have no experiential queries. Excluded from analysis because a positive or negative label cannot be reliably determined.
B
Query Timelines
Positives:
α
β
Symptom lookup period
Period of diagnosis
Search Engine Activity Data Over Time
S0
Exp0
Negatives:
α
No pancreatic cancer searching
Symptom lookup period
Search Engine Activity Data Over Time
S0

FIG 1. (A) Venn diagram depicting the sets of searchers used in the search log analysis: pancreatic adenocarcinoma searchers (*A*), pancreatic adenocarcinoma searchers with experiential diagnostic queries (*B*), and those who searched for pancreatic adenocarcinoma symptoms (*C*). $|A \cup C|$ (ie, the total number of searchers in our original, prefiltered data set) was 9.2 million. Positives are sourced from $B \cap C$ and negatives are sourced from $C \setminus A$. Relative set sizes in the diagram are not to scale. (B) Schematic illustrating 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. α is the duration of the symptom lookup period, which is meant to be approximately equal in the aggregate for the positives and negatives. β is the duration of the period of diagnosis, set to 1 week in the current study.

add noise to model training and testing), while allowing us to understand predictive performance with minimal lead times.

Negative cases

To generate a set of searchers we considered negative for pancreatic adenocarcinoma, we sampled from those who searched for pancreatic adenocarcinoma symptoms but who did not search for pancreatic adenocarcinoma directly anywhere in their timeline. We reduced the number of negatives via a sampling procedure to include only those with symptom lookup durations within three standard deviations of the mean of the positives (n = 3,025,046). The resultant positive and negative distributions are statistically indistinguishable using two-sample Kolmogorov-Smirnov tests for temporal duration (D = 0.005, P = .7017) and number of queries (D = 0.003, P = .7681), even though the latter was not a filtering criterion.

Early Detection

We framed early detection as a binary classification challenge using a statistical classifier. We trained the classifier on features from query timelines of experiential pancreatic adenocarcinoma searchers and symptom-only searchers. Given concerns about false positives and the rarity of pancreatic adenocarcinoma, we focused on maintaining low false-positive rates (FPRs; ie, one wrong prediction in 100,000 correctly identified cases), while retaining a high imbalance ratio of positives and negatives (ie, 1,000 positives *v* millions of negatives).

The set of observations or features extracted from the symptom lookup period are grouped into five categories as follows: (1) searcher demographic information, including age/sex predictions and dominant location (Demographics); (2) session characteristics, query classes, and URL classes, including activity characteristics and the topics of queries issued and resources accessed (Search Characteristics); (3) characteristics about symptoms searched, including generic symptom searching (eg, number of distinct symptoms; Symptom General) and features for each symptom (Symptom Specific); (4) features that capture the temporal dynamics of the features (eg, increasing/decreasing over time, rate of change; Temporal), and (5) risk factors, including their presence in queries (Risk Factors).

The learned statistical model is based on the gradient boosted trees[37] method. Regularization methods were used to minimize the risk of overfitting. See Paparrizos et al[38] for details on the construction of the classifier. We used the statistical classifier to study our ability to perform early identification of searchers who would later make experiential diagnostic queries for pancreatic adenocarcinoma. To characterize the predictive power, we used the area under the receiver operator curve (AUROC) and the recall (true-positive rate [TPR]) at low FPRs as evaluation metrics. Model generalizability was assessed using 10-fold cross validation, stratified by searcher.

RESULTS

Performance of the statistical classifier using data up to the period of diagnosis (ie, Exp_0 − 1 week) was strong (AUROC = 0.9003). Because low error rates are important when applying our model, the TPR (ie, fraction of positives recalled) at low FPRs (ie, 0.0001 or 0.00001) is also of interest. Focusing on FPRs in the range 0.00001 to 0.01, the model recalls 5% to 30% of the positives, depending on the FPR.

Performance by Week

Prediction performance can change as we increase the lead time between prediction and diagnostic query. We selected

Table 1. Performance at Early Prediction Task at 4-Week Intervals for the Set of Searchers for Whom Features Can Be Computed From Exp_0 − 1 Week to Exp_0 − 21 Weeks

No. of Weeks	TPR at FPRs Ranging From 0.00001 to 0.1					
Before Exp_0*	0.00001	0.0001	0.001	0.01	0.1	AUROC
1	7.122	10.386	20.772	36.202	71.810	0.9112
5	7.122	10.979	20.178	34.421	70.620	0.9047
9	7.122	10.683	18.991†	33.234†	70.023	0.8854†
13	7.122	9.792	17.804†	32.937†	67.359†	0.8700†
17	6.825	9.199†	17.209†	32.640‡	64.688‡	0.8539‡
21	6.528†	9.199†	16.319‡	32.345‡	61.424§	0.8315‡

NOTE. Values are averaged across the 10 folds of the cross-validation. Weeks denotes the week before first experiential diagnostic query when the prediction is made (eg, "5 weeks" means to train the model using data up to 5 weeks before the first experiential diagnostic query [Exp_0]).

Abbreviations: AUROC, area under the receiver operating characteristic curve; Exp_0, first-person diagnostic queries for pancreatic adenocarcinoma; FPR, false-positive rate; TPR, true-positive rate.

*β in Figure 1B.

†$P < .01$, ‡$P < .001$, and §$P < .0001$.

Table 2. Top 10 Features, Ranked in Descending Order by Evidential Weight

Observation Type	Weight	Direction	Class
No. of distinct symptoms searched	1.0000	Positive	Symptom general
Fraction of search queries that are health related	0.8253	Positive	Query topic
No. of distinct symptom synonyms searched	0.6899	Positive	Symptom general
Probability that searcher's age is 50-85 years	0.6889	Positive	Demographic
Searcher has searched for back pain	0.6622	Negative	Symptom specific
Searcher has searched for indigestion	0.6432	Negative	Symptom specific
Searcher has searched for indigestion, then abdominal pain	0.6349	Positive	Temporal
Gradient of best-fit line for no. of distinct symptoms searched	0.6154	Positive	Temporal
Searcher has searched for back pain, then yellowing sclera or skin	0.6004	Positive	Temporal
Probability that searcher's age is < 18 years	0.5869	Negative	Demographic

NOTE. Weights are relative to the top-weighted feature, "No. of distinct symptoms searched," which was assigned a weight of 1.0000. Direction of positive or negative means that the feature correlates positively or negatively with ground truth.

337 positives and 945,394 negatives who were still observed in the logs many weeks before Exp_0, and reported results for β = 1 to 21 weeks. Because feature generation requires 4 weeks of data, for inclusion at $Exp_0 - 21$ weeks, a searcher needs to be observed at $Exp_0 - 25$ weeks.

We trained a model for the filtered set of searchers as for all searchers. Table 1 reports the TPR at different FPRs for this same set of searchers at different 4-week increments, as well as the AUROC. Performance dropped consistently with increased lead time, but even at 21 weeks before Exp_0, the predictive performance was still strong (AUROC = 0.8315, TPR [at FPR = 0.00001] = 6.528%).

Contributions by Observation Type

Table 2 shows the observation types (features) with the highest evidential weight. Direction is based on correlations between the feature and training data labels. The number of distinct pancreatic adenocarcinoma symptoms searched is most important, representing a high level of concern. Also important are temporal features, including sequence ordering of symptom pairs, inferred age, and searches for back pain and indigestion (which are common ailments and have many explanations).

Observations also varied in predictive power at FPR = 0.00001, for example, temporal dynamics (AUROC = 0.8391, TPR = 0.2985%), specific symptoms (AUROC = 0.8176, TPR = 2.800%), and demographic information (AUROC = 0.6565, TPR = 0.2800%), differing significantly from the full model (at $P < .01$ using paired t tests).

Symptoms and Risk Factors

The presence of specific symptoms and risk factors in searchers' query timelines could affect early detection performance. Risk factors include pancreatitis, smoking, and obesity, as well as cancer syndromes such as hereditary intestinal polyposis syndrome or familial atypical multiple mole melanoma syndrome, which can all increase the likelihood of developing pancreatic adenocarcinoma.[26,39-43]

We applied cross-validation. For training, we learned a model on searchers in the nine folds allocated to training. For testing, we iterated through symptoms and risk factors and isolated searchers in the test fold who searched for those symptoms or risk factors at $Exp_0 - 1$ week or earlier. In each case, the number of positives and negatives is less than the full set. Appendix Table A1 (online only) presents statistics on the performance for each model with ≥ 10 positives (to help ensure that AUROC calculations were meaningful). TPRs at different FPRs are shown, as are the percentage of positives or negatives with symptom or risk factor searches. The last three columns present the estimated number of true positives (TPs) or false positives (FPs) that would be observed at FPR = 0.00001, and capture cost estimates in terms of numbers of searchers correctly and falsely alerted. 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 and rule-out. Such an optimization would leverage a careful characterization of the value of early intervention and details of designs of methods for engaging people.

Appendix Table A1 shows that considering only searchers seeking information related to risk factors such as smoking, hepatitis, and obesity leads to better overall performance. Fewer than 10 searchers searched for each cancer syndrome (eg, hereditary nonpolyposis colorectal cancer), and these cases were excluded from Appendix Table A1. We found terms for symptoms and risk factors that are more likely to occur in positives (eg, pancreatitis is six times more likely, smoking is four times more likely). If we fixed FPR = 0.00001, we would correctly detect 52 searchers (TPs) but would mistakenly alert 30 searchers (FPs; capture cost ratio = 1.72). Appendix Table A1 also shows that conditionalizing on specific symptoms/risk factors markedly improves the capture cost ratio. For example, for alcoholism or obesity, we found 20 to 30 times more TPs than FPs.

DISCUSSION

Web search logs may offer a useful source of signals for pancreatic adenocarcinoma screening, with significant lead time (eg, 5 months before the diagnostic query, TPR is 6% to 32% at extremely low FPRs). Because pancreatic adenocarcinoma may progress from stage I to stage IV in just over 1 year,[4] this screening capability could increase 5-year survival. Model performance on some symptoms and risk factors is even stronger. There are others (such as nausea, vomiting, chills, or fever) where the costs in mistakenly recommending that searchers seek medical attention could outweigh the benefits.

For completeness, we re-ran the analysis with an equally-balanced set of positives and negatives, and also learned a model using all positives/negatives and applied it to separate set of Bing logs where nonexperiential pancreatic adenocarcinoma searchers (gray region in Fig 1A) were included to mimic a realistic application scenario. Both studies yielded results similar to those reported herein. A final experiment where nonexperiential searchers were included as negatives for training (and testing occurred on the same separate set of logs) revealed a drop in AUROC and TPR. Including the nonexperiential pancreatic adenocarcinoma searchers may add noise to model training.[38]

We acknowledge that this study has several limitations. Per log anonymity, we lack explicit ground truth about diagnoses and rely on implicit self-reporting in queries. We note that streams of queries following the experiential queries provide confirmatory evidence of a pancreatic adenocarcinoma diagnosis. In the weeks immediately following Exp_0, $>$ 40% of searchers queried for treatment options, with many using sophisticated terminology (eg, Whipple procedure, pancreaticoduodenectomy, neoadjuvant therapy) and $>$ 20% searched for related medications (eg, gemcitabine, fluorouracil). In contrast, only 0.5% and 0.02% of our negative cases searched for treatments and medications, respectively, at any point in their query timeline. The impact of additional risk factors such as race,[23] family history,[24,25] and medical history[26,44] needs to be understood. Oncologists and patients also need to be directly involved in future studies.

To understand how particular symptoms or risk factors influence model performance, we excluded searchers who lacked supporting evidence for each symptom or risk factor in their search histories. An alternative is to train a separate model for each symptom or risk factor. However, there were insufficient positive examples in each data set with which to train a robust model. In addition, training a generic model and conditioning its application on the presence of symptoms and risk factors is more similar to how the model would be applied in practice.

Our approach leverages low-cost passive observation rather than active screening. This could be generalized to other diseases where noticeable symptoms appear and evolve over periods of time before diagnoses are made. Active screening is not cost effective unless there is a reasonable probability of detecting invasive or preinvasive disease (eg, at least 16%[45]). Search log–based (retrospective) methodologies support the characterization of individuals' longitudinal behaviors at a scale that is infeasible in other studies, which are typically much smaller, for example, Huxley et al[46] and Renehan et al.[47] Comparisons against baselines, where suspicions about the presence of pancreatic adenocarcinoma are raised via direct screening, are needed to determine changes in screening costs associated with our method. Clinical trials are necessary to understand whether our learned model has practical utility, including in combination with other screening methods.

Alerting people about the potential value of seeking medical care can be challenging. Surveillance systems need to convey the uncertainties associated with detection outcomes while balancing other issues such as alarm and anxiety for searchers and liability for search providers. Systems could summarize historic symptom search activity as talking points for discussion with medical professionals or alert physicians separately from patients. JOP

Authors' Disclosures of Potential Conflicts of Interest
Disclosures provided by the authors are available with this article at jop.ascopubs.org.

Author Contributions
Conception and design: All authors
Collection and assembly of data: All authors
Data analysis and interpretation: All authors
Manuscript writing: All authors
Final approval of manuscript: All authors

Corresponding author: Ryen W. White, PhD, Microsoft Research, One Microsoft Way, Redmond, WA 98052; e-mail: ryenw@microsoft.com.

References

1. Michaud DS: Epidemiology of pancreatic cancer. Minerva Chir 59:99-111, 2004

2. Hruban RH, Goggins M, Parsons J, et al: Progression model for pancreatic cancer. Clin Cancer Res 6:2969-2972, 2000

3. Li D, Xie K, Wolff R, et al: Pancreatic cancer. Lancet 363:1049-1057, 2004

4. Yu J, Blackford AL, Dal Molin M, et al: Time to progression of pancreatic ductal adenocarcinoma from low-to-high tumour stages. Gut 64:1783-1789, 2015

5. Bilimoria KY, Bentrem DJ, Ko CY, et al: Validation of the 6th edition AJCC Pancreatic Cancer Staging System: Report from the National Cancer Database. Cancer 110:738-744, 2007

6. Fox S, Duggan M: Health Online 2013. Washington, DC, Pew Research Center's Internet & American Life Project, 2013. www.pewinternet.org/2013/01/15/health-online-2013/

7. Bader JL, Theofanos MF: Searching for cancer information on the internet: Analyzing natural language search queries. J Med Internet Res 5:e31, 2003

8. Castleton K, Fong T, Wang-Gillam A, et al: A survey of Internet utilization among patients with cancer. Support Care Cancer 19:1183-1190, 2011

9. Helft PR: Patients with cancer, internet information, and the clinical encounter: A taxonomy of patient users. Am Soc Clin Oncol Educ Book 35:e89-e92, 2012

10. Ofran Y, Paltiel O, Pelleg D, et al: Patterns of information-seeking for cancer on the internet: An analysis of real world data. PLoS One 7:e45921, 2012

11. Paul MJ, White RW, Horvitz E: Search and breast cancer: On episodic shifts of attention over life histories of an illness. ACM Trans Web 10:2, 2016

12. White RW, Horvitz E: Cyberchondria: Studies of the escalation of medical concerns in web search. ACM Trans Inf Syst 27:23, 2009

13. Lauckner C, Hsieh G: The presentation of health-related search results and its impact on negative emotional outcomes, in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY, ACM, 2013, pp 333-342

14. White RW, Horvitz E: From health search to healthcare: Explorations of intention and utilization via query logs and user surveys. J Am Med Inform Assoc 21:49-55, 2014

15. White RW, Drucker SM: Investigating behavioral variability in web search, in Proceedings of the World Wide Web Conference. New York, NY, ACM, 2007, pp 21-30

16. Lau T, Horvitz E: Patterns of search: analyzing and modeling web query refinement, in Proceedings of the User Modeling Conference. Vienna, Austria, Spring, 1999, pp 119-128

17. Downey D, Dumais ST, Horvitz E: Models of searching and browsing: Languages, studies, and applications, in Proceedings of the International Joint Conference on Artificial Intelligence. San Francisco, CA, Morgan Kaufmann, 2007, pp 2740-2747

18. Dupret G, Piwowarski B: A user browsing model to predict search engine click data from past observations, in Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, ACM, 2008, pp 331-338

19. Joachims T: Optimizing search engines using clickthrough data, in Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, NY, ACM, 2002, pp 133-142

20. Tan B, Shen X, Zhai C: Mining long-term search history to improve search accuracy, in Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, NY, ACM, 2006, pp 718-723

21. Richardson M: Learning about the world from long-term query logs. ACM Trans Web 2:21, 2009

22. Klapman J, Malafa MP: Early detection of pancreatic cancer: Why, who, and how to screen. Cancer Contr 15:280-287, 2008

23. Coughlin SS, Calle EE, Patel AV, et al: Predictors of pancreatic cancer mortality among a large cohort of United States adults. Cancer Causes Control 11:915-923, 2000

24. Brand RE, Lynch HT: Hereditary pancreatic adenocarcinoma. A clinical perspective. Med Clin North Am 84:665-675, 2000

25. Lynch HT, Smyrk T, Kern SE, et al: Familial pancreatic cancer: A review. Semin Oncol 23:251-275, 1996

26. Lowenfels AB, Maisonneuve P, Cavallini G, et al: Pancreatitis and the risk of pancreatic cancer. N Engl J Med 328:1433-1437, 1993

27. Mertz HR, Sechopoulos P, Delbeke D, et al: EUS, PET, and CT scanning for evaluation of pancreatic adenocarcinoma. Gastrointest Endosc 52:367-371, 2000

28. Müller MF, Meyenberger C, Bertschinger P, et al: Pancreatic tumors: Evaluation with endoscopic US, CT, and MR imaging. Radiology 190:745-751, 1994

29. Legmann P, Vignaux O, Dousset B, et al: Pancreatic tumors: Comparison of dual-phase helical CT and endoscopic sonography. AJR Am J Roentgenol 170:1315-1322, 1998

30. Chari ST, Kelly K, Hollingsworth MA, et al: Early detection of sporadic pancreatic cancer: Summative review. Pancreas 44:693-712, 2015

31. Melo SA, Luecke LB, Kahlert C, et al: Glypican-1 identifies cancer exosomes and detects early pancreatic cancer. Nature 523:177-182, 2015

32. Yeo CJ, Abrams RA, Grochow LB, et al: Pancreaticoduodenectomy for pancreatic adenocarcinoma: Postoperative adjuvant chemoradiation improves survival. A prospective, single-institution experience. Ann Surg 225:621-633, discussion 633-636, 1997

33. Mayo SC, Nathan H, Cameron JL, et al: Conditional survival in patients with pancreatic ductal adenocarcinoma resected with curative intent. Cancer 118: 2674-2681, 2012

34. Lowenfels AB, Maisonneuve P: Epidemiology and risk factors for pancreatic cancer. Best Pract Res Clin Gastroenterol 20:197-209, 2006

35. White RW, Harpaz R, Shah NH, et al: Toward enhanced pharmacovigilance using patient-generated data on the internet. Clin Pharmacol Ther 96:239-246, 2014

36. Bennett PN, Svore K, Dumais ST: Classification-enhanced ranking, in Proceedings of the World Wide Web Conference. New York, NY, ACM, 2010, pp 111-120

37. Friedman JH: Greedy function approximation: A gradient boosting machine. Ann Stat 29:1189-1232, 2001

38. Paparrizos J, White RW, Horvitz E: Detecting devastating diseases in search logs, in Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (in press)

39. Fuchs CS, Colditz GA, Stampfer MJ, et al: A prospective study of cigarette smoking and the risk of pancreatic cancer. Arch Intern Med 156:2255-2260, 1996

40. Talamini G, Bassi C, Falconi M, et al: Alcohol and smoking as risk factors in chronic pancreatitis and pancreatic cancer. Dig Dis Sci 44:1303-1311, 1999

41. Goldstein AM, Fraser MC, Struewing JP, et al: Increased risk of pancreatic cancer in melanoma-prone kindreds with p16INK4 mutations. N Engl J Med 333:970-974, 1995

42. Gold EB, Goldin SB: Epidemiology of and risk factors for pancreatic cancer. Surg Oncol Clin N Am 7:67-91, 1998

43. Giardiello FM, Brensinger JD, Tersmette AC, et al: Very high risk of cancer in familial Peutz-Jeghers syndrome. Gastroenterology 119:1447-1453, 2000

44. Everhart J, Wright D: Diabetes mellitus as a risk factor for pancreatic cancer. A meta-analysis. JAMA 273:1605-1609, 1995

45. Rulyak SJ, Kimmey MB, Veenstra DL, et al: Cost-effectiveness of pancreatic cancer screening in familial pancreatic cancer kindreds. Gastrointest Endosc 57: 23-29, 2003

46. Huxley R, Ansary-Moghaddam A, Berrington de González A, et al: Type-II diabetes and pancreatic cancer: A meta-analysis of 36 studies. Br J Cancer 92: 2076-2083, 2005

47. Renehan AG, Tyson M, Egger M, et al: Body-mass index and incidence of cancer: A systematic review and meta-analysis of prospective observational studies. Lancet 371:569-578, 2008

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Screening for Pancreatic Adenocarcinoma Using Signals From Web Search Logs: Feasibility Study and Results

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or jop.ascopubs.org/site/misc/ifc.xhtml.

John Paparrizos
No relationship to disclose

Ryen W. White
No relationship to disclose

Eric Horvitz
No relationship to disclose

Appendix

Table A1. Performance of the Models Conditioned on a Variety of Symptom and Risk Factors

		TPR at FPRs Ranging From 0.00001 to 0.1						No.	No.	% All	% All	FPR = 0.00001		
Symptom or Risk Factor	Condition	0.00001	0.0001	0.001	0.01	0.1	AUROC	Positive	Negative	Positive	Negative	Capture	Cost	Capture Cost
Dark or tarry stool	Symptom	7.692	7.692	23.077	38.462	46.154	0.7173*	13	58,597	1.213	1.937	1	0.5860	1.7066
Abdominal swelling/pressure	Symptom	4.167	8.333	16.667	20.833	45.833	0.7735*	24	45,083	2.239	1.490	1	0.4508	2.2183
Ulcers	Risk factor	0.000	0.000	0.000	7.895	50.000	0.7894*	38	16,081	3.545	0.532	0	0.1608	0.0000
Dark urine	Symptom	0.000	5.556	16.667	27.778	50.000	0.8129†	18	51,236	1.679	1.694	0	0.5124	0.0000
Pancreatitis	Risk factor	6.061	9.091	12.121	24.242	54.546	0.8220†	33	34,184	3.078	1.130	2	0.3418	5.8514
Abdominal pain	Symptom	5.385	10.000	16.923	32.308	60.000	0.8343†	130	311,266	12.127	10.290	7	3.1127	2.2489
Enlarged gall bladder	Symptom	0.885	2.655	9.735	25.664	53.982	0.8358†	113	98,454	10.541	3.255	1	0.9845	1.0157
Constipation	Symptom	3.529	7.059	9.412	22.353	57.647	0.8469†	85	317,300	7.929	10.489	3	3.1730	0.9455
Smoking	Risk factor	3.846	3.846	7.692	15.385	53.846	0.8585	26	27,817	2.425	0.920	1	0.2782	3.5945
Blood clot	Symptom	4.494	10.112	14.607	31.461	61.798	0.8589	89	351,385	8.302	11.616	4	3.5139	1.1383
High blood sugar	Symptom	6.135	8.896	16.564	31.595	60.429	0.8611	326	429,543	30.410	14.200	20	4.2954	4.6561
Nausea or vomiting	Symptom	3.200	8.800	17.600	30.400	63.200	0.8706	125	639,502	11.660	21.140	4	6.3950	0.6255
Chills or fever	Risk factor	3.636	7.273	20.909	30.909	65.455	0.8727	110	357,536	10.261	11.819	4	3.5754	1.1188
Loose stool	Symptom	4.615	7.692	18.462	35.385	72.308	0.8756	65	74,720	6.063	2.470	3	0.7472	4.0150
Indigestion	Symptom	7.547	12.264	20.755	38.679	68.868	0.8932	106	504,462	9.888	16.676	8	5.0446	1.5859
Itchy skin	Symptom	18.750	25.000	25.000	25.000	75.000	0.8982	16	79,448	1.493	2.626	3	0.7945	3.7760
Back pain	Symptom	7.801	14.184	19.858	34.752	69.504	0.9047	141	223,586	13.153	7.391	11	2.2359	4.9197
Yellowing sclera or skin	Symptom	2.174	5.439	19.565	38.044	73.913	0.9217	92	85,805	8.582	2.836	2	0.8581	2.3307
Hepatitis	Risk factor	7.692	10.256	20.513	38.462	71.795	0.9275	39	25,158	3.638	0.832	3	0.2516	11.9237
Alcoholism	Risk factor	12.500	16.667	27.083	41.667	89.583	0.9494†	48	32,333	4.478	1.069	6	0.3233	18.5586
Obesity	Risk factor	20.690	20.690	37.931	62.069	82.7590	0.9572†	29	22,153	2.705	0.732	6	0.2215	27.0880
Overall	None	4.851	8.302	17.258	36.474	72.015	0.9003	1,072	3,025,046	100.000	100.000	52	30.2505	1.7190

NOTE. Values below the dashed line have a higher AUROC than Overall. Capture represents the number of true-positive cases in the cohort of positives ∪ negatives at FPR = 0.00001. Cost represents the number of false-positive cases in that same set at FPR = 0.00001. A capture cost ratio > 1.0 means that more people could benefit from an alert than could be mistakenly alerted. Statistically significant differences with Overall model (using DeLong's test [DeLong ER, et al: Biometrics 44:837-845, 1988]) are marked using $^*P < .0001$ and $†P < .001$ (where the significance threshold following a Bonferroni correction is .002).

Abbreviations: AUROC, area under the receiver operating curve; FPR, false-positive rate; TPR, true-positive rate.