Screening for Pancreatic Adenocarcinoma using Signals from Web Search Logs: Feasibility Study and Results

John Paparrizos[1], Ryen W. White, PhD[2*], and Eric Horvitz, MD, PhD[2*]
[1] Columbia University, New York City, NY 10027 USA
[2] Microsoft Research, Redmond, WA 98052 USA

[*]Authors to whom correspondence should be addressed;
E-mail: {ryenw,horvitz}@microsoft.com.

Abstract

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

Methods: We identify searchers in logs of online search activity who issue special queries that are suggestive of a recent diagnosis of pancreatic adenocarcinoma. We jump back many months prior to these landmark queries to examine patterns of symptomatology expressed as searches about concerning symptoms. We build statistical classifiers that predict the future appearance of the landmark queries based on patterns of signals seen in search logs.

Results: We find 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 show specifically that we can identify 5–15% of cases while preserving extremely low false positive rates (0.00001–0.0001).

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 over time.

1. Introduction

Pancreatic adenocarcinoma poses a difficult and resistant challenge in oncology. It is the fourth leading cause of cancer death in men and women in the United States and is the sixth leading cause of 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 one year [4]. Approximately 75% of pancreatic adenocarcinoma patients will die within a year of diagnosis, and only 4% survive for five years post diagnosis (for non-surgical candidates) [5].

Early signs and symptoms of pancreatic adenocarcinoma are subtle and often present as non-specific symptoms that appear and evolve over time. The symptoms often do not become salient until the disease has reached an advanced stage, leaving the patient with fewer options at time of diagnosis. We study a non-traditional, yet promising direction for the early detection of pancreatic adenocarcinoma. The approach centers on the analysis of signals from web search logs. Specifically, we examine 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 who perform searches about sets of relevant symptoms.

People frequently turn to Web search to locate health-related information [6]. For example, searchers concerned about new symptoms often input terms to search engines describing their observations and retrieve ranked lists of results on medical conditions linked to the symptoms. In related work, searchers with cancer diagnoses have been found to perform searches post-diagnosis [7,8,9], revealing strong similarities between temporal patterns in logs and behaviors observed in practice [10,11]. Analyses of symptom and illness-related searches for people over a period of time yields insights about medical concerns and anxieties [12,13] and can provide evidence of healthcare utilization [14]. More generally, search logs have been used to study how people search [15], to predict their next online actions [16, 17], to predict their future interests [18], to improve search engines [19, 20], and to understand in-world activities from long-term activity traces [21].

Screening for pancreatic adenocarcinoma is aimed at detecting the disease at a pre-invasive 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 at the phase in which 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, computer tomography scans, and magnetic resonance imaging [27,28] are useful to diagnose pancreatic adenocarcinoma once the tumor is large enough to cause symptoms that make people seek medical attention, but at this point the disease is more likely to be advanced and unresectable [29]. Earlier diagnosis of pancreatic adenocarcinoma leads to earlier stage [30,31] and improved chance of survival [32,33]. While patients diagnosed early enough to have a curative resection have a higher five-year survival rate, survival at five years for them is under 25% [32].

Surveillance and screening programs for pancreatic adenocarcinoma face the challenge 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 symptomatology.

Patterns of onset and persistence 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 point of view of traditional screening, there are few salient symptoms in early stages of the disease, and the symptoms are not specific enough 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 over 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 input into search engines that provide strong evidence of searchers having recently received a diagnosis from a professional. These are distinct from *exploratory* queries, including searches on symptoms or diseases that 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 users in search logs. We specifically seek evidence of first-person assertions such as the query, “i was just diagnosed with pancreatic adenocarcinoma,” which when associated with prior queries about symptoms, identifies searchers that we label as positive for adenocarcinoma. Searchers who search for one or more related symptoms of interest but show no evidence over time of searches for pancreatic adenocarcinoma constitute the negatives.

2. Methods and Materials

Search services track characteristics of people’s searching and clicking activities to capture intentions, improve their responses, and personalize content. Every such interaction corresponds to a log entry which, apart from the query and the results selected, includes a timestamp and unique, anonymized identifier associated with each browser. This identifier enables the extraction of the search log history comprising queries and clicks of an individual for up to 18 months. The identifier is tied to the machine, so may comprise the search activity of multiple users on shared machines and does not consolidate activity from a single user across multiple machines. We use proprietary search logs from Bing.com from users in the English-speaking United States locale, from October 2013 until May 2015 inclusive.

2.1 Symptoms and Risk Factors

We reviewed the signs, symptoms, and risk factors associated with pancreatic adenocarcinoma. We developed a symptom set covering the following set of concerns: yellow sclera or skin, blood clot, light stool, loose stool, enlarged gallbladder, 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 of the symptoms were identified (e.g., symptom: yellow skin or eyes, synonym: jaundice; symptom: abdominal pain, synonyms: belly pain, stomach ache). We also identified risk factors (e.g., pancreatitis, alcoholism) and their associated synonyms (see [34]), that describe attributes, characteristics or exposure that may increase the likelihood of developing pancreatic adenocarcinoma. The symptoms and the risk factors are mapped to terms in search queries. Searching activities provide streams of data to construct a statistical model that can be used to risk-stratify searchers for screening.

2.2 Extracting Pancreatic Adenocarcinoma Searchers and Symptom Searchers

To identify positive and negative cases in generating a learned model, we built a dataset of users from two groups (Figure 1(a)). Pancreatic adenocarcinoma searchers (*A*) includes all searchers that input one or more queries matching the expression [('pancreas' OR 'pancreatic') AND 'cancer']. We consider as searchers with a diagnosis of pancreatic adenocarcinoma (*B*) the subset of searchers *A* who issue one or more experiential diagnostic queries. Symptom searchers (*C*) includes all users with one or more queries related to pancreatic adenocarcinoma symptoms or synonyms (Section 2.1).

[Figure 1 goes here]

The full search histories (queries, clicked results) of 9.2 million searchers comprise the union of *A* and *C*. 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 the age and gender of searchers. The demographic classifiers are trained on ground truth demographics provided by a separate set of users. Using these statistical models as filters, we identified users for whom > 20% of their queries are health-related and removed these searchers from the study as likely being healthcare professionals. The 20% threshold has been shown in earlier work to accurately identify HCPs [35]. In total, 7.4 million users remained, from which 479,787 were pancreatic adenocarcinoma searchers. As additional features for the statistical analysis of searchers, we used a topic classifier that provides distributions of topics for sets of queries and clicked results [36]. We also considered the dominant geolocation for searcher by using a table that links the IP address of the searcher's browser to locations.

2.3 Positive and Negative Cases

We create *query timelines* for users labeled as experiential diagnostic and experiential symptom searchers, and then draw sets of observations from these timelines for use in the construction of a risk-stratification model. Figure 1(b) summarizes the strategies for identifying positives and negatives. Query timelines are aligned across users based on the point where people issue the first experiential diagnostic query. To ensure sufficient data about each user, we removed from the study searchers with fewer than five search sessions[1] spanning five different days. This reduced the user population to 6.4 million users, with a mean total duration (between first and last queries) of 210.32 days (standard deviation (SD) of 182.93 days and interquartile range of

[1] Search sessions comprise a sequence of search actions with no more than 30-minutes between actions [17].

120 days).

2.3.1 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 include 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 users, 3,203 matched the diagnostic query patterns. Experiential users must search for at least one symptom *prior to* 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 users). The symptom lookup period starts when the first symptom is detected in our symptom set (mean duration (α) = 109.34 days, SD = 49.66 days). For positives, the symptom lookup period terminates at least one week before diagnosis (β=1 week) to reduce the likelihood of overlap between them (which could add noise to model training and testing), while still allowing us to understand predictive performance with minimal lead times.

2.3.2 Negative Cases

To generate a set of users we consider as negative for pancreatic adenocarcinoma, we sample from the users who searched for pancreatic adenocarcinoma symptoms but who did not search for pancreatic adenocarcinoma directly anywhere in their timeline. For flexibility, observational features comprised both absolute and relative values. The use of absolute feature values meant that our model would perform well if there are differences in symptom lookup durations between positives and negatives.

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 = 0.7017) and number of queries (D = 0.003, p = 0.7681), even though the latter was not a filtering criterion.

2.4 Early Detection

We frame early detection as a binary classification challenge using a statistical classifier. We train 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 focus on maintaining a very low false-positive rates (i.e., 1 misprediction in 100k correctly identified cases) while retaining a high imbalance ratio of positives and negatives (i.e., one thousand positives vs. millions of negatives).

The set of observations or features extracted from the symptom lookup period are grouped into five categories: (i) user demographic information, including age/gender predictions, and dominant location (*Demographics*); (ii) characteristics about user sessions, query classes, and URL classes, including activity characteristics and the topics of queries issued and resources accessed (*SearchCharacteristics*); (iii) characteristics about symptoms searched, including generic symptom searching (e.g., number of distinct symptoms) (*SymptomGeneral*)

and features for each symptom (*SymptomSpecific*); (iv) features that capture the temporal dynamics of the features (e.g., increasing/decreasing over time, rate of change) (*Temporal*), and (v) 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 employed to minimize the risk of overfitting. See [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 use the area under the receiver operating characteristic curve (AUROC) and the recall (TPR, true positive rate) at low false positive rates (FPRs) as evaluation metrics. Model generalizability is assessed using 10-fold cross validation, stratified by user.

3 Results

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

3.1 Performance by Week

Prediction performance can change as we increase the lead time between prediction and diagnostic query. We selected 337 positives and 945,394 negatives who were still observed in the logs many weeks prior to Exp_0, and report results for β = 1–21 weeks. Since feature generation requires four weeks of data, for inclusion at Exp_0 – 21 weeks a user needs to be observed at Exp_0 – 25 weeks.

We trained a model for the filtered set of users as for all searchers. Table 1 reports the TPR at different false positive rates for this same set of users at different four-week increments, as well as the AUROC. Performance drops consistently with increased lead time, but even 21 weeks before Exp_0 the predictive performance is still strong (AUROC = 0.8315, TPR (FPR=0.00001) = 6.528%).

[Table 1 goes here]

3.2 Contributions by Observation Type

Table 2 shows the observation types (features) with highest evidential weight. Direction is based on correlations between the feature and training data labels. Number of distinct pancreatic adenocarcinoma symptoms was 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 and have many explanations).

[Table 2 goes here]

Observations also vary in predictive power, e.g., temporal dynamics (AUROC = 0.8391, TPR (FPR=0.00001) = 0.2985%), specific symptoms (AUROC = 0.8176, TPR = 2.800%), demographics (AUROC = 0.6565, TPR = 0.2800%), differing significantly from the full model ($p < 0.01$).

3.3 Symptoms and Risk Factors

The presence of specific symptoms and risk factors in user's query timelines could impact 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 [26,39,40,41,42,43].

We applied cross-validation. For training, we learned a model on those in the nine folds allocated to training. For testing, we iterated through symptoms and risk factors and isolated users 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. Table 3 presents statistics on the performance for each model with ≥ 10 positives (to help ensure that AUROC calculations were meaningful). TPRs at different false positive rates are shown, as are the percentage of positives or negatives with symptom or risk factor searches. In the last three columns are 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 users 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 about details of designs of methods for engaging people.

Table 3 shows that considering only users who search for information related to risk factors such as smoking, hepatitis, and obesity leads to better overall performance. Fewer than ten users searched for each of the cancer syndromes (e.g., hereditary nonpolyposis colorectal cancer) and these cases were excluded from Table 3. We found terms for symptoms and risk factors that are more likely to occur in positives (e.g., pancreatitis is almost seven times as likely, smoking is three times as likely). If we fix FPR = 0.00001, overall we correctly detect 52 users (TPs) but mistakenly alert 30 users (FPs) (capture/cost ratio = 1.72). Table 3 also shows that conditionalizing on specific symptoms / risk factors have significantly higher capture/cost ratios. For example, for alcoholism or obesity, we find 20–30 times as many TPs as FPs.

4. Discussion

Web search logs may offer a useful source of signal for pancreatic adenocarcinoma screening, with significant lead time (e.g., five months before the diagnostic query TPR is 6–32% at extremely low FPRs)[2]. Since pancreatic adenocarcinoma may progress from stage I to stage IV in

[2] 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 non-experiential pancreatic adenocarcinoma users (gray region in Figure 1(a)) were included to mimic a realistic application scenario. Both studies yielded results similar to those reported herein. A final experiment where non-experiential users were included as negatives for training (and testing occurred on the same separate set of logs) revealed a

just over one year [4], this screening capability could increase five-year survival. There are others such as nausea or vomiting, or chills or fever, where costs in mistakenly identifying users and recommending that they seek professional medical attention could outweigh the benefits.

We acknowledge several limitations. Per log anonymity we lack explicit ground truth about diagnoses and rely on implicit self-reporting in queries. Streams of queries following the experiential queries provide confirmatory evidence of a pancreatic adenocarcinoma diagnosis. In the weeks immediately following Exp_0, over 40% of users searched for treatment options, with many using sophisticated terminology (e.g., Whipple procedure, pancreaticoduodenectomy, neoadjuvant therapy) and over 20% searched for related medications (e.g., gemcitabine, 5-fu). In contrast, only 0.5% and 0.02% of those in our negative set 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 need to be directly involved in our studies going forward.

To understand how particular symptoms or risk factors impacted model performance, in applying the model we excluded users without supporting evidence for each symptoms or risk factor in their search histories. An alternative would be to train a separate model for each symptom or risk factor. However, there were insufficient positive examples in each dataset 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 in search histories is more similar to how the model will be employed in practice.

Our approach leverages low-cost passive observation rather than active screening. This could generalize to other chronic diseases for which noticeable symptoms are present. Active screening is not cost effective unless there is a reasonable probability of detecting invasive or preinvasive disease (e.g., at least 16% [45]). Search log-based (retrospective) methodologies support the characterization of individuals' longitudinal behaviors at a scale infeasible in other studies, which are typically much smaller, e.g., [46,47]. 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 patients to the need to seek medical care is a challenging emerging area. Surveillance systems need to convey the uncertainties associated with detection outcomes, while also balancing other issues such as searcher alarm and anxiety, 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.

---

drop in AUROC and TPR. Including the non-experiential pancreatic adenocarcinoma users may add noise to model training. See [31] for more details.

References

1. D. Michaud, Epidemiology of pancreatic cancer, *Minerva Chirurgica* 59(2), 99–111 (2004).
2. R. H. Hruban, M. Goggins, J. Parsons, S. E. Kern, Progression model for pancreatic cancer, *Clinical Cancer Research* 6(8), 2969–2972 (2000).
3. D. Li, K. Xie, R. Wolff, J. L. Abbruzzese, Pancreatic cancer, *The Lancet* 363(9414), 1049– 1057 (2004).
4. J. Yu, A. L. Blackford, M. dal Molin, C. L. Wolfgang, M. Goggins, Time to progression of pancreatic ductal adenocarcinoma from low-to-high tumour stages, *Gut* (2015).
5. K. Y. Bilimoria, D. J. Bentrem, C. Y. Ko, J. Ritchey, A. K. Stewart, D. P. Winchester, M. S. Talamonti. Validation of the 6th edition AJCC pancreatic cancer staging system, *Cancer* 110(4), 738-744 (2007).
6. S. Fox, M. Duggan, *Health online* 2013 (2013).
7. J. L. Bader, M. F. Theofanos, Searching for cancer information on the internet: analyzing natural language search queries, *Journal of Medical Internet Research* 5(4) (2003).
8. K.Castleton, et al., A survey of internet utilization among patients with cancer, *Supportive Care in Cancer* 19(8), 1183–1190 (2011).
9. P. R. Helft, Patients with cancer, internet information, and the clinical encounter: a taxonomy of patient users, *American Society of Clinical Oncology Educational Book*, e89–92. (2011).
10. Y. Ofran, O. Paltiel, D. Pelleg, J. M. Rowe, E. Yom-Tov, Patterns of information-seeking for cancer on the internet: an analysis of real world data, *PLoS ONE* 7(9) e45921(2012).
11. M. J. Paul, R. W. White, E. Horvitz, Search and breast cancer: On disruptive shifts of attention over life histories of an illness, ACM Transactions on the Web 10(2), (2016).
12. R. W. White, E. Horvitz, Cyberchondria: studies of the escalation of medical concerns in web search, *ACM Transactions on Information Systems* 27(4), 23 (2009).
13. C. Lauckner, G. Hsieh, The presentation of health-related search results and its impact on negative emotional outcomes, *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems*, 333–342 (2013).
14. R. W. White, E. Horvitz, From health search to healthcare: explorations of intention and utilization via query logs and user surveys, *Journal of the American Medical Informatics Association* 21(1), 49–55 (2014).
15. R. W. White, S. M. Drucker, Investigating behavioral variability in web search, *Proc. World Wide Web Conference*, 21–30 (2007).
16. T. Lau, E. Horvitz. Patterns of search: analyzing and modeling web query refinement, *Proc. User Modeling Conference*, 119–128 (1999).
17. D. Downey, S.T. Dumais, E. Horvitz. (2007). Models of searching and browsing: languages, studies, and applications. *Proc. International Joint Conference on Artificial Intelligence*, 2740–2747.
18. G. Dupret, B. Piwowarski, A user browsing model to predict search engine click data from past observations. *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval*, 331–338 (2008).
19. T. Joachims, Optimizing search engines using clickthrough data. *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 133–142 (2002).

20. B. Tan, X. Shen, C. Zhai. Mining long-term search history to improve search accuracy. *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 718–723 (2006).
21. M. Richardson. Learning about the world from long-term query logs. *ACM Transactions on the Web*, 2(4), 21 (2009).
22. J. Klapman, M. P. Malafa, Early detection of pancreatic cancer: why, who, and how to screen, *Cancer Control* 15(4), 280–287 (2008).
23. S. S. Coughlin, E. E. Calle, A. V. Patel, M. J. Thun, Predictors of pancreatic cancer mortality among a large cohort of united states adults, *Cancer Causes and Control* 11(10), 915–923 (2000).
24. R. E. Brand, H. T. Lynch, Hereditary pancreatic adenocarcinoma: a clinical perspective, *Medical Clinics of North America* 84(3), 665–675 (2000).
25. H. T. Lynch, T. Smyrk, S. E. Kern, R. H. Hruban, C. J. Lightdale, S. J. Lemon, J. F. Lynch, L. R. Fusaro, R. M. Fusaro, P. Ghadirian, Familial pancreatic cancer: a review, *Seminars in Oncology*, 23(2), 251–275 (1996).
26. A. B. Lowenfels, P. Maisonneuve, G. Cavallini, R. W. Ammann, P. G. Lankisch, J. R. Andersen, E. P. Dimagno, A. Andren-Sandberg, L. Domellof, Pancreatitis and the risk of pancreatic cancer, *New England Journal of Medicine* 328(20), 1433–1437 (1993).
27. H. R. Mertz, P. Sechopoulos, D. Delbeke, S. D. Leach, EUS, PET, and CT scanning for evaluation of pancreatic adenocarcinoma, *Gastrointestinal Endoscopy* 52(3), 367–371 (2000).
28. M. Müller, C. Meyenberger, P. Bertschinger, R. Schaer, B. Marincek, Pancreatic tumors: evaluation with endoscopic US, CT, and MR imaging, *Radiology* 190(3), 745–751 (1994).
29. P. Legmann, O. Vignaux, B. Dousset, A. J. Baraza, L. Palazzo, I. Dumontier, J. Coste, A. Louvel, G. Roseau, D. Couturier, A. Bonnin, Pancreatic tumors: comparison of dual-phase helical CT and endoscopic sonography, *American Journal of Roentgenology* 170(5), 1315–1322 (1998).
30. S. T. Chari, K. Kelly, M.A. Hollingsworth, S.P. Thayer, D. A. Ahlquist, D. K. Andersen, S. K. Batra, T. A. Brentnall, M. Canto, D. F. Cleeter, M. A. Firpo, Early detection of sporadic pancreatic cancer: summative review, *Pancreas* 44(5), 693 (2015).
31. S. A. Melo, L. B. Luecke, C. Kahlert, A. F. Fernandez, S. T. Gammon, J. Kaye, V. S. LeBleu, E. A. Mittendorf, J. Weitz, N. Rahbari, C. Reissfelder, Glypican-1 identifies cancer exosomes and detects early pancreatic cancer, *Nature* 523 (177–182) (2015).
32. C. J. Yeo, R. A. Abrams, L. B. Grochow, T. A. Sohn, S. E. Ord, R. H. Hruban, M. L. Zahurak, W. C. Dooley, J. Coleman, P. K. Sauter, H. A. Pitt, Pancreaticoduodenectomy for pancreatic adenocarcinoma: postoperative adjuvant chemoradiation improves survival. a prospective, single-institution experience, *Annals of Surgery* 225(5), 621 (1997).
33. S. C. Mayo, H. Nathan, J. L. Cameron, K. Olino, B. H. Edil, J. M. Herman, K. Hirose, R. D. Schulick, M. A. Choti, C. L. Wolfgang, T. M. Pawlik, Conditional survival in patients with pancreatic ductal adenocarcinoma resected with curative intent, *Cancer* 118(10), 2674–2681 (2012).
34. A. B. Lowenfels, P. Maisonneuve, Epidemiology and risk factors for pancreatic cancer, *Best Practice & Research Clinical Gastroenterology* 20(2), 197–209 (2006).
35. R. W. White, R. Harpaz, N. H. Shah, W. DuMouchel, E. Horvitz, Toward enhanced pharmacovigilance using patient-generated data on the internet, *Nature Clinical Pharmacology and Therapeutics* 96(2), 239–246 (2014).

36. P. N. Bennett, K. Svore, S. T. Dumais, Classification-enhanced ranking, *Proc. World Wide Web Conference*, 111–120 (2010).
37. J. H. Friedman, Greedy function approximation: a gradient boosting machine, *Annals of Statistics* (2001), 1189–1232.
38. J. Paparrizos, R. W. White, E. Horvitz, Detecting devastating diseases in search logs, *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, in press (2016).
39. C. S. Fuchs, G. A. Colditz, M. J. Stampfer, E. L. Giovannucci, D. J. Hunter, E. B. Rimm, W. C. Willett, F. E. Speizer, A prospective study of cigarette smoking and the risk of pancreatic cancer, *Archives of Internal Medicine* 156(19), 2255–2260 (1996).
40. G. Talamini, C. Bassi, M. Falconi, N. Sartori, R. Salvia, L. Rigo, A. Castagnini, V. Di Francesco, L. Frulloni, P. Bovo, B. Vaona, Alcohol and smoking as risk factors in chronic pancreatitis and pancreatic cancer, *Digestive Diseases and Sciences* 44(7), 1303–1311 (1999).
41. A. M. Goldstein, M. C. Fraser, J. P. Struewing, C. J. Hussussian, K. Ranade, D. P. Zametkin, L. S. Fontaine, S. M. Organic, N. C. Dracopoli, W. H. Clark Jr, M. A. Tucker, Increased risk of pancreatic cancer in melanoma-prone kindreds with p16 ink4 mutations, *New England Journal of Medicine* 333(15), 970–975 (1995).
42. E. Gold, S. Goldin, Epidemiology of and risk factors for pancreatic cancer, *Surgical Oncology Clinics of North America* 7(1), 67–91 (1998).
43. F. M. Giardiello, J. D. Brensinger, A. C. Tersmette, S. N. Goodman, G. M. Petersen, S.V. Booker, M. Cruz–Correa, J. A. Offerhaus, Very high risk of cancer in familial Peutz–Jeghers syndrome, *Gastroenterology* 119(6), 1447–1453 (2000).
44. J. Everhart, D. Wright, Diabetes mellitus as a risk factor for pancreatic cancer: a metaanalysis, *Journal of American Medical Association* 273(20), 1605–1609 (1995).
45. S. J. Rulyak, M. B. Kimmey, D. L. Veenstra, T. A. Brentnall, Cost-effectiveness of pancreatic cancer screening in familial pancreatic cancer kindreds, *Gastrointestinal Endoscopy* 57(1), 23–29 (2003).
46. R. Huxley, A. Ansary-Moghaddam, A. B. De González, F. Barzi, M. Woodward, Type-ii diabetes and pancreatic cancer: a meta-analysis of 36 studies, *British Journal of Cancer* 92(11), 2076–2083 (2005).
47. A. G. Renehan, M. Tyson, M. Egger, R. F. Heller, M. Zwahlen, Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies, *The Lancet* 371(9612), 569–578 (2008).
48. E. R. DeLong, D. M. DeLong, D. L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, *Biometrics*, 837–845 (1988).

Table 1. Performance at early prediction task at four-week intervals for the set of users for whom features can be computed from Exp_0 – 1 week to Exp_0 – 21 weeks. Values are averaged across the ten folds of the cross-validation. Significance of differences in AUROC and TPR using paired t-tests for each week versus Exp_0 – 1 is indicated as * $p < 0.01$, ** $p < 0.001$, and *** $p < 0.0001$. Weeks denotes the week prior to first experiential diagnostic query when the prediction is made (e.g., "5 weeks" means learn from data up to five weeks before the first experiential diagnostic query (Exp_0)).

Weeks before Exp_0 (β in Figure 1(b))	**TPR 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**

Table 2. Top 10 features, ranked in descending order by evidential weight. Weights are relative to the top weighted feature, *number of distinct symptoms searched*, assigned a weight of 1.000. Positive/Negative means that the feature correlates positively/negatively with ground truth.

Observation Type	**Weight**	**Direction**	**Class**
Number of distinct symptoms searched	1.0000	Positive	SymptomGeneral
Fraction of search queries that are health related	0.8253	Positive	QueryTopic
Number of distinct symptom synonyms searched	0.6899	Positive	SymptomGeneral
Probability that user age is 50–85 years	0.6889	Positive	Demographic
User has searched for back pain	0.6622	Negative	SymptomSpecific
User has searched for indigestion	0.6432	Negative	SymptomSpecific
User has searched for indigestion *then* abdominal pain	0.6349	Positive	Temporal
Gradient of best-fit line for number of distinct symptoms searched	0.6154	Positive	Temporal
User has searched for back pain *then* yellow skin or eyes	0.6004	Positive	Temporal
Probability that user age is < 18 years	0.5869	Negative	Demographic

Table 3 (ONLINE ONLY). Performance of the models conditioned on a variety of symptom and risk factors. 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 represents the number of FP cases in that same set at FPR = 0.00001. A capture-cost ratio of > 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 [48]) are marked using ** $p < 0.001$ and ***$p < 0.0001$ (where the significance threshold following a Bonferroni correction is 0.002).

Symptom or risk factor	Condition	TPR at FPRs ranging from 0.00001–0.1					AUROC	# pos	# neg	% all pos	% all neg	False positive rate = 0.00001		
		0.00001	0.0001	0.001	0.01	0.1						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	1.7065
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
Pancreatitis	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
Ulcers	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 gallbladder	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
Yellow skin or eyes	Symptom	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
Loose stool	Symptom	3.636	7.273	20.909	30.909	65.455	0.8727	110	357,536	10.261%	11.819%	4	3.5754	1.1188
Chills or fever	Risk factor	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
Smoking	Risk factor	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

Figure 1. (a) Venn diagram depicting the sets of users employed in the search log analysis: pancreatic cancer searchers (*A*), pancreatic adenocarcinoma searchers with experiential diagnostic queries (*B*), and those who search for pancreatic adenocarcinoma symptoms (*C*). $|A \cup C|$ (i.e., the total number users in our original, pre-filtered dataset) is 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.

Figure 1(a) – User sets

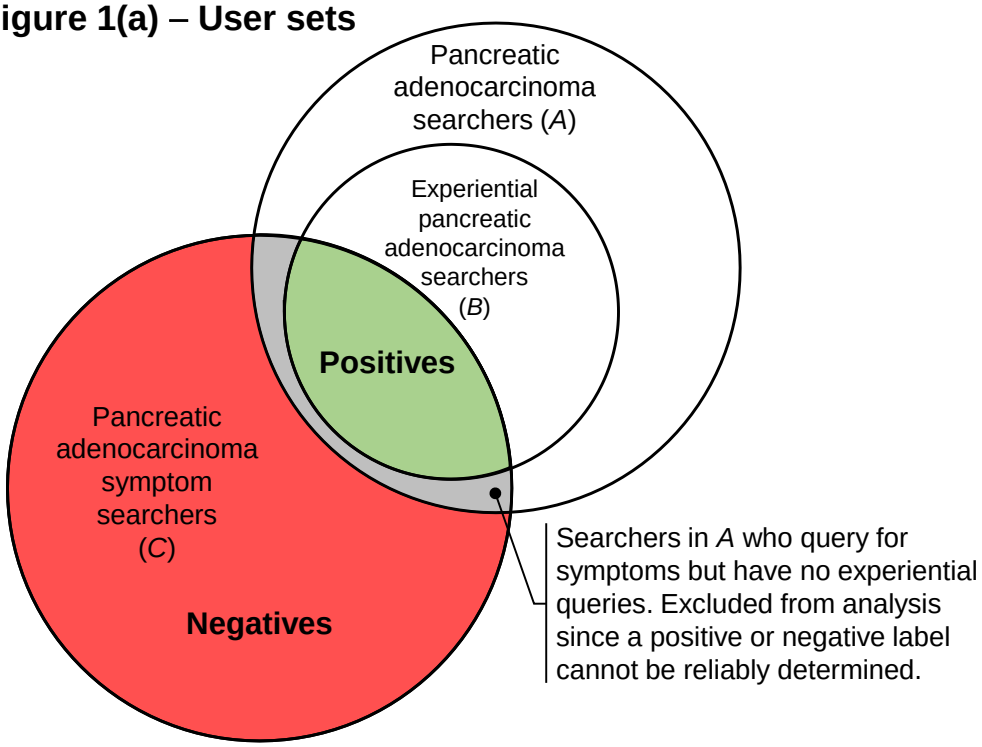

Figure 1(b) – Query timelines