JAMA Oncology | Brief Report

Evaluation of the Feasibility of Screening Patients for Early Signs of Lung Carcinoma in Web Search Logs

Ryen W. White, PhD; Eric Horvitz, MD, PhD

IMPORTANCE A statistical model that predicts the appearance of strong evidence of a lung carcinoma diagnosis via analysis of large-scale anonymized logs of web search queries from millions of people across the United States.

OBJECTIVE To evaluate the feasibility of screening patients at risk of lung carcinoma via analysis of signals from online search activity.

DESIGN, SETTING, AND PARTICIPANTS We identified people who issue special queries that provide strong evidence of a recent diagnosis of lung carcinoma. We then considered patterns of symptoms expressed as searches about concerning symptoms over several months prior to the appearance of the landmark web queries. We built statistical classifiers that predict the future appearance of landmark queries based on the search log signals. This was a retrospective log analysis of the online activity of millions of web searchers seeking health-related information online. Of web searchers who queried for symptoms related to lung carcinoma, some (n = 5443 of 4 813 985) later issued queries that provide strong evidence of recent clinical diagnosis of lung carcinoma and are regarded as positive cases in our analysis. Additional evidence on the reliability of these queries as representing clinical diagnoses is based on the significant increase in follow-on searches for treatments and medications for these searchers and on the correlation between lung carcinoma incidence rates and our log-based statistics. The remaining symptom searchers (n = 4 808 542) are regarded as negative cases.

MAIN OUTCOMES AND MEASURES Performance of the statistical model for early detection from online search behavior, for different lead times, different sets of signals, and different cohorts of searchers stratified by potential risk.

RESULTS The statistical classifier predicting the future appearance of landmark web queries based on search log signals identified searchers who later input queries consistent with a lung carcinoma diagnosis, with a true-positive rate ranging from 3% to 57% for false-positive rates ranging from 0.00001 to 0.001, respectively. The methods can be used to identify people at highest risk up to a year in advance of the inferred diagnosis time. The 5 factors associated with the highest relative risk (RR) were evidence of family history (RR = 7.548; 95% CI, 3.937-14.470), age (RR = 3.558; 95% CI, 3.357-3.772), radon (RR = 2.529; 95% CI, 1.137-5.624), primary location (RR = 2.463; 95% CI, 1.364-4.446), and occupation (RR = 1.969; 95% CI, 1.143-3.391). Evidence of smoking (RR = 1.646; 95% CI, 1.032-2.260) was important but not top-ranked, which was due to the difficulty of identifying smoking history from search terms.

CONCLUSIONS AND RELEVANCE Pattern recognition based on data drawn from large-scale web search queries holds opportunity for identifying risk factors and frames new directions with early detection of lung carcinoma.

Supplemental content

Author Affiliations: Microsoft Research, Redmond, Washington.

Corresponding Author: Ryen W. White, PhD, Microsoft Research, One Microsoft Way, Redmond, WA 98052 (ryenw@microsoft.com).

JAMA Oncol. doi:10.1001/jamaoncol.2016.4911
Published online November 10, 2016.

Lung carcinoma is the leading cause of cancer death in the United States.[1] Patient prognosis is strongly correlated with stage at diagnosis.[2] Most (>75%) present with stage III or IV disease and are rarely curable with current therapies.[3] In the absence of resection, 5-year survival rates are low.[4,5] Cost-effective methods for earlier detection of lung carcinoma could increase these rates significantly. Early signs often present as nonspecific symptoms that appear and evolve longitudinally. Symptoms are not typically salient until the disease has metastasized.

Screening for lung carcinoma involves identifying high-risk individuals and subsequent studies to detect tumors. Possibilities for screening for lung carcinoma have emerged from recent developments in biology and radiology, and from better understanding of high-risk populations.[6] Methods such as low-dose computed tomography (LDCT) can reduce mortality[7] but can also lead to many false-positive results.[8] Other tests, such as sputum cytology and chest radiography, have limited effectiveness.[9] Standing challenges of false-positives and false-negatives, and the costs associated with screening and follow-up, motivate the pursuit of new and complementary methods for early identification of lung carcinoma.

We examined the feasibility of a nontraditional yet promising direction for detecting early signs of lung carcinoma. We studied online signals connected to known risk factors for lung carcinoma and identified new patterns of evidence. The approach analyzes signals from web data, an area of research growing in prominence.[10,11] Population-scale statistical analyses of web search engine log data have already yielded clues for early detection of pancreatic adenocarcinoma.[12]

Key Points

Question Are statistical models learned from large-scale web search logs effective in detecting lung carcinoma in advance of a clinical diagnosis?

Findings In this modeling study, study, a statistical classifier accurately identified web searchers who later input queries that provide evidence of a recent clinical diagnosis of lung carcinoma. The methods can help identify people at highest risk up to a year in advance of the inferred diagnosis time, and identify new risk factors (eg, house age, air travel patterns) expressed as evidence in people's search activity and geographic location.

Meaning Pattern analysis and recognition based on search log data holds opportunity for identifying risk factors and for framing new directions with the early detection of lung carcinoma.

Methods

We harnessed web search log data to build a statistical classifier to stratify searchers per lung carcinoma risk. Logs from the Bing.com service from searchers in the English-speaking United States (May 2014 until October 2015) were used. Logs contained anonymized user identifiers, queries, and timestamps.

Positive and Negative Cases

We defined a set of searchers who issued lung carcinoma queries (*A*) and a set who issued queries on related symptoms (eg, bronchitis, cough, chest pain; see eTable 1 in the Supplement) (*B*). We take as positives those who provide strong evidence of a recent diagnosis via special queries referred to as *experiential* (vs *exploratory*) queries. Such queries include first-person statements (eg, "I was just diagnosed with lung cancer"). Follow-on queries (eg, on specific treatments and adverse effects) can also provide additional evidence of a diagnosis.[13] The intersection of *A* and *B* comprised the experiential searchers (5443 positives), and the remaining subset of *B* comprised exploratory searchers (4 808 542 negatives). Each searcher was associated with a timeline between their first query and a terminal query (*E*), either the experiential query (positives) or their last query in the logs (negatives). The objective was to predict the later appearance of strong indications of a lung carcinoma diagnosis based on data up to *E* minus *L* weeks lead time, where *L* varied from 1 to 52 weeks.

Risk Factors

Long-term tobacco smoking is associated with 85% to 90% of lung carcinoma cases.[14] Other risk factors include exposure to radon gas, asbestos, second-hand smoke, air pollution, and nutritional supplement use.[15] eTable 2 in the Supplement enumerates the risk factors analyzed. Evidence for the presence of many factors was obtained via query terms. Searcher age and sex were inferred via automated classifiers from Bing.com. Location data from reverse internet protocol lookup were used in location-specific factors, such as radon and air travel.

Early Detection

The classifier was trained on search terms and related information from searcher timelines. The set of observations or features extracted prior to *E* were grouped into risk factors and symptoms. Symptoms were identified via query terms matching a symptom set defined via literature review, including presence or absence and timing. Predictive power is measured via recall (true-positive rate [TPR]) at different target maximally tolerated false-positive rates (FPRs) and area under the receiver operating characteristic curve (AUROC).

Results

Overall

We report performance for different FPR thresholds, ranging from FPR = 0.00001 (1 error in 100 000 cases) to FPR = 0.1 (1 error in 10 cases). We performed the predictions using data up to the first experiential query (ie, *E* − 1 week). Model performance (overall) was strong, with AUROC = 0.9535, and TPRs ranged from 3% to 57% for FPRs from 0.00001 to 0.001. Searched symptoms were informative, especially those about bronchitis and coughing. Risk factors were also valuable: most informative are the likelihood that the searcher was male and was below the poverty line (proxies for smoking). Other informative risk factors include the number of older homes in the searcher's geographic region (which may lack radon mitigation) and higher frequencies of air travel.

Table. Performance at Early Prediction at 4-Week Intervals for the Set of Searchers for Whom Features Can Be Computed From 1 to 52 Weeks Before the First Experiential Query[a]

Weeks Before First Experiential Query	TPR (%) at FPRs Ranging From 0.00001 to 0.1					AUROC
	0.00001	0.0001	0.001	0.01	0.1	
1 wk	3.315	18.907	57.397	73.542	86.433	0.9415
13 wk (about 3 mo)	2.947	16.943	52.977	67.526	85.267	0.9311
26 wk (about 6 mo)	2.333[b]	14.549	49.110[b]	63.781[b]	83.917[b]	0.9120
39 wk (about 9 mo)	1.842[c]	12.277[b]	44.260[c]	57.950[b]	75.752[c]	0.8891[b]
52 wk (12)	1.473[c]	10.068[c]	39.288[c]	52.363[c]	69.613[c]	0.8662[c]

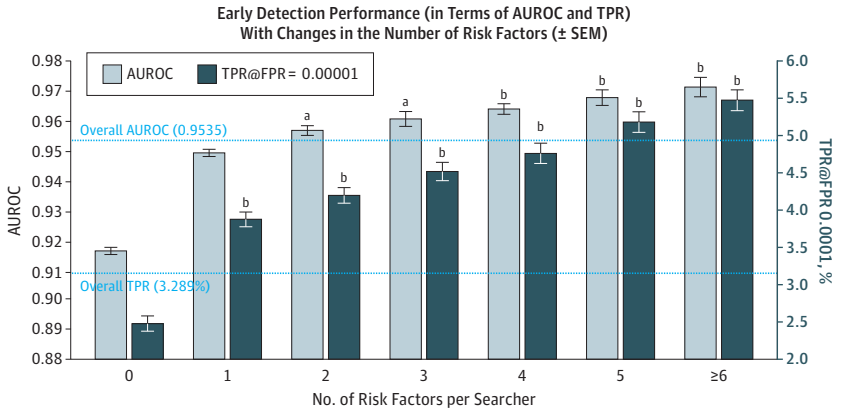
Abbreviations: AUROC, area under the receiver operating characteristic curve; FPR, false-positive rate; TPR, true-positive rate.

[a] Values are averaged across the 10-folds of the cross-validation. Weeks denotes the lead time prior to first experiential query, when the prediction is made.

[b] Significance of differences in AUROC and TPR using paired *t*-tests for each week vs 1 week before first experiential query was $P < .01$.

[c] Significance of differences in AUROC and TPR using paired *t*-tests for each week vs 1 week before first experiential query was $P < .001$.

Figure. Plots of the Area Under the Receiver Operating Characteristic Curve (AUROC) and True-Positive Rate (TPR) When the False-Positive Rate (FPR) Is Limited to 0.00001 (TPR@FPR) for Different Numbers of Risk Factors per Web Searcher



Risk factors, No.	0	1	2	3	4	5	≥6	≥0 (All cases)
Positive cases, No.	1251	1496	1072	664	420	193	347	5443
Negative cases, No.	1153900	1254239	844987	620332	405045	174244	355795	4808542
Positive cases, %	0.108	0.119	0.127	0.107	0.104	0.111	0.097	0.113

FPR = 0.00001 means that the model makes a prediction error 1 time in every 100 000 cases. Searchers with no risk factors (0 in the plot) search only for symptoms and have no risk factors that are significantly different from the background population. Performance of the overall model trained on all cases has an AUROC of 0.9535. (The TPR [percentage of positives in the set that are recalled by the model] is 3.289% when the FPR is limited to 0.00001). Error bars denote standard error of the mean. Also shown are number of searchers in each group and number and percentage of each group that is a positive case.

[a] Significance of differences in AUROC and TPR vs overall model computed using independent measures *t*-tests are denoted using $P < .01$.

[b] Significance of differences in AUROC and TPR vs overall model computed using independent measures *t*-tests was $P < .001$.

Lead Time Variation

To understand classifier performance with increasing lead times, we backtrack to $L = 52$ and consider recall moving forward at 3-month intervals to the first experiential query for the 1629 positives and 57 583 negatives observed over 52 weeks. The **Table** shows that the classifier performs effectively up to 1 year before the first experiential query.

Risk Factors

The 5 factors associated with the highest relative risk (RR) were evidence of family history (RR = 7.548; 95% CI, 3.937-14.470), age (RR = 3.558; 95% CI, 3.357-3.772), radon (RR = 2.529; 95% CI, 1.137-5.624), primary location (RR = 2.463; 95% CI, 1.364-4.446), and occupation (RR = 1.969; 95% CI, 1.143-3.391). Evidence of smoking (RR = 1.646; 95% CI, 1.032-2.260) was important but not top-ranked, highlighting the difficulty of identifying smoking history from search terms.

The "Overall" subsection in this section reported classifier performance on all features, including all risk factors. We also explored levels of recall for distinct cohorts of searchers at highest risk, based on considering risk factors. Whether the risk factor applied to a searcher was based on the presence of a significant difference ($P < .01$) between that searcher's feature value and the overall average. The desired directionality varied based on the connection with heightened risk for that feature (eg, for incidence rates, higher values indicate increased risk; for smoking bans, lower values indicate increased risk). We re-ran the detection task 1 week before the first experiential query for searchers with different numbers of risk factors, ranging from zero (only evidence of symptoms) to 6 or more risk factors (**Figure**). The classifier performs best for those at high risk. The presence of at least 1 risk factor significantly improves detection performance, and there are consistent marginal gains for each additional risk factor.

Discussion

The feasibility study highlights opportunity to leverage online behavioral data in prescreening or screening for lung carcinoma, perhaps to complement more traditional screening methods. Our classifier performs best for cohorts exhibiting evidence of key risk factors and makes accurate predictions up to 1 year prior to experiential diagnostic queries (eg, detecting 10% of positives while being incorrect 1 in 10 000 times). The decision threshold can be adjusted per desired model operating characteristics.

A limitation of our methods is the lack of ground truth about diagnosis. The alignment between the dates for experiential queries and actual diagnosis needs to be determined via additional studies. Risk factors inferred from online data were valuable. However, several factors (eg, age, smoking habits, family history) may be best obtained directly from searchers.

Conclusions

In a real-world deployment, web search engines could serve as a filter to identify patients who would benefit from clinical screening. Health-conscious patients may volunteer to receive alerts if concerning activity is detected. Communicating early detection outcomes with searchers without causing unnecessary alarm and associated costs needs more attention. Also related is whether such communication is necessary when outcomes could be passed directly to physicians for consideration and patient follow-up in a clinical setting. The costs and benefits associated with such broader prescreening and screening require detailed study.

ARTICLE INFORMATION

Accepted for Publication: September 12, 2016.

Published Online: November 10, 2016. doi:10.1001/jamaoncol.2016.4911

Author Contributions: Drs White and Horvitz had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: Both authors.
Acquisition, analysis, or interpretation of data: Both authors.
Drafting of the manuscript: Both authors.
Critical revision of the manuscript for important intellectual content: Both authors.
Statistical analysis: Both authors.

Conflict of Interest Disclosures: None reported.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. *CA Cancer J Clin*. 2015;65(1):5-29.

2. Henschke CI, Yankelevitz DF, Libby DM, Pasmantier MW, Smith JP, Miettinen OS; International Early Lung Cancer Action Program Investigators. Survival of patients with stage I lung cancer detected on CT screening. *N Engl J Med*. 2006;355(17):1763-1771.

3. Ihde DC. Chemotherapy of lung cancer. *N Engl J Med*. 1992;327(20):1434-1441.

4. Flehinger BJ, Kimmel M, Melamed MR. The effect of surgical treatment on survival from early lung cancer. Implications for screening. *Chest*. 1992;101(4):1013-1018.

5. Sobue T, Suzuki T, Matsuda M, Kuroishi T, Ikeda S, Naruke T; Japanese Lung Cancer Screening Research Group. Survival for clinical stage I lung cancer not surgically treated: comparison between screen-detected and symptom-detected cases. *Cancer*. 1992;69(3):685-692.

6. Mulshine JL, Henschke CI. Prospects for lung-cancer screening. *Lancet*. 2000;355(9204):592-593.

7. Bach PB, Mirkin JN, Oliver TK, et al. Benefits and harms of CT screening for lung cancer: a systematic review. *JAMA*. 2012;307(22):2418-2429.

8. Aberle DR, Abtin F, Brown K. Computed tomography screening for lung cancer: has it finally arrived? Implications of the national lung screening trial. *J Clin Oncol*. 2013;31(8):1002-1008.

9. Oken MM, Hocking WG, Kvale PA, et al; PLCO Project Team. Screening by chest radiograph and lung cancer mortality: the Prostate, Lung, Colorectal, and Ovarian (PLCO) randomized trial. *JAMA*. 2011;306(17):1865-1873.

10. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*. 2009;457(7232):1012-1014.

11. Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection—harnessing the Web for public health surveillance. *N Engl J Med*. 2009;360(21):2153-2157.

12. Paparrizos J, White RW, Horvitz E. Screening for pancreatic adenocarcinoma using signals from web search logs: feasibility study and results. *J Oncol Pract*. 2016;12(8):737-744.

13. White RW, Horvitz E. Screening for lung carcinoma with web search data. MSR Technical Report. 2016. MSR-TR-2016-51.

14. Molina JR, Yang P, Cassivi SD, Schild SE, Adjei AA. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. *Mayo Clin Proc*. 2008;83(5):584-594.

15. Biesalski HK, Bueno de Mesquita B, Chesson A, et al; Lung Cancer Panel. European Consensus Statement on Lung Cancer: risk factors and prevention. *CA Cancer J Clin*. 1998;48(3):167-176.