Screening for Lung Carcinoma with Web Search Data

Ryen W. White, PhD* and Eric Horvitz, MD, PhD
Microsoft Research, Redmond, WA 98052 USA

*Author to whom correspondence should be addressed;
E-mail: ryenw@microsoft.com

Word counts:

Abstract: 350 words
Body: 3000 words

Abstract

Importance: A statistical model that predicts the future appearance of diagnoses of lung carcinoma via analysis of large-scale anonymized logs of Web search queries input by millions of people across the United States.

Objective: To evaluate the technical feasibility of performing background screening and alerting for patients at risk of lung carcinoma via analysis of online search behavior.

Design: We identify people who issue special queries that provide strong evidence of a recent diagnosis of lung carcinoma. We then consider patterns of symptomatology expressed as searches about concerning symptoms over several months prior to the appearance of landmark queries. We build statistical classifiers that predict the future appearance of landmark queries based on the search log signals.

Setting: Retrospective log analysis of the online activity of millions of Web searchers seeking health-related information online.

Participants: Web searchers who query for symptoms related to lung carcinoma. Some (n=5,443) later issue 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 Outcome Measure(s): Classification performance of our 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 has strong performance in identifying searchers who will later input queries consistent with a lung carcinoma diagnosis, with a true-positive rate ranging from 3-32% for false-positive rates ranging from 0.00001 to 0.001, respectively. We can employ the methods to identify people at highest risk up to a year in advance of the inferred diagnosis time. Beyond studies of predictive power, we identify new risk factors expressed as evidence in people's search activity and location.

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

1. Introduction

Lung carcinoma is the leading cause of cancer death in the United States [1]. Patient prognosis is strongly correlated with the stage of disease at diagnosis time [2]. Most (>75%) present with stage III or IV disease and are rarely curable with current therapies [3]. In the absence of resection, survival rate is only 12% [4,5]. Devising cost-effective methods for earlier detection of lung carcinoma is important for improving survival. Early signs often present as non-specific symptoms that appear and evolve over time. The symptoms typically do not become salient until the disease has metastasized. We study a non-traditional, yet promising direction for early detection of lung carcinoma. The approach centers on the analysis of signals from Web search logs. Recent research has shown that population-scale statistical analyses of the Web search and browsing behavior of millions of people can yield clues about the early detection of pancreatic adenocarcinoma [6]. We examine the feasibility of applying similar methods for detecting the emergence of lung carcinoma from people performing searches about sets of symptoms.

Screening for lung carcinoma involves the identification of high risk individuals (e.g., older with a history of heavy smoking), and subsequent screening to detect tumors. Possibilities for screening for lung carcinoma have come to the fore as a consequence of recent developments in biology and radiology, and better understanding of high-risk populations [7]. Low-dose computed tomography (LDCT) [8] has been demonstrated to reduce the chance of death from lung carcinoma [9]. However, LDCT leads to many false positive tests [10]; for each true positive scan there are 19 falsely positives scans [9,11], resulting in overdiagnosis [12,13,14], and unnecessary tests and surgeries with associated risks and costs. Recurrent LDCT studies present challenges of cumulative radiation exposure [10] and can lead to costly follow-up [15]. Other tests, such as sputum cytology and chest radiography, have limited effectiveness [16,17,18,19,20]. Analysis of molecular genetics is promising [22,23,24], but relies on access to appropriate data to identify patients at risk. Patients may also be missed by these traditional methods (e.g., one study found 26% of negative screenings were false negatives [21]). The standing challenges of false positives and 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 consider the feasibility of identifying the presence of undiagnosed lung carcinoma via information gleaned from Web search logs. People frequently utilize Web search to locate health-related information, including their pursuit of explanations for concerning symptoms [25]. Web searching is common in cancer patients [26,27,28]; patterns of information access over time captured in search logs can align with phases of screening, diagnosis, and treatment [29,30]. Analyses of logged symptom- and illness-related searches yields insights about medical concerns and anxieties [31,32], and can offer evidence about healthcare utilization [33]. More generally, search logs enable search providers and

researchers to learn about search behavior [34], to predict future actions and interests [35,36,37], to improve search engines [38,39], and to understand in-world activities [40].

In addition to analyzing searches for relevant symptoms, we can also consider online signals linked to known risk factors for developing lung carcinoma. The primary related risk factor is long-term tobacco smoking, which is associated with 85-90% of lung carcinomas [41]. However, significant portion (10-15%) of cases are people who have never smoked [42]. There are a range of other risk factors for lung carcinoma including a history of exposure to radon gas [43,44], asbestos [45,46], second-hand smoke [47,48], other forms of air pollution [49,50,51], and the use of certain nutritional supplements [52]. We use signals from these and other risk factors as part of the early-detection methodology.

2. Methods and Materials

Search services such as Bing and Google log people's searching and clicking activities to capture intentions, improve their responses, and personalize content. We harness these streams of data to build a statistical classifier capable of stratifying searchers for risk of lung carcinoma. Every interaction corresponds to a log entry containing the query 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. We use proprietary logs from Bing.com from searchers in the English-speaking United States locale, from May 2014 until October 2015 inclusive. Searcher geolocation data is available based on reverse IP lookup, which provides searcher locations at query time at city and state level. These locations are converted to U.S. county level for risk factor analysis.

2.1 Symptom Searchers

We focused on people searching for one or more lung carcinoma symptoms. These searchers comprised the superset of all searchers from which the positive and negative cases were drawn for our studies. Synonyms for each symptom were identified via literature review (Table 1).

Table 1. Symptoms associated with lung carcinoma and examples of associated synonyms used to generate evidential features in the *Symptom* class.

Symptom	Example synonyms
cough	coughing, cough getting worse, persistent cough, cough lingering
chest pain	chest pains, pain in chest, sore chest, chest hurts when breathing
hoarseness	hoarseness, hoarse voice, deep cough, raspy voice
weight loss	rapid weight loss, unexpected weight loss
loss of appetite	losing appetite, lost appetite, appetite gone, not hungry
coughing up blood	coughing up blood, cough up blood, orange phlegm, orange spit
short of breath	shortness of breath, breathless, breathlessness, easily winded
tiredness	tiredness, feeling tired, feel tired, tired
weakness	weakness, feeling weak, feel weak, weak feeling, weak
bronchitis	persistent bronchitis, chronic bronchitis
wheezing	wheeze, whistling sound breathing
bone pain	bone pain, pain bone, bones sore, sore bones

2.2 Experiential Searchers

Per anonymity of searchers in the logs, we lack access to ground truth data about diagnoses of lung carcinoma. We take as a proxy for diagnosis the issuance of first-person *experiential* search queries such as "just diagnosed with lung cancer" or "I have just been diagnosed with lung cancer." To bolster robustness, we consider only searchers who engage in at least five sessions on at least five distinct days. We perform two studies that provide support for using experiential queries as information about clinical diagnosis. First, we perform correlation analysis with state-level incidence rates published by the U.S. Centers for Disease Control (CDC). We compute the percentage of Bing searchers in each state who issue an experiential diagnostic query for lung carcinoma. The Pearson correlation between these percentages and state-level lung cancer incidence rates is high (r=0.722, $t(47)$=7.154, $p < 0.0001$ (r with 95% lower/upper confidence intervals (CIs) for incidence rates=0.771−0.811), showing that experiential diagnostic queries have alignment with clinical diagnoses (Figure 1). Second, we examine search activities following the experiential diagnostic queries. The searchers marked as positives were significantly more likely to search for lung carcinoma treatment options (e.g., lobectomy, wedge resection) (32.2% of searchers) and related medications (e.g., tarceva, gilotrif) (10.4% of searchers) after their first experiential diagnostic query than the negatives (1.8% and 0.8% of negative cases searched for treatments or medications at any point, $p < 0.01$ with a two-proportion Z-test).

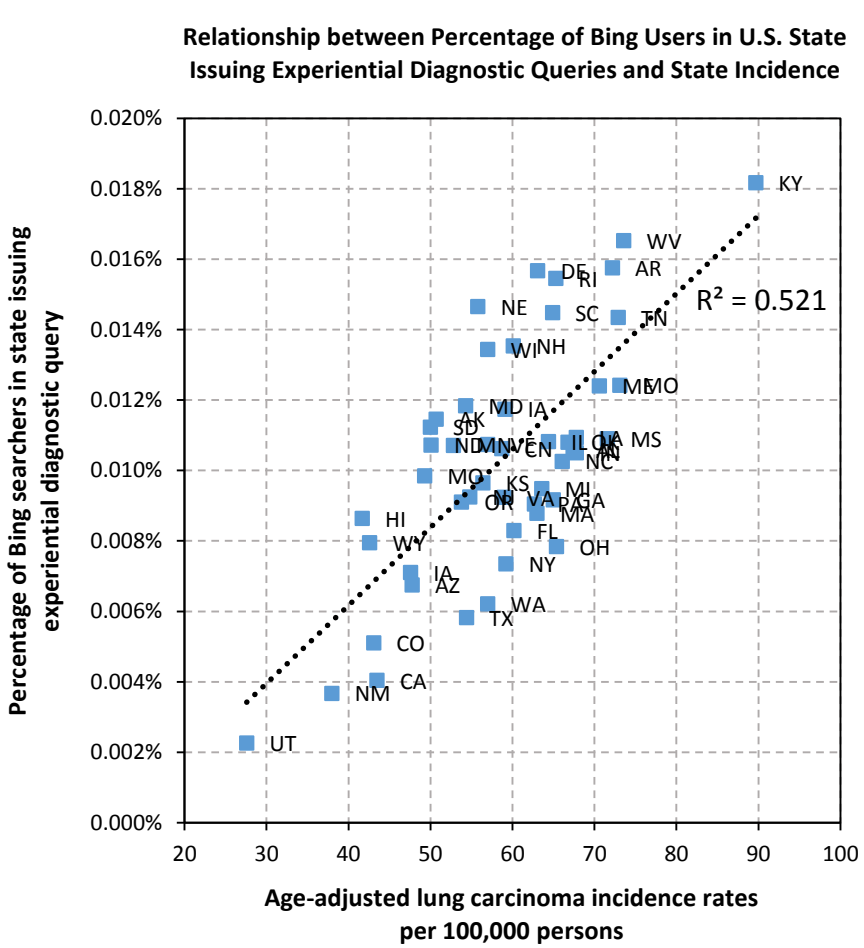


Figure 1. Scatterplot of (i) U.S. state incidence rates of lung carcinoma per 100,000 persons, age-adjusted to the 2000 U.S. standard population (available from the CDC) and (ii) the percentage of all Bing searchers in each state who issue experiential diagnostic queries related to lung carcinoma. States are labeled in the scatterplot using two-letter state name abbreviations. Lung carcinoma incidence rate data for the state of Nevada (NV) was not available.

2.3 Positive and Negative Cases

We defined a set of lung carcinoma searchers (A) and a set of related symptom searchers (B). We focus on the subset of $A \cap B$ comprising the experiential diagnostic searchers (positives) and the remaining subset of B (symptom searchers who do not issue experiential diagnostic queries) as negatives. In total, the dataset contained 5,443 positives and 4,808,542 negatives. For each searcher we defined a query timeline between the first query observed in the logs and the experiential query (positives) or the end of the logs (negatives), as the terminal query (E). The objective is to make the prediction based on data up to E minus L weeks lead time, where L varies from 1 week to 52 weeks. Query timelines had an average duration of

222 days (standard deviation=171.9 days, interquartile range (IQR)=317.8 days, median=180.76 days). We therefore had many months of search history from each searcher for use in early detection analysis.

2.4 Risk Factors

There are a broad range of different risk factors for lung carcinoma. Modeling these in an online setting can be challenging given limited data availability. The occurrence of risk factors is identified either via searchers' queries or derived via special analyses employing external sources. The following risk factors were used, grouped according to distinct classes: (i) incidence and mortality rates for the counties from which the searcher is observed searching, (ii) environmental (e.g., radon gas concentrations, average age of homes, air pollution, presence of naturally occurring asbestos), (iii) geographic (e.g., searcher's primary county, land area of searcher's primary county), (iv) demographic (e.g., age, gender), (v) socioeconomic (e.g., poverty rates, education rates), (vi) habits and activities (e.g., ever smoked, quit smoking, indoor air pollution), and (vii) medical (e.g., specific procedures, incidence rates of HIV or diabetes) (see Table 2 for details).

Table 2. Classes of risk factors and associated data used to operationalize these in statistical model. Risk factors computed at the county level, using data from external sources and averaged across all counties from where the searcher is observed to be searching, or at the searcher level, based on specific search queries issued by the searcher in their search history. References are included for risk factors where studies have demonstrated a potential link with increased likelihood of lung cancer.

Class	Data	Granularity	Description
Incidence and Mortality	Incidence rates	County	US Centers for Disease Control (CDC) incidence rate data on lung cancer.
	Mortality rates	County	CDC mortality rate data on lung cancer.
Environmental	Radon [53,54]	County	US Environmental Protection Agency (EPA) data on radon levels per county, split into three categories: (i) Category 1: > 4 pCi/L (picocuries per liter), (ii) Category 2: 2-4 pCi/L, (iii) Category 3: < 2 pCi/L.
		Individual searcher	Searcher queries for "radon venting", "radon mitigation", etc.
	Year house built	County	House age statistics from responses to US Census American Community Survey (ACS).
	Outdoor air pollution [49,50,51]	County	Air pollution levels (as air quality index and other pollutants) from EPA.
		County	Carbon dioxide equivalent (CO_2e) emissions data from the EPA.
	Arsenic in drinking water [55,56]	County	Levels of arsenic in groundwater from the US Geological Survey (USGS).
		Individual searcher	Searches for "arsenic drinking water", etc.
	Weather	County	Weather statistics from National Oceanic and Atmospheric Administration (NOAA), namely temperature, wind speed, rain, snow, and atmospheric pressure.
	Asbestos [45,46]	County	Locations of asbestos mines, prospects, and occurrences from the USGS.
	Cellphone towers [57]	County	Locations of cellphone towers from Federal Communications Commission (FCC).
		Individual searcher	Searches for "near cellphone tower", etc.
	Powerlines [58,59]	Individual searcher	Searches for "near powerline", etc.
	Nuclear power plants [60]	County	Locations of nuclear power plants from the US Nuclear Regulatory Commission (NRC).
	Passive smoking [47,48]	County	Counties with smoking bans in public places.
		Individual searcher	Searches for "second hand smoke", "passive smoking", "passive smoke", etc.
Geographic	Primary state	Individual searcher	US state from where searcher issues most queries.
	Percentage of queries in primary state	Individual searcher	Percentage of all queries from searcher originating in primary state.
	Primary county	Individual searcher	US county from where searcher issues most queries.
	Percentage of queries in primary county	Individual searcher	Percentage of all queries from searcher originating in primary county.
	Land area total	County	US Census data on the total land area of county (in square miles).
	Percentage of *land area* that is (i) urbanized, (ii) in urban clusters, or (iii) rural [61]	County	US Census data on the percentage of the county land area that is (i) urbanized (50,000 or more

			people), (ii) in urban clusters (2,500 to 49,999 people), and (iii) rural (< 2,500 people).
	Population density	County	US Census data on population density (people per square mile).
	Percentage of *county residents* who are (i) urbanized, (ii) in urban clusters, and (iii) rural	County	US Census data on the percentage of county residents who reside in (i) urbanized areas (50,000 or more people), (ii) urban clusters (2,500 to 49,999 people), and (iii) rural settings (< 2,500 people).
Demographic	Age [62]	Individual searcher	Classification of searcher age into discrete age buckets based on proprietary Bing classifier.
	Gender [63]	Individual searcher	Classification of searcher gender based on proprietary Bing classifier.
	Race or ethnicity [64,65,66]	County	US Census data on distributions of racial/ethnic background of county residents.
Socioeconomic [64,66,67,68]	Poverty	County	US Census data on the percentage of county residents below the poverty line.
	Educational attainment	County	US Census data on fraction of county residents aged 25 or older who have (i) high school graduate or more, (ii) some college or more, (iii) associate's degree of more, (iv) bachelor's degree or more, and (v) advanced degree.
	Unemployment	County	US Census data on fraction of county residents who are unemployed.
	Occupation	Individual searcher	Searches for "steel mill jobs", "electrician jobs", "shipbuilding jobs", etc.
Habits and Activities	Smoking [41]	County	US Census data on fraction of county residents who have ever smoked.
		County	US Census data on fraction of county residents who currently smoke.
		Individual searcher	Searches for "smoker", "cigar", "cigarette", "tobacco", "marlboro", "pall mall", "smokers lounge", "smokers room", etc.
		Individual searcher	Smoking cessation: Searches for "nicotine patch", "nicotine gum", "nicorette", "nicotine replacement therapy", etc.
		Individual searcher	Marijuana smoking: Searches for "cannabis", "marijuana", "smoking pot", etc.
	Nutritional supplements [52]	Individual searcher	Searches for "beta-carotene supplements", "solatene", "lumitene", etc.
	Talcum powder [69]	Individual searcher	Searches for "talcum powder", "baby powder", "sclerosol", etc.
	Indoor air pollution [70]	Individual searcher	Searches for "coal fire", "coal burning stove", "wood burning stove", etc.
	Flights [71,72]	Individual searcher	Flight statistics (number of flights, average flight duration, nature (domestic or international)) over the preceding one-year period. Derived from inferences on inter-query times and geographic location data from long-term activity on Bing.
Medical	HIV [73,74]	County	CDC data on human immunodeficiency virus (HIV) incidence rates per county.
		Individual searcher	Searches for "hiv", "aids", "human immunodeficiency virus", etc.
	Diabetes [75]	County	CDC data on diabetes incidence rates per county.
		Individual searcher	Searches for "diabetes", "diabetic", etc.
	Lung disease [76,77,78,79]	Individual searcher	Searches for "copd" "emphysema" "pneumonia" "silicosis" "tuberculosis", "lung fibrosis", etc.
	Autoimmune disease [80]	Individual searcher	Searches for "organ transplant", "autoimmune diseases", "lupus", etc.

	Diagnostic and therapeutic radiation [81,82]	Individual searcher	Searches for “chest radiation”, “low dose ct”, “radiotherapy”, etc.
	Oophorectomy [83]	Individual searcher	Searches for “oophorectomy”, “ovaries removed”, etc.
	Family history [84]	Individual searcher	Searches for “father lung cancer”, “brother diagnosed with lung cancer”, etc.

2.5 Early Detection

We frame early detection as a binary classification challenge using a statistical classifier. We train the classifier on evidential features from query timelines of the positive and negative cases. Given concerns about false positives and the large class imbalance (about five thousand positives vs. millions of negatives), we examine abilities to capture positive cases at very low false-positive rates.

The set of observations or features extracted from the time prior to *E* are grouped into two categories: (i) characteristics of each symptom searched, including their presence/absence and their temporal dynamics (e.g., increasing/decreasing over time, rate of change) (*Symptom*), and (ii) risk factors, including presence in queries (*Risk Factors*). All per-searcher feature computations are normalized by the sum of all queries from the searcher. For each per-county feature (e.g., lung carcinoma incidence rates) we determine its value based on searcher geolocation (reverse IP lookup) for each query, and calculate the mean, median, mode, maximum, and minimum feature value for that searcher across all of their queries.

The learned statistical model is based on gradient boosted trees [85]. Regularization methods were used to minimize overfitting. The statistical classifier lets us study how effectively we identify searchers who would later make experiential diagnostic queries for lung carcinoma. To characterize the predictive power, we report the area under the receiver operating characteristic curve (AUROC) and recall (TPR, true positive rate) at a set of low false positive rates (FPRs). Model generalizability is assessed using ten-fold cross validation, stratified by searcher.

3. Results

3.1 General Performance

We report performance in terms of overall AUROC and TPR at different FPR thresholds, ranging from FPR=0.00001 (1 error in 100,000 cases) to FPR=0.1 (1 error in 10 cases). We perform the predictions using data up to *E*–1 week to reduce the likelihood of experiential signals leaking into feature generation and training. Model performance (*Overall*) is strong, with AUROC=0.9535, and TPRs range from 3-32% for FPRs from 0.00001 to 0.001.

We experimented with adding a requirement that experiential diagnostic searchers (positives) also searched for either treatments or medications following *E*. This reduced the number of positives to 1,905 (35% of the 5,443 total), with the remainder added to the negative set. Significant performance

improvements are observed over the model trained and tested on all searchers (AUROC=0.9664, $TPR@FPR_{0.00001}$=5.039%, Z=13.421, $p < 0.0001$ [86]). The scoped positive set may also be more reliable given the need to observe follow-up treatment and medication searching.

3.2 Varying Lead Times

We sought to understand the impact of increasing the lead time on the performance of the algorithm. We backtrack to a period 52 weeks before *E* and target the 1,629 positives and 57,583 negatives observed 52 weeks in our dataset at that time. We then track these searchers forward in time, recording early detection performance at 13-week (three-month) intervals (Table 3).

Table 3. Performance at early prediction task at four-week intervals for the set of searchers for whom features can be computed from E – 1 week to E – 52 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 E–1 week is indicated as * $p < 0.01$, ** $p < 0.001$, and *** $p < 0.0001$. Weeks denotes the lead time prior to first experiential diagnostic query, when the prediction is made.

Weeks before *E*	**TPR (as %) at FPRs ranging from 0.00001–0.1**					**AUROC**
	0.00001	**0.0001**	**0.001**	**0.01**	**0.1**	
1 week	3.315	18.907	57.397	73.542	86.433	0.9415
13 weeks (~3 months)	2.947	16.943	52.977	67.526	85.267	0.9311
26 weeks (~6 months)	2.333*	14.549	49.110*	63.781*	83.917*	0.9120
39 weeks (~9 months)	1.842**	12.277*	44.260**	57.950*	75.752**	0.8891*
52 weeks (12 months)	1.473**	10.068**	39.288**	52.363**	69.613***	0.8662**

Results show that methods perform effectively up to one year before the experiential query. A one-year period is significant for early detection. Left untreated, the mean average survival time for stage I lung carcinoma is approximately one year [87,88,89,90].

3.3 Feature Analysis

We also pursued an understanding of the diagnostic power of features capturing terms about symptoms and those representing risk factors. Table 4 presents features ranked based on their absolute weights, capturing a measure of discrimination. We computed correlations with training data to understand the directionality of each feature.

Table 4. Top 30 features ranked by evidential weight, scored relative to the top-ranked feature (fraction of searches related to bronchitis), and ranked in terms of absolute weight of the feature. Weights are normalized with respect to the top-ranked feature. Direction is based on the correlation between the feature values and labels in the training data. Primary location is defined as the county in which the searcher issues the majority of their queries.

Feature	Weight	Direction	Class
Fraction of searches related to bronchitis	1.0000	Positive	Symptom
Fraction of searches related to hoarseness	0.8632	Positive	Symptom
Fraction of searches related to smoking	0.7840	Positive	Habits and Activities
Fraction of searches related to coughing	0.6492	Positive	Symptom
Probability that searcher age is 50–85 years	0.6307	Positive	Demographic
Fraction of searches related to chest pain	0.6062	Positive	Symptom
Fraction of searches related to smoking cessation	0.6033	Positive	Habits and Activities
Probability that user searcher age is 18–24 years	0.5941	Negative	Demographic
Fraction of searches related to shortness of breath	0.5721	Positive	Symptom
Fraction of searches related to tiredness	0.5581	Positive	Symptom
Fraction of searches related to wheezing	0.5371	Positive	Symptom
Annual lung cancer mortality rate in primary county	0.5371	Positive	Incidence and Mortality
Fraction of searches related to lung disease	0.5370	Positive	Medical
Radon levels in searcher's primary county	0.5345	Positive	Environmental
Fraction of searches related to family history of lung carcinoma	0.5344	Positive	Medical
Probability that searcher is male	0.4501	Positive	Demographic
Fraction of black or African American residents in primary county	0.4108	Positive	Demographic
US state associated with searcher's primary location	0.3925	Negative	Geographic
Average levels of air pollution in primary county	0.3178	Positive	Environmental
Fraction of searches related to weakness	0.2878	Positive	Symptom
Poverty rate in searcher's primary county	0.2736	Positive	Socioeconomic
Number of flights searcher presumed to take in one year prior	0.2404	Positive	Habits and Activities
Fraction of searches related to diagnostic / therapeutic radiation	0.2350	Positive	Medical
Searches for coughing up blood trending upwards over time	0.2174	Positive	Symptom
Fraction of residents in primary county who smoke	0.2050	Positive	Habits and Activities
Average wind speed in primary county	0.1895	Negative	Environmental
Smoking ban in effect in public places in primary county	0.1806	Negative	Environmental
Greenhouse gas emissions in primary county	0.1730	Positive	Environmental
Fraction of houses in primary location built ≤ 1939	0.1590	Positive	Environmental
Fraction of searches related to HIV	0.1368	Positive	Medical

The searched symptoms emerge as important, especially those related to bronchitis and coughing. We also find that sets of risk factors provide valuable signals. Discriminatory risk factors include expected factors, some of which may be proxies for smoking behavior, such as the likelihood that the searcher is male or below the poverty line [68], and less expected factors, such as the number of older homes in searchers' locations (homes which may lack radon mitigation or have old pipes) and the correlation with increased numbers of inferred trips on airplanes (inferred by frequency of large shifts in location in short time periods).

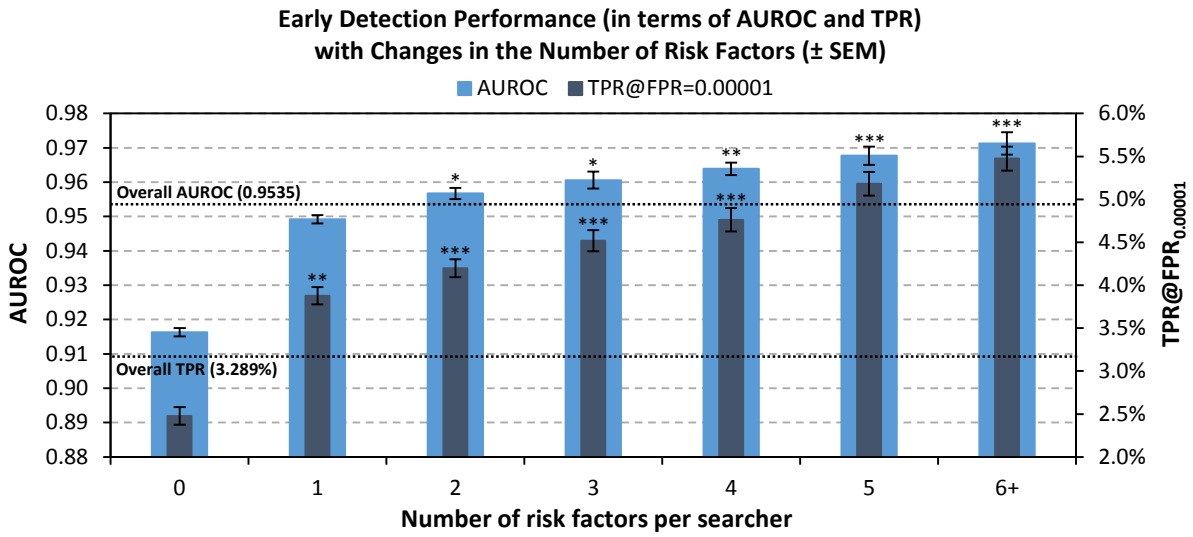
Since symptom terms were so important, we re-ran the training with symptom features only, yielding a reduction in early detection performance (AUROC=0.9237, $TPR@FPR_{0.00001}$=2.138%,

Z=26.223, $p < 0.0001$). The symptoms of lung carcinoma, such as coughing or chest pain, can be evidence of a range of medical conditions. It is not until the condition becomes more advanced that highly concerning symptoms (such as coughing up blood) emerge, and even then it is not clear that people will turn to search engines versus seeking professional medical advice [25].

3.4 Risk Factors

Table 2 shows the large number of risk factors under consideration. The five risk factors with the largest *relative risk* (RR) of experiential diagnostic searches over expected were: 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). Smoking (RR=1.646, 95% CI=1.032–2.260) was an important risk factor but not top-ranked despite strong connections with lung carcinoma incidence [41,47]. It is challenging to identify smokers in Web search logs; only 2.22% of searchers in our dataset queried for smoking-related materials per our definition (while recent statistics show that 16.8% of U.S. adults currently smoke [91]).

Results in the previous section suggest that risk factors are important for early detection. We were interested in how the model performed for searchers at highest risk, i.e., those affected by most risk factors. Whether the risk factor applied to a searcher was based on the relationship between the searcher's feature value and the average, and whether it was significantly ($p < 0.01$) above or below varied based on the specific feature (under "Data" in Table 2) and its association with *heightened risk* (e.g., for incidence rates, higher values indicate increased risk, whereas for smoking bans, lower values indicate increased risk). We re-ran the detection task at E−1 week for searchers with different numbers of significant risk factors, ranging from zero (searchers only present evidence of symptoms) to six or more risk factors. Figure 2 presents the results, including the number of positive and negatives, the AUROC, and the TPR at FPR=0.00001.



# risk factors	0	1	2	3	4	5	6+	≥ 0 (All cases)
# pos cases	1,251	1,496	1,072	664	420	193	347	5,443
# neg cases	1,153,900	1,254,239	844,987	620,332	405,045	174,244	355,795	4,808,542
% pos cases	0.108%	0.119%	0.127%	0.107%	0.104%	0.111%	0.097%	0.113%

Figure 2. Plots of the AUROC and TPR@$FPR_{0.00001}$ for different numbers of risk factors per searcher. 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 (TPR@$FPR_{0.00001}$=3.289%). Error bars denote standard error of the mean (± SEM). Also shown are number of searchers in each group and number and percentage of each group that is a positive case. Significance of differences in AUROC and TPR versus *Overall* model computed using independent measures *t*-tests are denoted using * $p < 0.01$, ** $p < 0.001$, and *** $p < 0.0001$.

At least one risk factor significantly improves detection performance over the symptom-only model, and there are consistent but smaller marginal gains for each additional factor. The model performs best for those at particularly high risk. These results can provide general insights about screening policies for lung carcinoma.

4. Discussion

The feasibility study highlights the potential value of employing search logs in pre-screening or screening for lung carcinoma. Search engines are a pervasive technology that can help people understand illness and to alert them or their physicians about the value of screening. The strong performance of the statistical classifier shows how methods based on logs might complement more traditional screening methods connected with high false positive rates [9,11] and that rely on access to data (e.g., sputum [24,92]) that is

unavailable from large populations. Search log analysis can be coupled with engagement methods that provide a useful alternative or complement to existing screening methods. The early detection model performs best for those at greatest risk, a finding that can guide the selective application of the method. The model makes accurate predictions up to one year in advance of the experiential diagnostic query (e.g., detecting 10% of positives while being incorrect 1 in 10,000 times). The decision threshold can be adjusted depending on the desired model operating characteristics (e.g., raise threshold to obtain higher-precision predictions, but lower recall).

We acknowledge several limitations. First, we lack access to ground truth data about whether searchers were actually diagnosed; we employ a similar method to prior work [6] leveraging the identification of experiential searchers. Experiential queries match well with both incidence rates and expected patterns of searching for treatments and medications. We found that performance improves further with additional diagnostic evidence of treatment or medication searching following the first experiential diagnostic query. On another limitation, we require searchers to issue at least one symptom query to be included in our sample. This was motivated by the envisaged application of these models in search engines to alert searchers and/or physicians when a positive prediction is made. An alternative would be to ignore the symptoms and look for risk factors in all searchers. This may boost coverage, but would also increase false positives dramatically, given the massive imbalance of positives and negatives in the open world of online search. For example, re-running model training with risk factors only reduced performance versus *Overall* (AUROC=0.7249, TPR@$FPR_{0.00001}$=0.772%, Z=79.404, $p < 0.0001$) compared to the symptom-only model or *Overall*. Symptom-related features are important in our early-detection model (Table 4). Finally, the extent of the alignment between dates associated with experiential diagnostic queries and the dates of the actual diagnosis needs to be determined via additional studies.

Of particular interest were a set of risk factors spanning different classes. These factors can provide new epidemiological insights about lung carcinoma. The identified factors can also guide both the selective application of the statistical models and future refinements of the methods. The broad range of risk factors tested highlights the flexibility in this type of analysis, and the importance of considering risk factors. We see opportunities ahead in refining the accuracy of signals pertaining to risk factors, e.g., using GPS traces of near-exact location data from mobile devices to model time spent in specific venues. Some risk factors may be best obtained from searchers directly (e.g., via active engagement and dialog during search). Three examples are searcher age, searcher smoking habits, and family history of lung carcinoma. Inferences about these factors are shown in Section 3.4 to have strong discriminative value. Prediction performance versus *Overall* drops if these features are not obtained (AUROC=0.9073, TPR@$FPR_{0.00001}$=2.021%, Z=35.584, $p < 0.0001$), but there remains sufficient information in the remaining features to make reliable predictions.

In a real-world deployment, search engines could serve as an initial filter to identify patients who are candidates for clinical screening. Health-conscious patients may volunteer to receive health alerts of this nature if concerning activity is detected. Our focus has been on performance at low false positive rates and we did not consider false negatives directly. Real-world systems would need to communicate to searchers some illness instances may be missed. More broadly, there is the associated issue of communicating early detection outcomes with searchers without causing unnecessary alarm and associated costs, and whether such communication is necessary when outcomes could be passed directly to physicians for consideration and patient follow-up.

References

1. Siegel, R.L., Miller, K.D., and Jemal, A. (2015). Cancer statistics, 2015. *CA: A Cancer Journal for Clinicians*, 65(1): 5–29.
2. Henschke, C.I., Yankelevitz, D.F., Libby, D.M., et al. (2006). Survival of patients with stage I lung cancer detected on CT screening. *New England Journal of Medicine*, 355(17): 1763–1771.
3. Ihde, D.C. (1992). Chemotherapy of lung cancer. *New England Journal of Medicine*, 327(20): 1434–1441.
4. Flehinger, B.J., Kimmel, M., and Melamed, M.R. (1992). The effect of surgical treatment on survival from early lung cancer. Implications for screening. *Chest*, 101(4): 1013–1018.
5. Sobue, T., Suzuki, R., Matsuda, M., Kuroishi, T., Ikeda, S., and Naruke, T. (1992). Survival for clinical stage I lung cancer not surgically treated: comparison between screen-detected and symptom-detected cases. *Lung Cancer*, 8(3): 237.
6. Paparrizos, J., White, R.W., and Horvitz, E. (2016). Screening for pancreatic adenocarcinoma using signals from web search logs: Feasibility study and results. *Journal of Oncology Practice*, in press.
7. Mulshine, J.L. and Henschke, C.I. (2000). Prospects for lung-cancer screening. *Lancet*, 355(9204): 592–593.
8. Naidich, D.P., Marshall, C.H., Gribbin, C., Arams, R.S., and McCauley, D.I. (1990). Low-dose CT of the lungs: preliminary observations. *Radiology*, 175(3): 729–731.
9. Bach, P.B., Mirkin, J.N., Oliver, T.K., et al. (2012). Benefits and harms of CT screening for lung cancer: A systematic review. *Journal of the American Medical Association*, 307(22): 2418–2429.
10. Aberle, D.R., Abtin, F., and Brown, K. (2013). Computed tomography screening for lung cancer: has it finally arrived? Implications of the national lung screening trial. *Journal of Clinical Oncology*, 31(8): 1002–1008.
11. National Lung Screening Trial Research Team. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5): 395.
12. Marcus, P.M., Bergstralh, E.J., Fagerstrom, R.M., et al. (2000). Lung cancer mortality in the Mayo Lung Project: Impact of extended follow-up. *J. Natl Cancer Inst.* 92: 1308–1316.
13. Crestanello, J.A., Allen, M.S., Jett, J., et al. (2004). Thoracic surgical operations in patients enrolled in a computed tomographic screening trial. *Journal of Thoracic and Cardiovascular Surgery*, 128(2): 254–259.
14. Patz, E.F., Pinsky, P., Gatsonis, C., et al. (2014). Overdiagnosis in low-dose computed tomography screening for lung cancer. *JAMA Internal Medicine*, 174(2): 269–274.

15. Croswell, J.M., Baker, S.G., Marcus, P.M., Clapp, J.D., and Kramer, B.S. (2010). Cumulative incidence of false-positive test results in lung cancer screening: a randomized trial. *Annals of Internal Medicine*, 152(8): 505–512.
16. Fontana, R.S., Sanderson, D.R., Taylor, W.F., et al. (1984). Early lung cancer detection: results of the initial (prevalence) radiologic and cytologic screening in the Mayo Clinic study. *American Review of Respiratory Disease*, 130(4): 561–565.
17. Kubik, A., Parkin, D.M., Khlat, M., Erban, J., Polak, J., and Adfamec, M. (1990). Lack of benefit from semi-annual screening for cancer of the lung: follow-up report of a randomized controlled trial on a population of high-risk males in Czechoslovakia. *International Journal of Cancer*, 45: 26-33.
18. Doria-Rose, V.P. and Szabo, E. (2010). Screening and prevention of lung cancer. In: Kernstine KH, Reckamp KL, editors. *Lung cancer: A Multidisciplinary Approach to Diagnosis and Management*. New York: Demos Medical Publishing, 53–72.
19. Oken, M.M., Hocking, W.G., Kvale, P.A., et al. (2011). Screening by chest radiograph and lung cancer mortality: the prostate, lung, colorectal, and ovarian (PLCO) randomized trial. *Journal of the American Medical Association*, 306: 1865–1873.
20. Manser, R., Lethaby., A., Irving., L.B., et al. (2013). Screening for lung cancer. *Cochrane Database of Systematic Reviews* 6(6): CD001991.
21. Swensen, S.J., Jett, J.R., Sloan, J.A., et al. (2002). Screening for lung cancer with low-dose spiral computed tomography. *American Journal of Respiratory and Critical Care Medicine*, 165(4): 508–513.
22. Mao, L., Hruban, R.H., Boyle, J.O., Tockman, M., and Sidransky, D. (1994). Detection of oncogene mutations in sputum precedes diagnosis of lung cancer. *Cancer Research*, 54(7): 1634–1637.
23. Belinsky, S.A., Nikula, K.J., Palmisano, W.A., et al. (1998). Aberrant methylation of p16INK4a is an early event in lung cancer and a potential biomarker for early diagnosis. *Proceedings of the National Academy of Sciences*, 95(20): 11891–11896.
24. Palmisano, W.A., Divine, K.K., Saccomanno, G., et al. (2000). Predicting lung cancer by detecting aberrant promoter methylation in sputum. *Cancer Research*, 60(21): 5954–5958.
25. Fox, S. and Duggan, M. (2013). *Health Online 2013*. http://www.pewinternet.org/2013/01/15/health-online-2013 [Accessed June 1, 2016].
26. Bader, J.L. and Theofanos, M.F. (2003). Searching for cancer information on the internet: analyzing natural language search queries, *Journal of Medical Internet Research*, 5(4).
27. Castleton, K., Fong., T., Wang-Gillam, A., et al. (2011). A survey of internet utilization among patients with cancer. *Supportive Care in Cancer*, 19(8): 1183–1190.

28. Helft, P.R. (2011). Patients with cancer, internet information, and the clinical encounter: a taxonomy of patient users, *American Society of Clinical Oncology Educational Book*, e89–92.
29. Ofran, Y., Paltiel, O., Pelleg, D., Rowe, J.M., and Yom-Tov, E. (2012). Patterns of information-seeking for cancer on the internet: an analysis of real world data. *PLoS ONE*, 7(9): e45921.
30. Paul, M.J., White, R.W., and Horvitz, E. (2014). Search and breast cancer: on disruptive shifts of attention over life histories of an illness. *ACM Transactions on the Web*, 10(2): 13.
31. White, R.W. and Horvitz, E. (2009). Cyberchondria: studies of the escalation of medical concerns in web search. *ACM Transactions on Information Systems*, 27(4): 23.
32. Lauckner, C. and Hsieh, G. (2013). 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.
33. White, R.W. and Horvitz, E. (2014). 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.
34. White, R.W. and Drucker, S.M. (2007). Investigating behavioral variability in web search. *Proc. World Wide Web Conference*, 21–30.
35. Lau, T. and Horvitz, E. (1999). Patterns of search: analyzing and modeling web query refinement, *Proc. User Modeling Conference*, 119–128.
36. Downey, D., Dumais, S.T., and Horvitz, E. (2007). Models of searching and browsing: languages, studies, and applications. *Proc. International Joint Conference on Artificial Intelligence*, 2740–2747.
37. Dupret, G. and Piwowarski, B. (2008). A user browsing model to predict search engine click data from past observations. *Proc. ACM Conference on Human Factors in Computing Systems*, 331–338.
38. Joachims, T. (2002). Optimizing search engines using clickthrough data. *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 133–142.
39. Tan, B., Shen, X., and Zhai, C. (2006). Mining long-term search history to improve search accuracy. *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 718–723.
40. Richardson, M. (2009). Learning about the world from long-term query logs. *ACM Transactions on the Web*, 2(4): 21.
41. Biesalski, H.K., Bueno de Mesquita, B., Chesson, A., et al. (1998). European Consensus Statement on Lung Cancer: risk factors and prevention. Lung Cancer Panel. *CA: A Cancer Journal for Clinicians*, 48(3): 167–176.
42. Thun, M.J., Hannan, L.M., Adams-Campbell, L.L., et al. (2008). Lung cancer occurrence in never-smokers: an analysis of 13 cohorts and 22 cancer registry studies. *PLoS Medicine*, 5(9): e185.

43. Schmid, K., Kuwert, T., and Drexler, H. (2010). Radon in indoor spaces: an underestimated risk factor for lung cancer in environmental medicine. *Deutsches Ärzteblatt International*, 107(11): 181–186.
44. Choi, H. and Mazzone, P. (2014). Radon and lung cancer: assessing and mitigating the risk. *Cleveland Clinic Journal of Medicine*, 81(9): 567–575.
45. Doll, R. (1993). Mortality from lung cancer in asbestos workers 1955. *British Journal of Industrial Medicine*, 50(6): 485.
46. Hodgson, J.T. and Darnton, A. (2000). The quantitative risks of mesothelioma and lung cancer in relation to asbestos exposure. *Annals of Occupational Hygiene*, 44(8), 565-601.
47. Taylor, R., Najafi, F., and Dobson, A. (2007). Meta-analysis of studies of passive smoking and lung cancer: effects of study type and continent. *International Journal of Epidemiology* 36(5): 1048–1059.
48. Wakelee, H.A., Chang, E.T., Gomez, S.L., et al. (2007). Lung cancer incidence in never smokers. *Journal of Clinical Oncology*, 25(5): 472–478.
49. Chen, H., Goldberg, M.S., and Villeneuve, P.J. (2008). A systematic review of the relation between long-term exposure to ambient air pollution and chronic diseases. *Reviews on Environmental Health*, 23(4): 243–297.
50. Pope III, C.A., Burnett, R.T., Thun, M.J., et al. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Journal of the American Medical Association*, 287(9): 1132–1141.
51. Nyberg, F., Gustavsson, P., Järup, L., et al. (2000). Urban air pollution and lung cancer in Stockholm. *Epidemiology*, 11(5): 487–495.
52. Omenn, G.S., Goodman, G.E., Thornquist, M.D., et al. (1996). Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease. *New England Journal of Medicine*, 334(18): 1150–1155.
53. Darby, S., Hill, D., Auvinen, A., et al. (2005). Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case-control studies. *British Medical Journal*, 330(7485): 223.
54. Krewski, D., Lubin, J.H., Zielinski, J.M., et al. (2005). Residential radon and risk of lung cancer: a combined analysis of 7 North American case-control studies. *Epidemiology*, 16(2): 137–145.
55. Chen, C.J., Chen, C.W., Wu, M.M., and Kuo, T.L. (1992). Cancer potential in liver, lung, bladder and kidney due to ingested inorganic arsenic in drinking water. *British Journal of Cancer*, 66(5): 888.
56. Smith, A.H., Goycolea, M., Haque, R., and Biggs, M.L. (1998). Marked increase in bladder and lung cancer mortality in a region of Northern Chile due to arsenic in drinking water. *American Journal of Epidemiology*, 147(7): 660–669.

57. Wolf, R. and Wolf, D. (2004). Increased incidence of cancer near a cell-phone transmitter station. *International Journal of Cancer*, 1(2): 123–128.
58. Verkasalo, P.K., Pukkala, E., Hongisto, M.Y., et al. (1993). Risk of cancer in Finnish children living close to power lines. *British Medical Journal*, 307(6909): 895–899.
59. Feychting, M. and Alhbom, M. (1993). Magnetic fields and cancer in children residing near Swedish high-voltage power lines. *American Journal of Epidemiology*, 138(7): 467–481.
60. Spix, C., Schmiedel, S., Kaatsch, P., Schulze-Rath, R., and Blettner, M. (2008). Case–control study on childhood cancer in the vicinity of nuclear power plants in Germany 1980–2003. *European Journal of Cancer*, 44(2): 275–284.
61. Campbell, N.C., Elliott, A.M., Sharp, L., Ritchie, L.D., Cassidy, J., and Little, J. (2001). Rural and urban differences in stage at diagnosis of colorectal and lung cancers. *British Journal of Cancer*, 84(7): 910.
62. Hart, C.L., Hole, D.J., Gillis, C.R., Smith, G.D., Watt, G.C., and Hawthorne, V.M. (2001). Social class differences in lung cancer mortality: risk factor explanations using two Scottish cohort studies. *International Journal of Epidemiology*, 30(2): 268–274.
63. Zang, E.A. and Wynder, E.L. (1996). Differences in lung cancer risk between men and women: examination of the evidence. *Journal of the National Cancer Institute*, 88(3-4), 183–192.
64. Devesa, S.S. and Diamond, E.L. (1983). Socioeconomic and racial differences in lung cancer incidence. *American Journal of Epidemiology*, 118(6): 818–831.
65. Fukuoka, M., Yano, S., Giaccone, G., et al. (2003). Multi-institutional randomized phase II trial of gefitinib for previously treated patients with advanced non–small-cell lung cancer. *Journal of Clinical Oncology*, 21(12): 2237–2246.
66. Ward, E., Jemal, A., Cokkinides, V., et al. (2004). Cancer disparities by race/ethnicity and socioeconomic status. *CA: A Cancer Journal for Clinicians*, 54(2), 78–93.
67. Mao, Y., Hu, J., Ugnat, A.M., Semenciw, R., Fincham, S., and Canadian Cancer Registries Epidemiology Research Group. (2001). Socioeconomic status and lung cancer risk in Canada. *International Journal of Epidemiology*, 30(4): 809–817.
68. Centers for Disease Control and Prevention. Current Cigarette Smoking Among Adults—United States, 2005–2014. Morbidity and Mortality Weekly Report 2015;64(44):1233–40 [accessed May 24 2016].
69. Thomas, T.L. (1989). Lung cancer mortality among pottery workers in the United States. *IARC scientific publications*, 97: 75–81.
70. Mumford, J.L., He, X.Z., Chapman, R.S., et al. (1987). Lung cancer and indoor air pollution in Xuan Wei, China. *Science*, 235(4785): 217–220.

71. Band, P.R., Le, N.D., Fang, R., et al. (1996). Cohort study of Air Canada pilots: mortality, cancer incidence, and leukemia risk. *American Journal of Epidemiology*, 143(2): 137–143.
72. Rafnsson, V., Hrafnkelsson, J., and Tulinius, H. (2000). Incidence of cancer among commercial airline pilots. *Occupational and Environmental Medicine*, 57(3): 175–179.
73. Kirk, G. D., Merlo, C., O'Driscoll, P., et al. (2007). HIV infection is associated with an increased risk for lung cancer, independent of smoking. *Clinical Infectious Diseases*, 45(1): 103–110.
74. Engels, E.A., Brock, M.V., Chen, J., Hooker, C.M., Gillison, M., and Moore, R.D. (2006). Elevated incidence of lung cancer among HIV-infected individuals. *Journal of Clinical Oncology*, 24(9): 1383–1388.
75. Steenland, K., Nowlin, S., and Palu, S. (1995). Cancer incidence in the National Health and Nutrition Survey I. Follow-up data: diabetes, cholesterol, pulse and physical activity. *Cancer Epidemiology Biomarkers and Prevention*, 4(8): 807–811.
76. Meyer, E.C. and Liebow, A.A. (1965). Relationship of interstitial pneumonia honeycombing and atypical epithelial proliferation to cancer of the lung. *Cancer*, 18(3): 322–351.
77. McDonald, J.C. (1989). Silica, silicosis, and lung cancer. *British Journal of Industrial Medicine*, 46(5): 289.
78. Young, R.P., Hopkins, R.J., Christmas, T., Black, P.N., Metcalf, P., and Gamble, G.D. (2009). COPD prevalence is increased in lung cancer, independent of age, sex and smoking history. *European Respiratory Journal*, 34(2): 380–386.
79. Wilson, D.O., Weissfeld, J.L., Balkan, A., et al. (2008). Association of radiographic emphysema and airflow obstruction with lung cancer. *American Journal of Respiratory and Critical Care Medicine*, 178(7): 738–744.
80. Hemminki, K., Liu, X., Ji, J., Sundquist, J., and Sundquist, K. (2012). Effect of autoimmune diseases on risk and survival in histology-specific lung cancer. *European Respiratory Journal*, 40(6): 1489–1495.
81. Hall, E.J. and Brenner, D.J. (2014). Cancer risks from diagnostic radiology. *British Journal of Radiology*, 81(965): 362–378.
82. de Gonzalez, A.B. and Darby, S. (2004). Risk of cancer from diagnostic X-rays: estimates for the UK and 14 other countries. *The Lancet*, 363(9406): 345–351.
83. Parker, W.H., Broder, M.S., Chang, E., et al. (2009). Ovarian conservation at the time of hysterectomy and long-term health outcomes in the nurses' health study. *Obstetrics and Gynecology*, 113(5): 1027.
84. Matakidou, A., Eisen, T., and Houlston, R.S. (2005). Systematic review of the relationship between family history and lung cancer risk. *British Journal of Cancer*, 93(7): 825-833.

85. Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189–1232.
86. Hanley, J.A. and McNeil, B.J. (1982). The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology*, 143(1): 29–36.
87. Kyasa, M.J. and Jazieh, A.R. (2002). Characteristics and outcomes of patients with unresected early-stage non-small cell lung cancer. *Southern Medical Journal*, 95(10): 1149–1153.
88. McGarry, R., Song, G., Rosiers, P., and Timmerman, R. (2002). Observation only management of early stage medically inoperable lung cancer poor outcome. *Chest*, 121(4): 1155–1158.
89. Chadha, A.S., Ganti, A.K., Sohi, J.S., Sahmoun, A.E., and Mehdi, S.A. (2005). Survival in untreated early stage non-small cell lung cancer. *Anticancer Research*, 25(5): 3517–3520.
90. Wao, H., Mhaskar, R., Kumar, A., Miladinovic, B., and Djulbegovic, B. (2013). Survival of patients with non-small cell lung cancer without treatment: A systematic review and meta-analysis. *Systematic Reviews*, 2(1): 1.
91. Centers for Disease Control and Prevention. Current Cigarette Smoking Among Adults—United States, 2005–2014. *Morbidity and Mortality Weekly Report 2015*, 64(44): 1233–1240.
92. Saccomanno, G., Archer, V., Auerbach, O., Saunders, R.P., and Brennan L.M. (1974). Development of carcinoma of the lung as reflected in exfoliated cells. *Cancer*, 33(1): 256–270.