ORIGINAL ARTICLE

Using socio-demographic and early clinical features in general practice to identify people with lung cancer earlier

Barbara Iyen-Omofoman,[1] Laila J Tata,[1] David R Baldwin,[2] Chris JP Smith,[1] Richard B Hubbard[1,2,3]

[1]Department of Epidemiology and Public Health, University of Nottingham, Nottingham, UK
[2]Nottingham University Hospitals NHS Trust, Nottingham, UK
[3]Respiratory Biomedical Research Unit, University of Nottingham, Nottingham, UK

Correspondence to
Dr Barbara Iyen-Omofoman, Department of Epidemiology and Public Health, University of Nottingham, Clinical Sciences Building, City Hospital, Nottingham NG5 1PB, UK; barboiyen@yahoo.com

Received 27 June 2012
Revised 23 November 2012
Accepted 18 December 2012
Published Online First
15 January 2013

ABSTRACT

Introduction In the UK, most people with lung cancer are diagnosed at a late stage when curative treatment is not possible. To aid earlier detection, the socio-demographic and early clinical features predictive of lung cancer need to be identified.

Methods We studied 12 074 cases of lung cancer and 120 731 controls in a large general practice database. Logistic regression analyses were used to identify the socio-demographic and clinical features associated with cancer up to 2 years before diagnosis. A risk prediction model was developed using variables that were independently associated with lung cancer up to 4 months before diagnosis. The model performance was assessed in an independent dataset of 1 826 293 patients from the same database. Discrimination was assessed by means of a receiver operating characteristic (ROC) curve.

Results Clinical and socio-demographic features that were independently associated with lung cancer were patients' age, sex, socioeconomic status and smoking history. From 4 to 12 months before diagnosis, the frequency of consultations and symptom records of cough, haemoptysis, dyspnoea, weight loss, lower respiratory tract infections, non-specific chest infections, chest pain, hoarseness, upper respiratory tract infections and chronic obstructive pulmonary disease were also independently predictive of lung cancer. On validation, the model performed well with an area under the ROC curve of 0.88.

Conclusions This new model performed substantially better than the current National Institute for Health and Clinical Excellence referral guidelines and all comparable models. It has the potential to predict lung cancer cases sufficiently early to make detection at a curable stage more likely by allowing general practitioners to better risk stratify their patients. A clinical trial is needed to quantify the absolute benefits to patients and the cost effectiveness of this model in practice.

Key massages

What is the key question?

- Can the early records of patients with lung cancer in general practice be used to develop a predictive model that will aid earlier identification of patients with lung cancer?

What is the bottom line?

- A model developed using a combination of patients' socio-demographic and clinical features was found to be predictive of lung cancer 4–12 months before diagnosis and outperformed the current National Institute for Health and Clinical Excellence referral guidelines and all comparable models.

Why read on?

- The model developed and validated in this study is the first risk-prediction model for lung cancer that incorporates the combination of patients' baseline characteristics and early clinical features by excluding records made in the months before diagnosis when general practitioners had initiated investigations for suspected lung cancer. Application of this model in practice should lead to earlier identification and an improved prognosis for patients with lung cancer in general practice.

INTRODUCTION

Lung cancer is the most common cancer and the leading cause of cancer deaths worldwide.[1] Survival from lung cancer is known to vary across Europe[2] and for patients in the UK, survival is lower than other comparable countries.[3–5] Delays in diagnosis are thought to contribute to this problem.[3 4] Since curative treatments for lung cancer are only available for the minority of people with cancers diagnosed in the early stages,[6] any change that results in earlier diagnosis is a priority. The National Awareness and Early Diagnosis Initiative established in 2008 has set up programmes to increase public awareness of symptoms of lung cancer.[7] There are currently no widely available screening tests for lung cancer, although several randomised controlled trials are ongoing,[8–11] one of which has shown a 20% reduction in mortality. There are also no clinical predictive models currently available that have been demonstrated to detect lung cancer at a stage early enough to improve clinical outcomes.

Most patients with lung cancer experience at least one symptom before diagnosis.[12] In a study of 22 people with recently diagnosed lung cancer, symptoms were recalled starting between 4 months and 2 years before diagnosis.[13] In the UK, the general practitioner (GP) acts as the gatekeeper to specialised healthcare and most people present with

To cite: Iyen-Omofoman B, Tata LJ, Baldwin DR, *et al*. *Thorax* 2013;**68**:451–459.

symptoms to their GP before the diagnosis of lung cancer is made.[13 14] A case–control study of 247 cases of lung cancer and 1235 controls from 21 practices in Exeter, UK showed that haemoptysis, dyspnoea, abnormal spirometry and smoking were independently associated with lung cancer up to 180 days before diagnosis.[14] The National Institute for Health and Clinical Excellence (NICE) referral guidelines[15] have provided a big step forward in aiding earlier diagnosis of lung cancer by facilitating urgent referral of suspected lung cancer cases; however, the guidelines were not designed to specifically improve patient ascertainment and were not based on a strong evidence base.[13 16 17] Because many lung cancer symptoms are non-specific, GPs need help to estimate the risk of lung cancer by taking into account a combination of socio-demographic features and clinical symptoms.

Although several risk prediction models have been developed to estimate the risk of lung cancer,[18–21] only one algorithm has been developed using a combination of patients' baseline risk factors and symptoms in primary care.[22] However, this model did not exclude symptoms in the period preceding lung cancer diagnosis when patients would likely be undergoing investigations for suspected cancer and it may therefore be limited in its ability to identify patients with lung cancer early enough to result in an improvement in outcome.

The aim of this study is to develop and validate a lung cancer risk prediction model that could be used to aid earlier diagnosis in general practice by identifying the socio-demographic factors and the pattern and frequency of symptoms and clinical investigations prior to diagnosis.

METHODS

The general practice data used in this study were from The Health Improvement Network[23] (THIN), a large nationally representative database of general practice records in the UK. Over 95% of the UK population is registered with a GP and general practices in THIN are broadly representative of general practices across the UK in terms of the patient demographics, geographical distribution and practice size.[24] THIN has a high level of completeness of lung cancer data and the characteristics of patients with lung cancer in THIN are representative of the UK lung cancer population.[25] At the time of this study, THIN had data from 446 UK general practices with a total of 8.2 million patients. To derive the lung cancer risk-prediction model, we identified all incident cases of lung cancer diagnosed between 1 January 2000 and 28 July 2009 (Read code list available). Patients who had less than 1 year of active records prior to their first diagnosis of lung cancer were removed to exclude prevalent cases. Since lung cancer is rare in patients younger than 40 years of age, these patients were also excluded. For each case, 10 randomly selected controls were identified. Controls were registered in the same general practice as the case, with at least 1 year of active data, and they were aged 40 years or older at the time of lung cancer diagnosis in their practice-matched case.

The variables analysed were 5-year age band, sex, socioeconomic status (Townsend deprivation quintiles) and smoking history. Smoking records made within 6 months preceding lung cancer diagnosis were excluded to account for a possible change in cigarette consumption in the months leading up to diagnosis. Patients were categorised as current smokers, ex smokers or non-smokers. Based on the highest ever recorded number of cigarettes smoked daily, the smoking records of current or ex smokers were further categorised as trivial (less than one cigarette daily), light (1–9 cigarettes daily), moderate (10–19 cigarettes daily), heavy (20–39 cigarettes daily) or very heavy (more than 40 cigarettes daily). Smokers who had no records of daily cigarette consumption were recorded as such and patients who had no recorded smoking information were coded in a separate category.

Symptoms that were analysed in cases and controls were those detailed in the NICE guidelines[15] (box 1). In addition, we assessed the six most common symptoms and diagnoses recorded in the records of patients with lung cancer prior to their diagnoses. These were upper and lower respiratory tract infections (URTI and LRTI), non-specific chest infections, constipation, depressive disorders and chronic obstructive pulmonary disease (COPD). Records of chest x-rays, blood tests and number of general practice consultations for symptoms other than those already assessed were also identified.

All symptoms, diagnoses and investigations over the 2-year period before lung cancer diagnosis (or matched date) were identified. Since a chest x-ray is the initial investigation for suspected lung cancer,[15] we examined the timing of chest x-rays prior to lung cancer diagnosis and found a steep increase in the chest x-ray frequency in cases (but not controls) within the 4 months prior to diagnosis; so all symptoms, blood tests and other general practice consultations recorded within this period were excluded.

To determine the independent early predictors for lung cancer, univariate logistic regression models were used to calculate ORs. These analyses for symptoms, diagnoses, blood tests and GP consultations were done separately for records made in the 4–12 month and 13–24 month periods prior to diagnosis. Multivariate modelling was done using only variables that were associated with lung cancer in univariate analyses, using a statistical significance cut-off level of $p<0.05$. Variables that were not statistically significant in the multivariate analysis were removed from the model and those that previously showed no association with lung cancer in the univariate model were rechecked for significance in the final model. In developing the risk probabilities for lung cancer, we weighted each variable according to the strength of its association in the multivariate logistic regression model and then applying the method used to develop the Thoracic Surgery Scoring System (Thoracoscore),[26] the β-coefficient values (log OR) from the multivariate model were used to compute aggregate scores for individual patients.

Validation of the model was carried out on a cohort all THIN patients who were 39 years or older and free from lung cancer on 29 July 2009. Eligibility in this cohort was limited to patients who had at least 1 year of general practice follow-up.

Box 1 Clinical features for which urgent referral for a chest x-ray should be offered for suspected lung cancer[15]

Haemoptysis
Any of the following unexplained or persistent symptoms or signs:
- Cough
- Chest/shoulder pain
- Dyspnoea
- Weight loss
- Hoarseness
- Finger clubbing*
- Features suggestive of metastasis from lung cancer*
- Cervical/supraclavicular lymphadenopathy*

*Clinical features not analysed in study.

Each person was given a lung cancer risk probability score on the basis of their records. The actual number of incident lung cancer cases within the year after 29 July 2009 were identified and then the performance of the model was assessed by comparing the sensitivity and specificity at different cut-offs. Additionally, a comparison of the sensitivity and specificity of this model with those of the NICE guideline symptoms was made. The discriminatory power of the model was assessed by means of a receiver operating characteristic (ROC) curve and an area under the curve (AUC) calculation.

All analyses were performed using Stata release SE11 and the study protocol was approved in 2009 by the Cegedim Strategic Data Medical Research Scientific Review Committee.

RESULTS

We identified 12 135 incident cases of lung cancer. After excluding 59 patients who were under 40 years old at the time of diagnosis (0.49%), 12 073 cases were matched with 10 controls each, two cases had no eligible controls and were excluded, and the remaining case had one eligible control, giving a total of 12 074 cases and 120 731 controls. The average follow-up time prior to diagnosis was similar in the cases and controls: 9.5 years (IQR 5.5–13.5 years) and 9.1 years (IQR 5.2–13.2 years) respectively. Compared with controls, people with lung cancer were more likely to be older men, live in households located in more deprived areas and more likely to be current or ex smokers (table 1).

A plot of the chest x-ray frequency among cases leading up to lung cancer diagnoses showed a stable pattern up to the fourth month preceding diagnosis. However, after this, there was a steep increase, implying that investigations for lung cancer were initiated by GPs (figure 1).

Analysis of the symptoms, diagnoses, blood tests and other GP consultations in the 4–12 month and 13–24 month periods preceding lung cancer diagnosis (table 2) showed greater ORs for lung cancer with all the symptoms recorded in the 4–12 month period than in the 13–24 month period. Furthermore, graphically, the increase in symptom presentations in cases occurred in the year before diagnosis (plot not shown), so the remaining analyses focused only on the 4–12 month period.

The symptoms with the highest frequency among cases were cough, non-specific chest infections, dyspnoea, chest pain and COPD. Although haemoptysis records were made for only 2% of cases in the 4–12 months before diagnosis, the OR for lung cancer among people who had haemoptysis in this period was

Table 1 Social, demographic and lifestyle characteristics of lung cancer cases and controls

	Cases n (%) N=12074	Controls n (%) N=120731	Unadjusted OR for lung cancer (95% CI)
Age at diagnosis (years)			
>80	2639 (21.86)	10797 (8.94)	48.80 (39.72 to 59.97)
75–80	2305 (19.09)	8191 (6.78)	56.19 (45.69 to 69.10)
70–75	2212 (18.32)	9940 (8.23)	44.43 (36.13 to 54.64)
65–70	1750 (14.49)	11201 (9.28)	31.20 (25.34 to 38.40)
60–65	1488 (12.32)	13475 (11.16)	22.05 (17.90 to 27.16)
55–60	896 (7.42)	15439 (12.79)	11.59 (9.37 to 14.33)
50–55	469 (3.88)	15963 (13.22)	5.87 (4.70 to 7.32)
45–50	220 (1.82)	16756 (13.88)	2.62 (2.06 to 3.34)
40–45	95 (0.79)	18969 (15.71)	1.00
Sex			
Men	7154 (59.25)	58034 (48.07)	1.57 (1.51 to 1.63)
Women	4920 (40.75)	62697 (51.93)	1.00
Townsend deprivation quintile			
5 (most deprived)	2234 (18.50)	15997 (13.25)	1.94 (1.82 to 2.07)
4	2640 (21.87)	21071 (17.45)	1.74 (1.64 to 1.85)
3	2421 (20.05)	23791 (19.71)	1.41 (1.33 to 1.50)
2	2236 (18.52)	26540 (21.98)	1.17 (1.10 to 1.25)
1 (least deprived)	2064 (17.09)	28681 (23.76)	1.00
Missing Townsend records	479 (3.97)	4651 (3.85)	1.43 (1.29 to 1.59)
Smoking status			
Current very heavy (40+/day)	471 (3.90)	1466 (1.21)	12.52 (11.14 to 14.09)
Current heavy (20–39/day)	2589 (21.44)	10928 (9.05)	9.24 (8.61 to 9.90)
Current moderate (10–19/day)	1,665 (13.79)	8247 (6.83)	7.87 (7.29 to 8.49)
Current light (1–9/day)	607 (5.03)	3765 (3.12)	6.28 (5.68 to 6.96)
Current trivial (<1/day)	7 (0.06)	144 (0.12)	1.89 (0.89 to 4.05)
Current, no qty recorded	439 (3.64)	4495 (3.72)	3.81 (3.40 to 4.26
Ex very heavy (40+/day)	221 (1.83)	841 (0.70)	10.24 (8.75 to 12.00)
Ex heavy (20–39/day)	1043 (8.64)	4258 (3.53)	9.55 (8.75 to 10.42)
Ex moderate (10–19/day)	777 (6.44)	4394 (3.64)	6.89 (6.27 to 7.57)
Ex light (1–9/day)	399 (3.30)	2837 (2.35)	5.48 (4.87 to 6.17)
Ex trivial (<1/day)	13 (0.11)	289 (0.24)	1.75 (1.00 to 3.06)
Ex, no qty recorded	1780 (14.74)	16027 (13.27)	4.33 (4.02 to 4.66)
Non-smoker	1300 (10.77)	50676 (41.97)	1.00
Missing smoking records	763 (6.32)	12364 (10.24)	2.41 (2.20 to 2.64)

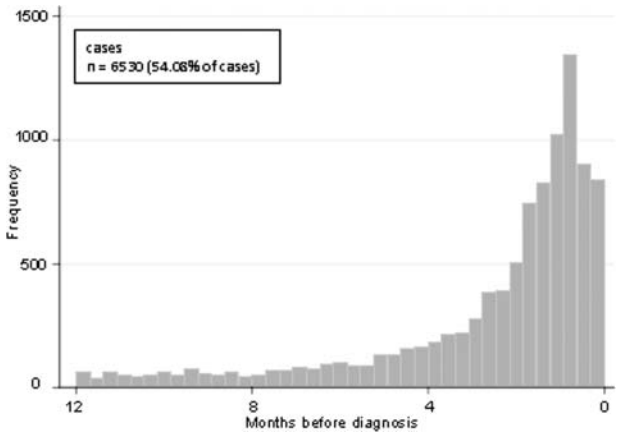

Figure 1 Frequency distribution of chest x-rays among cases in general practice, 12 months prior to lung cancer diagnoses. The plot for frequency of chest x-rays in controls is not shown but the pattern was consistent over the 12-month period and overall only 4% of controls had chest x-rays performed within the 12 months. This figure is only reproduced in colour in the online version.

20.15 (95% CI 16.24 to 25.01). Compared with controls, people with lung cancer consulted their GPs for other symptoms more often before diagnosis. Using fewer than 10 consultations as a reference value, the OR for cases to consult their GPs 21 times or more was 4.45 (95% CI 4.24 to 4.68) in the 4–12 months before diagnosis. There were also more blood investigations among cases than controls, with an increase in the number of normal and abnormal test results.

Development of the lung cancer risk model

Our model was developed using the independent predictors of lung cancer in the 4–12 month period before diagnosis (table 3). Variables that were independently associated with lung cancer and included in the final model were age, sex, Townsend deprivation quintiles, smoking (status and highest daily cigarette consumption), number of other GP consultations, and symptom presentations of cough, haemoptysis, dyspnoea, weight loss, LRTI, non-specific chest infections, COPD, chest/shoulder pain, voice hoarseness and URTI. Constipation, depression and blood tests were not independently associated with lung cancer. The odds of lung cancer increased with increasing age, male sex, greater socioeconomic deprivation and higher daily cigarette consumption. The association with daily cigarette consumption was stronger among current smokers than ex smokers. Haemoptysis and weight loss were relatively uncommon symptoms among lung cancer cases but they were associated with the greatest risk of lung cancer.

Using β-coefficient values derived from multivariate logistic regression (shown in table 3), aggregate risk probabilities were computed for individual patients in the dataset using the equation:

$$\text{Risk score} = \text{constant} + \text{sum of } \beta \text{ coefficients at different values of the exposure variables.}$$

The validation cohort comprised 1 826 293 patients in THIN who had no history of lung cancer up to the 29 July 2009 and with at least 1 year of follow-up data before and after 29 July 2009. There were 939 299 women (51.4%) and 886 994 men (48.6%). A total of 1728 incident cases of lung cancer (0.09% of the cohort) were identified during the 1-year of follow-up from 29 July 2009.

Risk probability scores were computed for all individuals in the dataset using the β-coefficient values derived from the logistic regression model. The number of patients identified by the score at different cut-off values, and the sensitivity and specificity of the risk model at the cut-off values are shown in table 4.

Table 5 shows, for different symptoms in the NICE guidelines, the number of patients in the validation cohort who will require a chest x-ray, the number of true positives identified and the sensitivity and specificity of the guideline symptoms in predicting lung cancer risk. Using haemoptysis alone as a trigger for chest x-rays, only 24 cases of lung cancer in the cohort population can be detected. Using the most commonly reported symptom, cough, as a trigger for investigations, 175 290 patients are identified to be at risk of lung cancer and 413 of these will be diagnosed with lung cancer. Therefore, using the NICE symptoms to identify a comparable number of true positives as the lung cancer risk model, a higher number of patients are required to undergo chest x-rays than the risk model. For example, at a cut-off to identify 610 cases of lung cancer in the validation cohort, the risk model identified 72 883 patients at high risk of lung cancer for whom chest x-ray investigations are indicated (119 chest x-rays per identified case), yet using a weighted combination of all the NICE symptoms, a total of 305 137 patients will have to undergo chest x-ray investigations to identify 724 cases of lung cancer (421 chest x-rays per identified case).

The ROC curve obtained from the application of the risk model in the validation cohort is shown in figure 2. The AUC is 0.88. Using a weighted combination of the NICE guideline symptoms alone to identify patients at high risk of lung cancer, the area under the ROC curve was 0.64.

DISCUSSION

We used a combination of patients' socio-demographic and clinical records in general practice to develop a lung cancer risk prediction model which can be used by GPs to aid earlier identification of patients at high risk of lung cancer. On validating this model in an independent dataset, it performed well and showed good discrimination, with an area under the ROC curve of 0.88.

The lung cancer risk prediction model was developed using the THIN database, which has previously been validated against other UK national lung cancer databases.[25] By incorporating information that is routinely collected and therefore readily available to GPs, application of the risk model from this study could allow an easy and practical means of identifying general practice patients who are at risk of lung cancer, at no extra cost to GPs. We excluded records made in the 4 months prior to diagnosis of lung cancer to avoid symptoms, diagnoses and investigations attributable to lung cancer rather than predictive of it, and we focused on the 4–12 month period because symptom records by cases increased in the 12 months before diagnosis. This ensured that our model would predict lung cancer at an earlier stage.

Some relevant information is not reliably recorded in THIN (occupational exposure to carcinogens such as asbestos) and so could not be included in the model. Although inclusion of these variables may improve the performance of the model, the validation analyses using the currently available variables have shown good discrimination and the model performed substantially better than the current NICE guidelines[15] when validated in an independent dataset. With further improvements in general

Table 2 Symptoms, blood investigations and number of general practice consultations recorded among cases and controls in the 4–12 and 13–24 month periods prior to lung cancer diagnosis

Variable in GP record	Cases n (%) N=12074	Controls n (%) N=120731	Unadjusted OR for lung cancer	95% CI	p Value*
Cough					
4–12 months	1938 (16.05)	7088 (5.87)	3.07	2.90 to 3.24	<0.001
13–24 months	1774 (14.69)	9087 (7.53)	2.12	2.00 to 2.24	<0.001
Haemoptysis					
4–12 months	247 (2.05)	125 (0.10)	20.15	16.24 to 25.01	<0.001
13–24 months	133 (1.10)	191 (0.16)	7.03	5.63 to 8.78	<0.001
Chest/shoulder pain					
4–12 months	1002 (8.30)	4880 (4.04)	2.15	2.00 to 2.31	<0.001
13–24 months	959 (7.94)	6540 (5.42)	1.51	1.40 to 1.62	<0.001
Voice hoarseness					
4–12 months	66 (0.55)	219 (0.18)	3.02	2.30 to 3.99	<0.001
13–24 months	56 (0.46)	326 (0.27)	1.72	1.30 to 2.29	<0.001
Dyspnoea					
4–12 months	1091 (9.04)	2479 (2.05)	4.74	4.40 to 5.10	<0.001
13–24 months	992 (8.22)	3,047 (2.52)	3.46	3.21 to 3.72	<0.001
Weight loss					
4–12 months	197 (1.63)	323 (0.27)	6.18	5.17 to 7.39	<0.001
13–24 months	139 (1.15)	416 (0.34)	3.37	2.78 to 4.09	<0.001
Constipation					
4–12 months	423 (3.50)	1469 (1.22)	2.95	2.64 to 3.29	<0.001
13–24 months	421 (3.49)	1848 (1.53)	2.32	2.09 to 2.59	<0.001
Depressive disorders					
4–12 months	365 (3.02)	3365 (2.79)	1.09	0.97 to 1.21	0.135
13–24 months	449 (3.72)	4705 (3.90)	0.95	0.86 to 1.05	0.333
URTI					
4–12 months	426 (3.53)	3082 (2.55)	1.40	1.26 to 1.55	<0.001
13–24 months	497 (4.12)	4274 (3.54)	1.17	1.06 to 1.29	<0.001
LRTI					
4–12 months	516 (4.27)	1585 (1.31)	3.36	3.03 to 3.71	<0.001
13–24 months	566 (4.69)	2218 (1.84)	2.63	2.39 to 2.89	<0.001
Non-specific chest infections					
4–12 months	1398 (11.58)	4350 (3.60)	3.50	3.29 to 3.73	<0.001
13–24 months	1356 (11.23)	5856 (4.85)	2.48	2.33 to 2.64	<0.001
COPD					
4–12 months	978 (8.10)	1349 (1.12)	7.80	7.17 to 8.49	<0.001
13–24 months	1024 (8.48)	1553 (1.29)	7.11	6.56 to 7.71	<0.001
Outcome of blood tests					
4–12 months					
No blood test record	6406 (53.06)	84997 (70.40)	1.00		
Test without results	5431 (44.98)	34295 (28.41)	2.10	2.02 to 2.18	
Abnormal	107 (0.89)	528 (0.44)	2.69	2.18 to 3.31	<0.001
Normal	130 (1.08)	911 (0.75)	1.89	1.57 to 2.28	
13–24 months					
No blood test record	6136 (50.82)	79446 (65.80)	1.00		
Test without results	5632 (46.65)	39255 (32.51)	1.86	1.79 to 1.93	
Abnormal	127 (1.05)	752 (0.62)	2.19	1.81 to 2.64	<0.001
Normal	179 (1.48)	1278 (1.06)	1.81	1.55 to 2.13	
Number of GP consultations					
4–12 months					
0–10	4316 (35.75)	77720 (64.37)	1.00		
11–20	4373 (36.22)	29327 (24.29)	2.69	2.57 to 2.81	<0.001
21 or more	3385 (28.04)	13684 (11.33)	4.45	4.24 to 4.68	
13–24 months					
0–10	3491 (28.91)	64881 (53.74)	1.00		
11–20	3492 (28.92)	29296 (24.27)	2.22	2.11 to 2.33	<0.001
21 or more	5091 (42.16)	26554 (21.99)	3.56	3.41 to 3.73	

*p Values for binary variables were obtained using Wald's test of significance. In variables with more than two categories, p values were obtained from the likelihood ratio test.
COPD, chronic obstructive pulmonary disease; GP, general practitioner; LRTI, Lower respiratory tract infections; URTI, upper respiratory tract infections.

Table 3 Multivariate model of factors associated with lung cancer 4–12 months before diagnosis

Risk factor variable	Adjusted OR (95% CI)	p Value*	β coefficient
Age at diagnosis (years)			
40–45	1.00		0.9164
45–50	2.50 (1.96 to 3.19)		1.6900
50–55	5.42 (4.34 to 6.78)		2.3669
55–60	10.67 (8.61 to 13.22)	<0.001	2.9746
60–65	19.59 (15.86 to 24.18)		3.3534
65–70	28.61 (23.17 to 35.32)		3.8006
70–75	44.74 (36.26 to 55.21)		4.0944
75–80	60.03 (48.62 to 74.12)		4.1828
>80	65.55 (53.10 to 80.93)		
Sex			
Men	1.62 (1.55 to 1.69)	<0.001	0.4805
Women			
Townsend score			
5 (most deprived)	1.10 (1.02 to 1.18)		0.0932
4	1.12 (1.05 to 1.20)	0.0017	0.1157
3	1.07 (1.00 to 1.14)		0.0640
2	1.00 (0.93 to 1.07)		-0.0009
1 (least deprived)	1.00		0.0099
Missing Townsend records	1.01 (0.90 to 1.13)		
Smoking status and 6-month qty			
Current very heavy (40+/day)	15.91 (13.90 to 18.21)		2.7664
Current heavy (20–39/day)	13.45 (12.44 to 14.54)		2.5984
Current moderate (10–19/day)	9.82 (9.04 to 10.68)		2.2845
Current light (1–9/day)	5.98 (5.36 to 6.68)		1.7885
Current trivial (<1/day)	2.68 (1.21 to 5.90)		0.9851
Current, no qty recorded	3.47 (3.08 to 3.91)		1.2432
Ex very heavy (40+/day)	5.33 (4.48 to 6.35)		1.6742
Ex heavy (20–39/day)	6.67 (6.06 to 7.35)		1.8980
Ex moderate (10–19/day)	4.50 (4.07 to 4.98)		1.5045
Ex light (1–9/day)	3.54 (3.12 to 4.02)	<0.001	1.2636
Ex trivial (<1/day)	1.21 (0.68 to 2.17)		0.1943
Ex, no qty recorded	2.57 (2.38 to 2.78)		0.9455
Missing smoking records	2.70 (2.45 to 2.97)		0.9922
Non-smoker	1.00		
Cough	1.63 (1.53 to 1.75)	<0.001	0.4915
Haemoptysis	8.70 (6.75 to 11.20)	<0.001	2.1630
Dyspnoea	1.41 (1.29 to 1.55)	<0.001	0.3449
Weight loss	2.66 (2.16 to 3.29)	<0.001	0.9794
LRTI	1.56 (1.38 to 1.76)	<0.001	0.4414
Chest infections	1.55 (1.44 to 1.68)	<0.001	0.4393
COPD	1.61 (1.46 to 1.78)	<0.001	0.4786
Chest/shoulder pain	1.39 (1.28 to 1.51)	<0.001	0.3296
Voice hoarseness	1.79 (1.28 to 2.49)	0.001	0.5806
URTI	1.15 (1.02 to 1.30)	0.020	0.1417
No. of GP consultations			
0–10	1.00		0.2032
11–20	1.23 (1.16 to 1.29)	<0.001	0.3069
21 or more	1.36 (1.28 to 1.44)		
Logistic regression constant			−7.2295

*p Values for binary variables were obtained using Wald's test of significance. In variables with more than two categories, p values were obtained from the likelihood ratio test.
COPD, chronic obstructive pulmonary disease; GP, general practitioner; LRTI, lower respiratory tract infection; URTI, upper respiratory tract infection.

practice data recording, a review of this model will be warranted to reflect more accurate lung cancer prediction. Another limitation in this study was the unavailability of information on cigarette pack-years for defining patients' lifetime cigarette exposure. As a proxy, we categorised patients' exposure to cigarette smoke using the highest recorded quantity of cigarettes smoked daily, which allowed us to classify patients' worst possible estimate of daily consumption. The results from analyses using these categories fit broadly with findings from the literature. Nevertheless, these pragmatic categorisations are using the

Table 4 Performance of the risk model at different cut-off values in the validation population (n=1 826 293)

Cut-off value	Patients at risk of lung cancer based on risk model	Patients not requiring a chest x-ray based on risk model	No. of true positives	No. of true negatives	Sensitivity (%)*	Specificity (%)†
−3	737390	1088903	1624	1088799	93.98	59.67
−2.5	541074	1285219	1526	1285017	88.31	70.43
−2	388040	1438253	1375	1437900	79.57	78.81
−1.5	255788	1570505	1182	1569959	68.40	86.05
−1.25	192433	1633860	1063	1633195	61.52	89.51
−1	144523	1681770	917	1680959	53.07	92.13
−0.5	72883	1752292	610	1752292	35.30	96.04
0	30994	1795299	367	1793938	21.24	98.32
0.5	11860	1814433	174	1812879	10.07	99.36

*Sensitivity=true positives / (true positives+false negatives).
†Specificity=true negatives / (true negatives+false positives).

information that would be available to GPs in standard practice for assessing their patients' risk.

Validation of the risk model in an independent cohort showed that a considerable number of patients need to undergo chest x-ray investigations to diagnose lung cancer cases. This is unsurprising considering that lung cancer was rare in our validation cohort and was only diagnosed in 1728 patients (0.09% of the population). Positive predictive values are not good measures of model accuracy, particularly with rare outcomes, as they are usually low even with good sensitivity and specificity.[27] A similar finding was shown in the randomised Danish lung cancer screening trial in which 980 CT scans were done to identify 69 lung cancer cases.[28] However, the model compared quite favourably with the NICE guideline symptoms, with about a quarter of chest x-rays required to detect a comparable number of lung cancers even than a weighted combination of the NICE guideline symptoms.

A number of models including the Bach,[18] Spitz[20] and the Liverpool Lung Project (LLP)[21] models have been developed to predict the risk of lung cancer using patients' baseline risk factors. The Bach model was developed to determine variation in lung cancer risk among current or former smokers aged between 55 and 74 years who were enrolled in a clinical trial of lung cancer prevention.[18] Since this model was developed using only data (age, sex, asbestos exposure and smoking history) from individuals with a smoking history, it is only applicable to smokers—a subset of individuals at risk of lung cancer. The expanded Spitz model was developed using information from 725 newly diagnosed cases of lung cancer and 615 healthy controls, on age, smoking history, family history of cancer, occupational exposure to carcinogens, previous history of respiratory disease and biomarker assays. This model is limited in that the biomarker assays included in the model derivation are select markers of host DNA repair capacity which require technical expertise and are not readily available in general practice.

A study that compared the discriminatory power of the Spitz, LLP and Bach models found an AUC statistic of 0.69 in the Spitz and LLP models and an AUC of 0.66 for the Bach model.[29] These are substantially lower than the AUC statistic value of 0.88 in our study. Compared with the Bach and Spitz models, the LLP model has been found to have a much higher rate of false positives and therefore falsely identifies more individuals who have low risk of lung cancer than the previous two models.[29] The LLP model is currently being used to select individuals who have a 5% risk of developing lung cancer over 5 years for inclusion in the UK lung screen trial of low-dose CT screening for lung cancer.[10] However, at a cut-off to capture 62% of cases of lung cancer, the LLP model falsely identifies 30% of non-lung cancer controls and does not perform as well as our risk model, which for accurately identifying 79.6% of lung cancer cases gives a false positive rate of 21.2%. However, LLP applies to asymptomatic patients.

Only one other model used patient records from a large primary care database to predict the risk of lung cancer.[22] In developing this model, patient records in the database were examined up to a certain time point to establish baseline risk, after which incident diagnoses of lung cancer over the subsequent 2 years were predicted. In the validation study of this model, it appeared to have a good discriminatory power with

Table 5 Sensitivity and specificity of NICE guideline symptoms alone in validation population (n=1 826 293)

Symptom	Patients requiring a chest x-ray based on NICE guideline	Patients not requiring a chest x-ray based on NICE guideline	No. of true positives	No. of true negatives	Sensitivity (%)	Specificity (%)
Haemoptysis	1843	1824450	24	1822746	1.39	99.90
Cough	175290	1651003	413	1649688	23.90	90.42
Chest/shoulder pain	107753	1718540	192	1717004	11.11	94.10
Dyspnoea	61631	1764662	315	1763249	18.23	96.64
Weight loss	7679	1818614	26	1816912	1.50	99.58
Voice hoarseness	5209	1821084	9	1819365	0.52	99.72

NICE, National Institute for Health and Clinical Excellence.

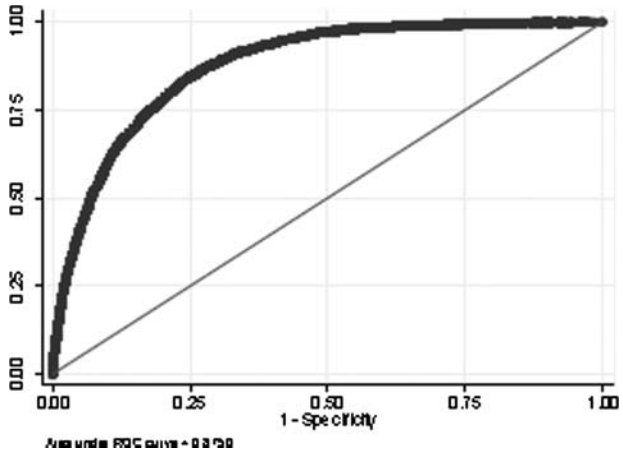

Figure 2 Receiver operating characteristic (ROC) curve for the lung cancer risk prediction model. The area under the curve is 0.88. The diagonal line represents the discrimination expected by chance alone. This figure is only reproduced in colour in the online version.

ROC values of 0.92 for men and women and at a threshold to identify the top 0.5% of patients at risk of lung cancer, the positive predictive value of the model was 1.3% (77 patients identified to be at risk of lung cancer for one true case). However, all GP records of patients recorded in the period leading up to lung cancer diagnosis were included in the algorithm development so it is likely that many symptoms and smoking records included were those after the point at which clinical lung cancer investigations were already underway and a diagnosis of lung cancer was actively being sought by the GPs. Our study has shown that in the 4-month period leading up to lung cancer diagnosis, the majority of patients with lung cancer start undergoing investigations in general practice. Therefore, it follows that the model developed by Hippisley-Cox and Coupland[22] will be predicting lung cancer in patients who are already being investigated in general practice and hence it is of limited value in diagnosing lung cancer at an earlier stage.

In conclusion, a combination of patients' socioeconomic characteristics, smoking status and early-stage symptoms appear to aid earlier identification of patients who are at an increased risk of lung cancer and who will benefit from further investigations such as chest x-rays. The weighting and inclusion of sociodemographic variables—age, sex, socioeconomic status and smoking history—and the weighting and inclusion of other clinical diagnoses—URTI, LRTI, non-specific chest infections, COPD and the frequency of general practice consultations—make our model a huge improvement on the NICE list[15] of symptoms. Evidence from past research has shown that a delay of 18–131 days (median of 54 days) between diagnosis and curative treatment for lung cancer was associated with an increase in cross-sectional tumour size and an increased risk of the cancer becoming incurable.[30] The outcomes of lung cancer are likely to be better in patients referred earlier and whose disease is diagnosed earlier because they may have earlier-stage disease and better performance status. A clinical trial, perhaps in conjunction with a screening trial, is needed to fully quantify the benefit of the model in practice.

CLINICAL IMPLICATIONS

There are several potential ways of applying this model clinically. Our primary aim is to develop the algorithm into a programme which could be incorporated into GP software and used by GPs to provide a rational estimate of patients' lung cancer risk during consultation. For example, if a patient presents with symptoms such as cough, chest pain and a history of weight loss, the GP with the aid of this algorithm, can calculate an estimate of the patient's risk of developing lung cancer taking account of the patient's background risk factors in addition to the current presenting symptoms and other clinical data within a preceding time frame. By incorporating the model into GP computer software, these risk assessments would not need to be directly calculated by GPs. Similar methods are already being used for the calculation of cardiovascular disease risk and the benefits of this as opposed to GPs working out the lung cancer risk for individual patients is that rather than making a risk estimation based on information collected by the GP during a consultation, the system takes account of all previous recorded data for patients, including records entered during consultation with other GPs in the same practice.

Another potential means of applying this model is by making the algorithm widely available to the general population to enable individuals to estimate their own risk of developing lung cancer. This could ultimately encourage earlier symptom presentation to general practice by high-risk patients who, following an assessment of their lung cancer risk, recognise the need for further investigation.

Contributors RH, LT and BI-O conceived the idea for and designed this study. DB provided advice on the study and extensively edited this paper. CS extracted the THIN data and ensured its accuracy. BI-O performed the statistical analysis and wrote the first draft of the manuscript. All authors critically revised and approved the final manuscript.

Funding This piece of research was funded by a PhD studentship from the Economic and Social Research Council, held by Barbara Iyen-Omofoman.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 WHO International Agency for Research on Cancer. *GLOBOCAN 2008: Lung Cancer Incidence, Mortality and Prevalence Worldwide in 2008*. Lyon: WHO International Agency for Research on Cancer, 2010. http://globocan.iarc.fr. accessed 31 Dec 2012.
2 Ferlay J, Parkin DM, Steliarova-Foucher E. Estimates of cancer incidence and mortality in Europe in 2008. *Eur J Cancer* 2010;46:765–81.
3 Janssen-Heijnen MLG, Gatta G, Forman D, *et al*. Variation in survival of patients with lung cancer in Europe, 1985–1989. *Eur J Cancer* 1998;34:2191–6.
4 Holmberg L, Sandin F, Bray F, *et al*. National comparisons of lung cancer survival in England, Norway and Sweden 2001–2004: differences occur early in follow-up. *Thorax* 2010;65:436–41.
5 Imperatori A, Harrison RN, Leitch DN, *et al*. Lung cancer in Teesside (UK) and Varese (Italy): a comparison of management and survival. *Thorax* 2006;61:232–9.
6 Mountain CF. Revisions in the international system for staging lung cancer. *Chest* 1997;111:1710–17.
7 Richards MA. The national awareness and early diagnosis initiative in England: assembling the evidence. *Br J Cancer* 2009;101(Suppl 2):S1–4.
8 Infante MV, Pedersen JH. Screening for lung cancer: are we there yet? *Curr Opin Pulm Med* 2010;16:301–6.
9 Pedersen JH, Ashraf H, Dirksen A, *et al*. The Danish randomized lung cancer CT screening trial—overall design and results of the prevalence round. *J Thorac Oncol* 2009;4:608–14.
10 Baldwin DR, Duffy SW, Wald NJ, *et al*. UK Lung Screen (UKLS) nodule management protocol: modelling of a single screen randomised controlled trial of low-dose CT screening for lung cancer. *Thorax* 2011;66:308–13.
11 Aberle DR, Adams AM, Berg CD, *et al*. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011;365:395–409.
12 Bjerager M, Palshof T, Dahl R, *et al*. Delay in diagnosis of lung cancer in general practice. *Br J Gen Pract* 2006;56:863–8.
13 Corner J, Hopkinson J, Fitzsimmons D, *et al*. Is late diagnosis of lung cancer inevitable? Interview study of patients' recollections of symptoms before diagnosis. *Thorax* 2005;60:314–19.
14 Hamilton W, Peters TJ, Round A, *et al*. What are the clinical features of lung cancer before the diagnosis is made? A population based case-control study. *Thorax* 2005;60:1059–65.

15 National Institute for Health and Clinical Excellence. The diagnosis and treatment of lung cancer: Methods, evidence and guidance 2005. http://www.nice.org.uk/nicemedia/pdf/cg024fullguideline.pdf. accessed 31 Dec 2012.
16 Hamilton W, Sharp D. Diagnosis of lung cancer in primary care: a structured review. *Fam Pract* 2004;21:605–11.
17 Allgar VL, Neal RD, Ali N, *et al*. Urgent GP referrals for suspected lung, colorectal, prostate and ovarian cancer. *Br J Gen Pract* 2006;56:355–62.
18 Bach PB, Kattan MW, Thornquist MD, *et al*. Variations in lung cancer risk among smokers. *J Natl Cancer Inst* 2003;95:470–8.
19 Spitz MR, Hong WK, Amos CI, *et al*. A risk model for prediction of lung cancer. *J Natl Cancer Inst* 2007;99:715–26.
20 Spitz MR, Etzel CJ, Dong Q, *et al*. An expanded risk prediction model for lung cancer. *Cancer Prev Res (Phila)* 2008;1:250–4.
21 Cassidy A, Myles JP, van Tongeren M, *et al*. The LLP risk model: an individual risk prediction model for lung cancer. *Br J Cancer* 2008;98:270–6.
22 Hippisley-Cox J, Coupland C. Identifying patients with suspected lung cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract* 2011;61:715–23.
23 CSD EPIC. *THIN Data from EPIC: A Guide for Researchers*. London: CSD EPIC, July 2009.
24 Bourke A, Dattani H, Robinson M. Feasibility study and methodology to create a quality-evaluated database of primary care data. *Inform Prim Care* 2004;12:171–7.
25 Iyen-Omofoman B, Hubbard RB, Smith CJ, *et al*. The distribution of lung cancer across sectors of society in the United Kingdom: a study using national primary care data. *BMC Public Health* 2011;11:857.
26 Falcoz PE, Conti M, Brouchet L, *et al*. The Thoracic Surgery Scoring System (Thoracoscore): risk model for in-hospital death in 15,183 patients requiring thoracic surgery. *J Thorac Cardiovasc Surg* 2007;133:325–32.
27 Freedman AN, Seminara D, Gail MH, *et al*. Cancer risk prediction models: a workshop on development, evaluation, and application. *J Natl Cancer Inst* 2005;97:715–23.
28 Saghir Z, Dirksen A, Ashraf H, *et al*. CT screening for lung cancer brings forward early disease. The randomised Danish lung cancer screening trial: status after five annual screening rounds with low-dose CT. Thorax 2012;67:296–301.
29 D'Amelio AM Jr, Cassidy A, Asomaning K, *et al*. Comparison of discriminatory power and accuracy of three lung cancer risk models. *Br J Cancer* 2010;103:423–9.
30 O'Rourke N, Edwards R. Lung cancer treatment waiting times and tumour growth. *Clin Oncol* 2000;12:141–4.