

# Will That Pulmonary Nodule Become Cancerous? A Risk Prediction Model for Incident Lung Cancer



Barbara Nemesure<sup>1</sup>, Sean Clouston<sup>1,2</sup>, Denise Albano<sup>3</sup>, Stephen Kuperberg<sup>4</sup>, and Thomas V. Bilfinger<sup>3</sup>

## Abstract

This prospective investigation derived a prediction model for identifying risk of incident lung cancer among patients with visible lung nodules identified on computed tomography (CT). Among 2,924 eligible patients referred for evaluation of a pulmonary nodule to the Stony Brook Lung Cancer Evaluation Center between January 1, 2002 and December 31, 2015, 171 developed incident lung cancer during the observation period. Cox proportional hazard models were used to model time until disease onset. The sample was randomly divided into discovery ( $n = 1,469$ ) and replication ( $n = 1,455$ ) samples. In the replication sample, concordance was computed to indicate predictive accuracy and risk scores were calculated using the linear predictions. Youden index was used to identify high-risk versus low-risk patients and cumulative lung cancer incidence was examined for high-risk and

low-risk groups. Multivariable analyses identified a combination of clinical and radiologic predictors for incident lung cancer including ln-age, ln-pack-years smoking, a history of cancer, chronic obstructive pulmonary disease, and several radiologic markers including spiculation, ground glass opacity, and nodule size. The final model reliably detected patients who developed lung cancer in the replication sample ( $C = 0.86$ , sensitivity/specificity = 0.73/0.81). Cumulative incidence of lung cancer was elevated in high-risk versus low-risk groups [HR = 14.34; 95% confidence interval (CI), 8.17–25.18]. Quantification of reliable risk scores has high clinical utility, enabling physicians to better stratify treatment protocols to manage patient care. The final model is among the first tools developed to predict incident lung cancer in patients presenting with a concerning pulmonary nodule.

## Background

Lung cancer claims the lives of more than 150,000 people in the United States each year and, with an age-standardized mortality rate of 40.6/100,000, it is the most common reason for cancer death (1). An individual's risk of developing lung cancer is thought to be the result of a combination of environmental, genetic, and behavioral factors (2). Although risk factors for lung cancer are relatively well understood, discrimination tools to predict who will develop lung cancer and who will not are primarily based on case-control screening studies (3–6),

or clinic-based studies of patients being evaluated for existing cancers (7–9). With the advent of the recent United States Preventive Services Task Force (USPSTF) recommendations for lung cancer screening (10), several prospective studies have begun to develop risk prediction models for patients with a history of heavy tobacco use (11–13). However, there have been few large-scale, prospective studies evaluating the risk that identified nodules convert to cancer (14). To our knowledge, no studies to date have examined the extent to which clinical and imaging factors might prospectively identify incident lung cancers among individuals with suspicious lung nodules identified on CT. The present investigation aimed to construct a prediction model of incident lung cancer for patients presenting with a lung nodule.

<sup>1</sup>Department of Family, Population and Preventive Medicine, Stony Brook Medicine, Stony Brook, New York. <sup>2</sup>Program in Public Health, Stony Brook Medicine, Stony Brook, New York. <sup>3</sup>Department of Surgery, Stony Brook Medicine, Stony Brook, New York. <sup>4</sup>Department of Medicine, Stony Brook Medicine, Stony Brook, New York.

**Corresponding Author:** Barbara Nemesure, Stony Brook University, Stony Brook, NY 11794-8036. Phone: 631-444-1293; E-mail: Barbara.Nemesure@stonybrookmedicine.edu

Cancer Prev Res 2019;12:463–70

doi: 10.1158/1940-6207.CAPR-18-0500

©2019 American Association for Cancer Research.

## Materials and Methods

### Setting

The Stony Brook Cancer Center's Lung Cancer Evaluation Center (LCEC) provides evaluation, diagnosis, treatment, and surveillance for patients who are referred as a result of abnormal pulmonary findings discovered on CT.

For the purposes of this study, enrollment and diagnostic characteristics were retrieved from the LCEC patient database during the period January 1, 2002 to December 31, 2015. Incident cases were included if they occurred before December 31, 2016. Data obtained from the Stony Brook Cancer Registry were used to confirm lung cancer diagnoses. Date of diagnosis, cancer stage, and cancer type were recorded. Patient records were excluded if they had been referred to the LCEC but had not attended initial visits at time of data retrieval, or had been evaluated by the LCEC but had been determined to have an unrelated condition including, for example, metastatic disease and chronic obstructive pulmonary disease (COPD) management.

Among those who were eligible ( $N = 4,272$ ), 1,348 patients with a history of lung cancer or whose cancers were diagnosed within 6 months of initial visits were excluded. The remaining 2,924 patients provided 20,104.42 person-years at risk. Among the eligible patients, 171 developed lung cancer during the observation period (i.e., 6 months after the initial visit through December 31, 2016). Patients were censored at time of death. In patients with incident cancer, 33 of 171 (19.3%) were diagnosed at later stages III/IV, whereas records showed that 82.5% had non-small cell cancer, 2.3% had carcinoids, 1.8% had small-cell cancer, and 12.9% had other forms of cancer.

### Measures

Age, gender, smoking status (current or former smoker, years since quitting, and pack-years of smoking), personal and family history of lung or other cancers, COPD diagnosed using the GOLD criteria (15), self-reported asbestos exposure and lung cancer status (yes vs. no) were recorded. Pack-years smoked and age at enrollment were transformed using the natural log because it improved model fit. Imaging features derived from CT scanning included size in millimeters of the largest nodule, nodule count and location (upper lobe vs. other), opacity type (ground glass vs. other), and spiculation (any vs. none).

### Statistical analyses

The LCEC data were initially stratified into incident cases and those who were not diagnosed with cancer during the observational period. Student  $t$  tests and  $\chi^2$  tests were used to examine differences between patients who developed incident cancer and those determined to be cancer free. Multivariable Cox proportional hazards models were used to examine predictors of incident cancer (16). The Breslow method was used to handle ties and Schoenfeld residuals tested the proportional hazards assumption. Power analyses revealed that this sample was capable of detecting standardized effect sizes of  $\|d\| > 0.05$ .

The goal for the risk model was to identify risk factors for incidence of lung cancer. To improve predictive reliability,

the sample was randomly divided in half to form discovery and replication samples. Comparative analyses used  $\chi^2$  and  $t$  tests to examine differences between the 2 samples but found no statistically significant differences. In discovery analyses, Cox models were completed and standardized effect sizes ( $d$ ) were calculated to facilitate variable exclusion. Standardized effect sizes are useful in that they are sensitive to rare exposures, which may be inadequately powered in some studies but may be important in replication (17). Effect sizes exceeding  $\|d\| = 0.05$  were retained. Multivariable-adjusted HRs and 95% confidence intervals were computed and baseline survival and hazards were reported to facilitate risk score calculation. Beta coefficients were used to calculate linear predictions of risk scores from the fitted model. An optimal cutoff to differentiated high-risk versus low-risk groups was determined in the discovery sample using the Youden method (18); sensitivity, specificity, and Youden's index were also reported. Strata-specific incidence rates, 1-/5-year proportion with incident cancer, and years until 10% of the at-risk population developed lung cancer (19) were calculated in the replication sample. Risk and cumulative incidence of lung cancer above/below the optimal cutoff, determined in the discovery sample, were examined for comparison purposes in the replication sample. Concordance ( $C$ ) between model and outcome was assessed to measure model accuracy. Significance testing adjusted for the (false discovery rate (FDR) = 0.05; ref. 20). Analyses were performed using Stata 15/SE (StataCorp).

Sensitivity analyses examined differences in risk calculation when adjusting for differential mortality using fine-gray competing risks hazards models (21); sub-HRs (SHR) were reported. The extent to which results changed when excluding individuals who were lost to follow up prior to incident cancer diagnosis was also assessed. Additional analyses were conducted considering specific variables in clinical and radiological data. Sensitivity analyses examined the predictive power of the model to identify cancer subtypes including non-small cell ( $n = 141$ ) and other ( $n = 30$ ) cancers, and also examined differences in capacity for identification by cancer stage at diagnosis.

## Results

### Sample characteristics

Compared with patients with nodules who did not develop lung cancer during the observation period (Table 1), patients developing lung cancer were older and had increased ln-pack-years smoking history. Patients who developed cancer were more likely to report a family history of lung and other cancers, and a personal history of non-lung cancer, COPD, and asbestos exposure. CT images of patients developing cancer were more likely to

**Table 1.** Characteristics of incident lung cancer cases and those who remained cancer-free during the observation period, LCEC 2002 to 2016

Patient characteristic	Incident cancer diagnosis (N = 171)	No cancer yet (N = 2,753)
Female sex	54.4	51.7
Age in years, mean (SD)	68 (10.7)	60.9 (14.4) <sup>a</sup>
Ln pack-years, mean (SD)	2.7 (1.8)	0.7 (1.4) <sup>a</sup>
Number of lesions, mean (SD)	1.4 (0.8)	1.2 (0.5) <sup>a</sup>
Largest nodule size, mean (SD)	12.5 (11)	7.4 (10.4) <sup>a</sup>
Asbestos exposure	9.36	2.80 <sup>a</sup>
Family history of lung cancer	8.19	3.16 <sup>a</sup>
Family history of other cancer	18.13	8.50 <sup>a</sup>
Personal history of other cancer	21.05	6.39 <sup>a</sup>
Chronic obstructive pulmonary disorder	21.64	2.32 <sup>a</sup>
Ground glass opacity	8.19	1.85 <sup>a</sup>
Spiculated	17.54	2.80 <sup>a</sup>
Upper lobe	40.94	13.77 <sup>a</sup>

NOTE: All statistics report percentages (%) or, when noted, means (SD). Ln, transformed using the natural logarithm. Results for categorical variables used  $\chi^2$  tests whereas results for continuous variables (reporting means and SD above) used 2-tailed *t* tests.

<sup>a</sup>*P* < 0.001 when adjusting for the FDR.

have an increased number and larger sized nodules, ground glass opacities, spiculation, and nodes in the upper lobes.

### Discovery analyses

Model development efforts revealed a saturated model that fit the data well (*C* = 0.83). Excluding covariates with small effects in the discovery sample resulted in similar model fit (*C* = 0.86). Ln-age, Ln-pack-years smoking, a history of cancer, COPD, as well as radiologic markers of spiculation, ground glass opacity, and nodule size were retained.

### Replication analyses

The final model (shown in Table 2) was also able to discriminate risk of cancer in the replication sample (*C* = 0.86). Predictors included Ln-age, Ln-pack-years of smoking, personal history of other cancers, COPD, ground glass opacity, and spiculation (Table 2). Risk scores were calculated in both samples using the results provided by the replication sample model.

**Table 2.** Multivariable-adjusted hazard ratios and 95% CIs examining cumulative incidence of lung cancer using Cox proportional hazards modeling for variables that passed exclusion criteria in the replication sample, LCEC 2002 to 2016

Patient characteristic	Adjusted HR (95% CI)
Ln age in years <sup>a</sup>	8.79 (2.44–31.71) <sup>a</sup>
Ln pack-years of smoking	1.59 (1.39–1.82) <sup>a</sup>
Personal history of other cancer	2.31 (1.38–3.89) <sup>a</sup>
Chronic obstructive pulmonary disorder	3.15 (1.76–5.64) <sup>a</sup>
Ground glass opacity	2.45 (1.04–5.80) <sup>a</sup>
Spiculated	1.99 (1.08–3.69) <sup>a</sup>
Size of largest node	1.02 (1.00–1.04)

Abbreviation: Ln, transformed using the natural logarithm.

<sup>a</sup>Statistically significant associations after accounting for the FDR. Baseline cumulative hazard was 0.0134037; baseline survival was 0.9866759.

### Risk score characteristics

Risk scores calculated in the discovery sample (mean = 9.47; SD = 1.04; min = 5.89; median = 9.36; max = 13.25) were used to develop an optimal cutoff (risk score  $\geq 10.17$ ), which had a high degree of accuracy (specificity = 0.81; sensitivity = 0.73; Youden index = 0.54). Risk scores calculated using the final model in the replication sample (mean = 9.60; SD = 1.40; range = 5.89–15.00) did not differ, on average, from those in the discovery sample (*P* = 0.952). In the replication sample, risk scores were higher (*P* < 0.001) in incident cases (mean = 11.55; SD = 1.24; range = 8.83–13.64) than in cancer-free cases. In addition, 24.0% of patients in the replication sample were in the high-risk group whereas 81.0% of incident cancers occurred in patients with risk scores above the cutoff (specificity = 0.79; sensitivity = 0.81).

### Outcomes among those deemed to have high risk

Incidence was elevated in the high-risk versus low-risk group (Table 3). As evidenced by separation in lines over time, cumulative incidence curves showed that incidence increased in high-risk patients, whereas the steepness of the curves clarified that much of this increased risk was concentrated in the first 5 years (Fig. 1).

Sensitivity analyses accounting for selective mortality revealed that the risk score was predictive of incident lung cancer [SHR = 14.29 (8.17–24.98)]. Additional analyses examining the predictive value of the risk score only among individuals who were reported to have stayed at the LCEC throughout their observation period resulted in similar overall results [HR = 16.72 (9.18–30.43)]. Next, analyses replicating those above were used to examine the utility of relying subsets of the data. Considered as solo predictors, the following clinical characteristics were predictive [Ln-age, *C* = 0.65, cutoff = 4.13 (62.30 years of age); personal history of other cancer, *C* = 0.59; COPD, *C* = 0.59; Ln-pack-years, *C* = 0.78, cutoff = 1.32 (3.74 pack-years)]. These factors, utilized together, were good at predicting incidence of lung cancer in the replication sample (*C* = 0.83; cutoff = 11.28). Examinations of cancer type revealed that although risk scores were similarly effective at identifying incidence of late-stage (stages III/IV; *C* = 0.86) and early-stage (*C* = 0.84) cancers, the risk score was better at identifying non-small cell cancer (*C* = 0.87) versus other cancers (*C* = 0.74).

### Discussion

The Stony Brook LCEC was established to diagnose and treat patients with lung cancer, as well as provide monitoring and follow-up for those presenting with a concerning nodule (but not cancer) at initial consult. In this investigation, the risk of developing lung cancer was calculated utilizing a set of clinical, demographic, and imaging-related risk factors including: Ln age, Ln-pack-years

**Table 3.** Posterior descriptive statistics examining measures of lung cancer risk, stratified into high-risk versus low-risk groups estimated using information from survival models in the replication sample, LCEC 2002 to 2016

Risk group	Mean risk score (95% CI)	Incidence rate (95% CI)	1-Year risk	5-Year risk	Estimated years until 10% of group has cancer	HR (95% CI)
Low risk	9.01 (8.97–9.05)	18.67 (11.26–30.97)	0.18%	1.08%	107.11	1.00
High risk	11.25 (11.16–11.33)	304.24 (238.13–388.7)	4.87%	17.19%	6.57	14.34 (8.17–25.18)

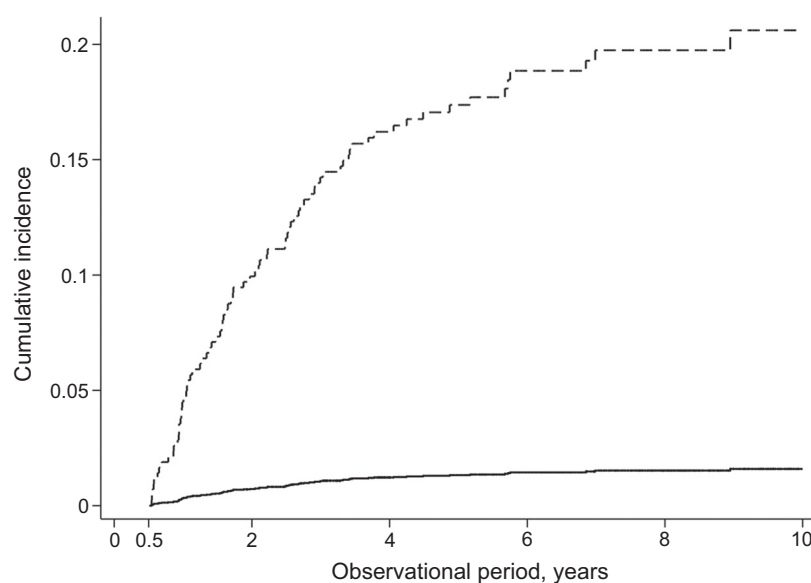
NOTE: Incidence rate is estimated per 1,000 person-years. HRs were calculated using Cox proportional hazards regression. Patients spent 10,137.02 years at risk in the replication sample; cases spent 183.86 years at risk. Years until 10% of group has cancer was calculated using Rothman waiting-period method.

of smoking, cancer history, COPD, ground-glass opacity, and spiculation. This study provides the first risk prediction model to our knowledge for the identification of incident lung cancer in a monitoring program of patients with pulmonary nodules discovered on CT. The final model fit well, and risk scores had good accuracy to identify high risk of incident lung cancer.

Accurate risk prediction tools for lung cancer are critical for efficient and effective cancer care. To date, existing models have been based on retrospective study designs, small samples, or heavy smoking populations, and included primarily epidemiologic factors (3–9, 22). Recent models have focused on utilizing population-based cohort studies: the European Prospective Investigation into Cancer Study derived a predictive model for lung cancer based on age and smoking history (23); the Carotene and Retinol Efficacy Trial derived a model that additionally included gender and asbestos exposure (24); the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial further incorporated education, body mass, COPD, and a personal history of cancer and family history of lung cancer (25). One study evaluating a large database (1.3 million men) in South Korea developed a risk prediction model for prevalent lung cancer based on age, smoking history, body mass, physical activity, and fasting glucose level (26). However, the risk of lung cancer in studies of the general population is deter-

mined, in part, by the risk of being screened. If physical health or levels of access to medical care are associated with the screening likelihood, then results from population data and lung cancer screening clinics may diverge. This study differed in prospectively monitoring individuals with an existing pulmonary nodule.

As a result of new USPSTF recommendations, several studies have begun to discuss prediction models in relation to lung cancer screening. To date, these have relied primarily upon subgroups of high-risk patients. First, the Liverpool Lung Project (LLP) utilized a cohort of 8,760 participants to show that age, gender, smoking duration, COPD, prior diagnosis of other cancer, and family history of early onset lung cancer were predictive of lung cancer, however radiologic features were not assessed (12). Second, in a study of heavy smokers without a diagnosis of cancer ( $n = 5,203$ ), a risk prediction model incorporating nodule characteristics, including nonsolid and large (>8 mm) nodes, improved prediction accuracy compared with models including epidemiologic factors alone (27). Although important, these studies relied on high-risk patients. Yet, many patients referred to evaluation clinics fall outside of the well-defined high-risk profile (28). This study, in contrast, relied on patients where risk of lung cancer was determined based on evidence of a pulmonary nodule. These data, although critical to improving future



**Figure 1.** Cumulative incidence stratified by risk grouping among individuals in the replication sample, LCEC 2002 to 2016. High-risk group (dashes); low-risk group (solid line).

guidelines, suggested that risk factors for lung cancer were predictive of having increased risk of lung cancer at follow-up. The best fitting model included both radiologic and clinical features. However, it is worth noting that contrary to prior studies (27), epidemiologic features were relatively good at identifying risk of incident lung cancer among patients with a known nodule. Notable among these, ln-pack-years was critical for identifying risk, though higher risk individuals were those with more than 3.74 pack-years.

This study is novel in examining risk of incident lung cancer in patients that may have no history of smoking. To date, there is only one study utilizing patients undergoing low dose CT screening to determine risk of prevalent lung cancer (11). That study relied on the Pan-Canadian Early Detection Lung Cancer Study ( $N = 7,008$  with 102 lung cancers) that utilized data from the British Columbia Cancer Agency ( $N = 5,021$  with 42 lung cancers) for validation. The study found nodule size, type, location, count, and spiculation, along with demographic factors and a history of COPD, were important to determining risk of lung cancer. Although prior studies have identified epidemiologic and imaging risk factors that are critical to lung cancer risk, they have not quantified that risk in relation to incident lung cancer among those deemed to be cancer-free at initial consult. The present investigation therefore extends this line of research to quantify predictive capability of a risk score in a study of lung cancer incidence. The resulting risk prediction model demonstrated high discriminant value for identifying risk of developing lung cancer. Whereas these results represent notable findings, further refinements may be necessary to optimize the clinical significance of such models, and to identify appropriate decision-making algorithms for use in clinical settings.

### Limitations

Although this investigation is strengthened by its prospective study design spanning over one decade, its well-defined large cohort of incident lung cancer cases and noncases and its inclusion of both epidemiologic and radiologic factors in the risk prediction models, the study has several limitations. There is potential for reporting bias, as data on smoking history, exposures, and other epidemiologic factors were self-reported. Although an internal replication sample in this investigation was included, due to the uniqueness of these data collection efforts, the final model was not validated in a separate sample. The risk score that was calculated indicated an average risk for individuals and was based on a number of indicators that may vary with time. Therefore, reliance on scores may be biased as variation changes. Further work may help to clarify the role of changing characteristics in modifying risk. In addition,

although a cutoff was provided to usefully differentiate between individuals deemed at high risk of incident lung cancer, a number of individuals in the low-risk group developed lung cancer. Decreasing the risk cutoff could increase the chances of identifying lung cancer at the cost of increased resources needed for monitoring. The applicability of this study to individuals without evidence of a pulmonary nodule is unclear. Given that this study was based on data originating from a single Cancer Center, the generalizability of the findings to other populations is unknown. Several studies have considered markers of DNA repair and particular SNPs associated with disease development (3, 4, 23, 29–34). Although useful, this study was unable to examine the role of genetic factors. Mortality data were used to censor individuals; however, it remains uncertain whether mortality information is systematically biased among individuals originally determined to be at low risk of cancer. While generalizability of the sample may be unclear, the risk factors identified here are not unique to this study but instead represent common risk factors for lung cancer incidence and mortality, thus suggesting that the biological mechanisms may generalize to other populations. Furthermore, the characteristics identified collate information from a number of related studies with different populations and results. Nevertheless, given that a number of factors, including for example race/ethnicity, were not observed, this remains an important limitation.

### Conclusions

Differentiating pulmonary nodules that will progress into cancer from those that will remain benign has high clinical utility. This study sought to quantify the risk of incident lung cancer among patients found to have lung nodules identified on CT. The risk prediction model developed in the present investigation estimates risk by utilizing a common modeling parameter to determine a set of significant epidemiologic and radiologic features. Risk prediction may reduce the burden of care in some subpopulations with very low risk, while ensuring that those deemed to be at high risk are closely monitored in a consistent and evidence-based manner. The final model may assist physicians in managing the care of patients who may be at increased risk of developing lung cancer.

### Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

### Authors' Contributions

**Conception and design:** B. Nemesure, S. Clouston, D. Albano, S. Kuperberg, T.V. Bilfinger

**Development of methodology:** B. Nemesure, S. Clouston, T.V. Bilfinger

**Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.):** B. Nemesure, S. Kuperberg, T.V. Bilfinger

**Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis):** B. Nemesure, S. Clouston, T.V. Bilfinger

**Writing, review, and/or revision of the manuscript:** B. Nemesure, S. Clouston, D. Albano, T.V. Bilfinger

**Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases):** B. Nemesure, S. Clouston

**Study supervision:** B. Nemesure

**Other (physician involved in conceptualization of strategy for nodule evaluation):** S. Kuperberg

## Acknowledgments

Secondary data analysis of de-identified clinical data was deemed by the Stony Brook University ethics review board to be not human subjects research.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received December 21, 2018; revised February 14, 2019; accepted May 13, 2019; published first June 27, 2019.

## References

- Centers for Disease Control and Prevention. Leading Causes of Death. Atlanta, GA: Centers for Disease Control and Prevention; 2017.
- Alberg AJ, Samet JM. Epidemiology of lung cancer. *Chest* 2003; 123(1 Suppl):21S–49S.
- Cassidy A, Myles JP, van Tongeren M, Page RD, Liloglou T, Duffy SW, et al. The LLP risk model: an individual risk prediction model for lung cancer. *Br J Cancer* 2008;98: 270–6.
- Spitz MR, Etzel CJ, Dong Q, Amos CI, Wei Q, Wu X, et al. An expanded risk prediction model for lung cancer. *Cancer Prev Res (Phila)* 2008;1:250–4.
- Spitz MR, Hong WK, Amos CI, Wu X, Schabath MB, Dong Q, et al. A risk model for prediction of lung cancer. *J Natl Cancer Inst* 2007; 99:715–26.
- Etzel CJ, Kachroo S, Liu M, D'Amelio A, Dong Q, Cote ML, et al. Development and validation of a lung cancer risk prediction model for African-Americans. *Cancer Prev Res (Phila)* 2008;1: 255–65.
- Schultz EM, Sanders GD, Trotter PR, Patz EF Jr, Silvestri GA, Owens DK, et al. Validation of two models to estimate the probability of malignancy in patients with solitary pulmonary nodules. *Thorax* 2008;63:335–41.
- Swensen SJ, Silverstein MD, Ilstrup DM, Schleck CD, Edell ES. The probability of malignancy in solitary pulmonary nodules. Application to small radiologically indeterminate nodules. *Arch Intern Med* 1997;157:849–55.
- Gould MK, Donington J, Lynch WR, Mazzone PJ, Midthun DE, Naidich DP, et al. Evaluation of individuals with pulmonary nodules: when is it lung cancer? Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest* 2013;143(5 Suppl):e93S–120S.
- Moyer VA. Screening for lung cancer: US Preventive Services Task Force recommendation statement. *Ann Intern Med* 2014;160: 330–8.
- McWilliams A, Tammemagi MC, Mayo JR, Roberts H, Liu G, Soghrati K, et al. Probability of cancer in pulmonary nodules detected on first screening CT. *N Engl J Med* 2013; 369:910–9.
- Marcus MW, Chen Y, Raji OY, Duffy SW, Field JK. LLPi: liverpool lung project risk prediction model for lung cancer incidence. *Cancer Prev Res (Phila)* 2015;8:570–5.
- Katki HA, Kovalchik SA, Berg CD, Cheung LC, Chaturvedi AK. Development and validation of risk models to select ever-smokers for CT lung cancer screening. *JAMA* 2016;315: 2300–11.
- Wiener RS, Gould MK, Woloshin S, Schwartz LM, Clark JA. 'The thing is not knowing': patients' perspectives on surveillance of an indeterminate pulmonary nodule. *Health Expect* 2015;18: 355–65.
- Fromer L, Cooper CB. A review of the GOLD guidelines for the diagnosis and treatment of patients with COPD. *Int J Clin Pract* 2008;62:1219–36.
- Cox DR, Oakes D. Analysis of survival data. Boca Raton, FL: CRC Press; 1984.
- Cohen J. Statistical power analysis for the behavioral sciences. Lawrence Erlbaum Associates: USA; 1977.
- Youden WJ. Index for rating diagnostic tests. *Cancer* 1950;3: 32–5.
- Hallas J, Gaist D, Bjerrum L. The waiting time distribution as a graphical approach to epidemiologic measures of drug utilization. *Epidemiology* 1997:666–70.
- Benjamini Y, Hochberg Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J Roy Stat Soc B Met* 1995;57:289–300.
- Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Statist Assoc* 1999;94: 496–509.
- Soardi GA, Perandini S, Larici AR, del Ciello A, Rizzardi G, Solazzo A, et al. Multicentre external validation of the BIMC model for solid solitary pulmonary nodule malignancy prediction. *Eur Radiol* 2017;27:1929–33.
- Hoggart C, Brennan P, Tjonneland A, Vogel U, Overvad K, Ostergaard JN, et al. A risk model for lung cancer incidence. *Cancer Prev Res (Phila)* 2012;5:834–46.
- Bach PB, Kattan MW, Thornquist MD, Kris MG, Tate RC, Barnett MJ, et al. Variations in lung cancer risk among smokers. *J Natl Cancer Inst* 2003;95:470–8.
- Tammemagi CM, Pinsky PF, Caporaso NE, Kvale PA, Hocking WG, Church TR, et al. Lung cancer risk prediction: prostate, lung, colorectal and ovarian cancer screening trial models and validation. *J Natl Cancer Inst* 2011;103: 1058–68.
- Park S, Nam BH, Yang HR, Lee JA, Lim H, Han JT, et al. Individualized risk prediction model for lung cancer in Korean men. *PLoS One* 2013;8:e54823.
- Maisonneuve P, Bagnardi V, Bellomi M, Spaggiari L, Pelosi G, Rampinelli C, et al. Lung cancer risk prediction to select smokers for screening CT—a model based on the Italian COSMOS trial. *Cancer Prev Res (Phila)* 2011;4:1778–89.
- Nemesure B, Plank A, Reagan L, Albano D, Reiter M, Bilfinger TV. Evaluating efficacy of current lung cancer screening guidelines. *J Med Screen* 2017;24:208–13.

29. Deng L, Kimmel M, Foy M, Spitz M, Wei Q, Gorlova O. Estimation of the effects of smoking and DNA repair capacity on coefficients of a carcinogenesis model for lung cancer. *Int J Cancer* 2009;124:2152–8.
30. Cassidy A, Myles JP, Liloglou T, Duffy SW, Field JK. Defining high-risk individuals in a population-based molecular-epidemiological study of lung cancer. *Int J Oncol* 2006;28:1295–301.
31. Li H, Yang L, Zhao X, Wang J, Qian J, Chen H, et al. Prediction of lung cancer risk in a Chinese population using a multifactorial genetic model. *BMC Med Genet* 2012;13:118.
32. Raji OY, Agbaje OF, Duffy SW, Cassidy A, Field JK. Incorporation of a genetic factor into an epidemiologic model for prediction of individual risk of lung cancer: the Liverpool Lung Project. *Cancer Prev Res (Phila)* 2010;3:664–9.
33. Raji OY, Duffy SW, Agbaje OF, Baker SG, Christiani DC, Cassidy A, et al. Predictive accuracy of the Liverpool Lung Project risk model for stratifying patients for computed tomography screening for lung cancer: a case-control and cohort validation study. *Ann Intern Med* 2012;157:242–50.
34. Young RP, Hopkins RJ, Hay BA, Epton MJ, Mills GD, Black PN, et al. Lung cancer susceptibility model based on age, family history and genetic variants. *PLoS One* 2009;4:e5302.

