Identifying patients with suspected lung cancer in primary care: derivation and validation of an algorithm

Abstract
Background
Lung cancer has one of the lowest survival outcomes of any cancer because more than two-thirds of patients are diagnosed when curative treatment is not possible. The challenge is to help earlier diagnosis of lung cancer and hence improve prognosis.

Aim
To derive and validate an algorithm incorporating information on symptoms, to estimate the absolute risk of having lung cancer.

Design and setting
Cohort study of 375 UK QResearch® general practices for development, and 189 for validation.

Method
Selected patients were aged 30–84 years and free of lung cancer at baseline and haemoptysis, loss of appetite, or weight loss in previous 12 months. Primary outcome was incident diagnosis of lung cancer recorded in the next 2 years. Risk factors examined were: haemoptysis, appetite loss, weight loss, cough, dyspnoea, tiredness, hoarseness, smoking, body mass index, deprivation score, family history of lung cancer, other cancers, asthma, chronic obstructive airways disease, pneumonia, asbestos exposure, and anaemia. Cox proportional hazards models with age as the underlying time variable were used to develop separate risk equations in males and females. Measures of calibration and discrimination assessed performance in the validation cohort.

Results
There were 3785 incident cases of lung cancer arising from 4,289,282 person-years in the derivation cohort. Independent predictors were haemoptysis, appetite loss, weight loss, cough, body mass index, deprivation score, smoking status, chronic obstructive airways disease, anaemia, and prior cancer (males only). On validation, the algorithms explained 72% of the variation. The receiver operating characteristic (ROC) statistics were 0.92 for both females and males. The D statistic was 3.25 for females and 3.29 for males. The 10% of patients with the highest predicted risks included 77% of all lung cancers diagnosed over the subsequent 2 years.

Conclusion
The algorithm has good discrimination and calibration and could potentially be used to identify those at highest risk of lung cancer, to facilitate early referral and investigation.

Keywords
diagnosis, lung cancer, primary care; reseach, risk prediction, symptoms.

INTRODUCTION
Lung cancer is the most common cancer worldwide, with 1.3 million new cases diagnosed every year. It has one of the lowest survival outcomes of any cancer because over two-thirds of patients are diagnosed when curative treatment is not possible. In addition to preventing lung cancer by promoting smoking cessation, the challenge is to help earlier diagnosis, since prognosis varies according to the stage of cancer at presentation.

Earlier diagnosis of lung cancers could be improved by a combination of systematic screening of high-risk individuals using spiral CT (computerised tomography) scanning, particularly where there is likely to be a favourable benefit-to-harm ratio, and by facilitating the earlier investigation and referral of high-risk symptomatic individuals who present to their family physician. Cancer symptoms present a very real challenge for family physicians, since the symptoms can be common and non-specific, making it difficult to reliably distinguish patients who need further investigation from those who can be reassured.

While smoking is a well-established major risk factor for lung cancer, a significant proportion of cancers develop in non-smokers, and not all long-term heavy smokers develop lung cancer, suggesting that other factors also play an important role. Evidence suggests that age, deprivation, previous diagnoses of other cancers, previous pneumonia, family history of lung cancer, and asbestos exposure also increase long-term risk independently of smoking. In addition, ‘red-flag’ symptoms such as haemoptysis, loss of appetite, dyspnoea, and cough might herald an existing condition of lung cancer, especially among individuals with a high baseline risk. However, an approach that focuses on individual ‘red-flag’ symptoms such as haemoptysis without taking account of other risk factors is likely to miss 80% of current lung cancers. A variety of factors, therefore, need to be combined to develop a risk-prediction algorithm to help clinicians better assess and prioritise patients at high risk of having lung cancer, for further investigation or referral. While the case for such models is accepted, and some models that estimate long-term risk have been published, there are no models that combine baseline risk and symptoms.

This study aimed to develop and validate an algorithm to estimate the individualised absolute risk of having lung cancer, incorporating both symptoms and baseline risk factors, to help identify those at highest risk for further investigation or referral. QResearch® (a large UK primary care database) was used to develop the risk-prediction models, since it contains robust data on many of the relevant exposures and outcomes. It is also representative of the population where such a model is likely to be used and has been used successfully to develop and validate a range of prognostic models for use in primary care. Once validated, the prediction models could be
integrated into clinical computer systems to help systematically identify those at high risk, and alert clinicians to those who might benefit most from further assessment or interventions. 16,18

METHOD
Study design and data source
A prospective cohort study was carried out in a large population of primary care patients from an open cohort study, using the QRResearch database (version 30). The study included all practices in England and Wales that had been using their EMIS® (Egton Medical Information Systems) computer system for at least a year. Two-thirds of practices were randomly allocated to the derivation dataset and the remaining one-third to a validation dataset. An open cohort of patients was identified aged 30–84 years, drawn from patients registered with practices between 1 Jan 2000 and 30 September 2010. The study excluded patients without a postcode-related Townsend score, patients with a history of lung cancer at baseline, and those with a first recorded ‘red-flag’ symptom in the 12 months prior to baseline; that is, symptoms of haemoptysis, appetite loss, weight loss, cough, body mass index, deprivation score, smoking status, chronic obstructive airways disease, anaemia, and prior cancer (females only). It has good discrimination and calibration and could be used to identify those at highest risk for early referral and investigation.

How this fits in
Lung cancer is the most common cancer worldwide and has poor survival, since many cancers are diagnosed late when curative treatment is not possible. Symptoms that might herald a diagnosis of lung cancer are common and non-specific, making it difficult for GPs to identify high-risk patients. The QLung® cancer algorithm developed in this study includes age, haemoptysis, appetite loss, weight loss, cough, body mass index, deprivation score, smoking status, chronic obstructive airways disease, anaemia, and prior cancer (females only). It has good discrimination and calibration and could be used to identify those at highest risk for early referral and investigation.

Clinical outcome definition
The study outcome was incident diagnosis of lung cancer during the subsequent 2 years, recorded either on the patient’s GP record using the relevant UK diagnostic codes or on their linked Office for National Statistics (ONS) cause-of-death record, using the relevant International Classification of Diseases (ICD)-9 codes or ICD-10 diagnostic codes (codes available from the authors). A 2-year follow-up was used, since this represents the period of time during which existing lung cancers are likely to become clinically manifest. 13,21 It was assumed that where lung cancer deaths occurred within 2 years, without a recorded diagnostic code in the GP record, the cancer would have been present at the start of the 2-year period.

Predictor variables
Established predictor variables were examined, focusing on those that are likely to be recorded in the patient’s electronic record and that the patient themself is likely to know. Three ‘red-flag’ symptoms were also included (haemoptysis, loss of appetite, and weight loss) as well as other symptoms that might herald a diagnosis of lung cancer. Separate analyses were carried out in males and females, and age was accounted for by using it as the underlying time variable in the analyses. The predictor variables examined were:

- currently consulting GP with first onset of haemoptysis (yes/no); 12
- currently consulting GP with first onset of loss of appetite (yes/no); 12
- currently consulting GP with first onset of weight-loss symptom (yes/no); 12
- recently consulted GP with first onset of any of:
  - cough in the past 12 months (yes/no); 12
  - dyspnoea in the past 12 months (yes/no); 12
  - tiredness in the past 12 months (yes/no);
  - hoarseness in the past 12 months (yes/no);
- body mass index (BMI, continuous);
- smoking status [non-smoker; ex-smoker; light smoker [1–9 cigarettes/day]; moderate smoker [10–19 cigarettes/day]; heavy smoker [≥20 cigarettes/day]], 11,14,15,22
- chronic obstructive airways disease diagnosed ever (yes/no), 8,14
- Townsend deprivation score (continuous), 23
- family history of lung cancer (yes/no), 8,14
Variables were included in the final model if they had a hazard ratio of $<0.80$ or $>1.20$ (for binary variables) and were statistically significant at the 0.01 level. Tests were also carried out for interactions between smoking and deprivation.

Derivation and validation of the models

Multiple imputation was used to replace missing values for smoking status and BMI. Fractional polynomials were used to model non-linear risk relations with BMI. Derivation cohort Validation cohort (n = 2 406 127) (n = 1 267 151)

Table 1. Baseline characteristics of patients in the derivation and validation cohorts. (Figures in the tables are number [%] unless otherwise specified)

| Sex         | Derivation cohort | Validation cohort |
|-------------|-------------------|-------------------|
| Female      | 1 205 833 (50.1)  | 634 629 (50.1)    |
| Male        | 1 200 294 (49.9)  | 632 522 (49.9)    |
| Mean age (SD), years | 49.7 (14.8)          | 49.6 (14.8)     |
| Mean Townsend score (SD) | -0.3 (3.4)                      | -0.1 (3.6)    |

BMI

| Derivation cohort | Validation cohort |
|-------------------|-------------------|
| BMI recorded prior to study entry | 1 585 199 (65.9) | 860 819 (67.9) |
| Mean BMI in kg/m^2 (SD) | 26.4 (4.6) | 26.4 (4.7) |

Smoking status

| Derivation cohort | Validation cohort |
|-------------------|-------------------|
| Non smoker        | 1 231 186 (51.2)  | 644 915 (50.9) |
| Ex-smoker         | 417 267 (17.3)    | 223 718 (17.7) |
| Current smoker, amount not recorded | 73 325 (3.0) | 40 439 (3.2) |
| Light smoker (<10/day) | 153 421 (6.4) | 82 973 (6.5) |
| Moderate smoker (10–19/day) | 187 856 (7.8) | 100 354 (7.9) |
| Heavy smoker (≥20/day) | 141 040 (5.9) | 77 679 (6.1) |
| Smoking status not recorded | 202 030 (8.4) | 97 073 (7.7) |

Medical history

| Derivation cohort | Validation cohort |
|-------------------|-------------------|
| Family history of lung cancer | 13 378 (0.6) | 8078 (0.6) |
| Prior cancer (apart from lung cancer) | 48 548 (2.0) | 25 547 (2.0) |
| Asthma | 182 800 (7.6) | 99 417 (7.8) |
| Chronic obstructive airways disease | 37 191 (1.5) | 21 458 (1.7) |
| Pneumonia | 31 840 (1.3) | 17 238 (1.4) |
| Asbestos exposure | 1649 (0.1) | 928 (0.1) |

Current symptoms and symptoms in the preceding year

| Derivation cohort | Validation cohort |
|-------------------|-------------------|
| Current haemoptysis | 13 980 (0.6) | 8010 (0.6) |
| Current appetite loss | 11 853 (0.5) | 6303 (0.5) |
| Current weight-loss symptom | 30 937 (1.3) | 17 365 (1.4) |
| Cough in last year | 55 434 (2.3) | 30 298 (2.4) |
| Dyspnoea in last year | 11 549 (0.5) | 5887 (0.5) |
| Tiredness in last year | 22 779 (0.9) | 12 854 (1.0) |
| Hoarseness | 1748 (0.1) | 966 (0.1) |
| Haemoglobin recorded in the last year | 351 100 (14.6) | 189 945 (15.0) |
| Haemoglobin <11 g/dl in the last year | 13 980 (0.6) | 8010 (0.6) |

SD = standard deviation.
(4.9%) without a recorded Townsend deprivation score, 18 with missing dates for the diagnoses of lung cancer, 1490 (0.1%) with a history of lung cancer, and a further 6522 patients (0.3%) with at least one ‘red-flag’ symptom (haemoptysis, loss of appetite, or weight loss) recorded in the 12 months prior to entry to the study at baseline, leaving 2,606,127 patients for analysis.

A total of 1,342,329 patients aged 30–84 years were identified in the validation cohort, and the following were excluded: 70,847 patients (5.3%) without a recorded Townsend score, eight (<0.1%) without a recorded date of diagnosis of lung cancer, 1,185,3 w ith m issing data supported the use of multiple imputation to replace missing values for smoking and BMI (not shown, available from the authors).

### Table 3. Adjusted hazard ratios (95% CI) for the final model for lung cancer for males and females in the derivation cohort

| Symptoms presented to GP | Adjusted hazard ratios for females (95% CI) | Adjusted hazard ratios for males (95% CI) |
|--------------------------|--------------------------------------------|------------------------------------------|
| Current haemoptysis*     | 23.9 (20.6 to 27.6)                         | 21.5 (19.3 to 23.9)                      |
| Current appetite loss*    | 4.14 (3.15 to 5.54)                         | 4.71 (3.69 to 6.00)                      |
| Current weight loss*      | 4.52 (3.80 to 5.38)                         | 6.09 (5.33 to 6.95)                      |
| New onset cough in last 12 months* | 1.90 (1.56 to 2.32) | 1.47 (1.23 to 1.75) |
| Recorded haemoglobin<11 g/dl in last 12 months* | 1.75 (1.38 to 2.22) | 1.89 (1.54 to 2.32) |

| Smoking status            | Adjusted hazard ratios for females (95% CI) | Adjusted hazard ratios for males (95% CI) |
|---------------------------|--------------------------------------------|------------------------------------------|
| Non smoker                |                                            |                                           |
| Ex-smoker                 | 3.37 (2.83 to 4.01)                         | 2.13 (1.87 to 2.43)                      |
| Light smoker (<10/day)    | 6.57 (5.37 to 8.03)                         | 3.70 (3.20 to 4.27)                      |
| Moderate smoker (10–19/day) | 8.32 (7.05 to 9.82) | 4.95 (4.26 to 5.76) |
| Heavy smoker (≥20/day)    | 10.6 (8.49 to 13.2)                        | 6.35 (5.43 to 7.43)                      |
| Prior diagnosis other cancer except lung cancer* | 1.33 (1.09 to 1.63) | NS |
| Chronic obstructive airways disease* | 1.82 (1.57 to 2.11) | 1.51 (1.34 to 1.69) |
| Townsend deprivation score (5 unit increase) | 1.17 (1.08 to 1.27) | 1.17 (1.10 to 1.24) |

*Compared with person without this characteristic. NS = not significant. Hazard ratios were adjusted for all other terms in the table and models accounted for age as underlying time function and also included fractional polynomial terms for body mass index (BMI). For females, the terms were BMI<sup>0.5</sup>, ln(BMI). For males the terms were BMI<sup>0.3</sup>, BMI<sup>0.9</sup>/ln(BMI).
The risk of lung cancer in females was significantly associated with decreasing BMI, increasing deprivation, and amount smoked each day. For example, compared with non-smokers, the risks were increased by 10.6-fold for heavy smokers, 8.3-fold for moderate smokers, 6.6-fold for light smokers, and 3.4-fold for ex-smokers. Risks were also elevated among females with current haemoptysis (26.5-fold higher), current appetite loss (4.1-fold higher), current weight loss symptom (4.5-fold higher), cough in the last 12 months (1.9-fold higher), chronic obstructive airways disease (1.8-fold higher), recorded Hb in the last year (1.6-fold higher), and a prior diagnosis of another cancer (1.3-fold higher). The other variables examined were not independent risk factors in females, so were not included in the final model.

The final model for males was similar to that for females, except that it did not include history of another cancer. Prior history of cancer was significant for males on univariate analysis (unadjusted hazard ratio = 4.3, 95% CI = 3.6 to 5.1), but not after adjustment for other factors in the model. The magnitudes of the hazard ratios were generally similar to those found for females, apart from smoking, where the hazard ratios for males were lower than those for females.

Validation

The validation statistics (Table 4) showed that the risk-prediction equations explained 71.7% (95% CI = 70.3 to 73.1) of the variation in time to diagnosis in females and 72.1% of the variation in males (95% CI = 71.0 to 73.2). The D statistic was 3.25 (95% CI = 3.15 to 3.37) for females and 3.29 (95% CI = 3.20 to 3.38) for males. The ROC statistics were 0.92 (95% CI = 0.91 to 0.93) for both females and males. Figure 1 shows the mean predicted scores and the observed risks at 2 years within each tenth of predicted risk, in order to assess the calibration of the model in the validation cohort. There was close correspondence between predicted and observed 2-year risks within each model tenth, indicating that the algorithm was well calibrated.

Individual risk assessment and thresholds

One potential use for this algorithm is within consultations with individual patients, particularly if they present with new onset of haemoptysis or unexplained anaemia. The results could help inform the decision to undertake further investigation such as a chest X-ray or spiral CT, and/or the degree of urgency for referring the patients to secondary care. Some clinical examples are shown in Box 1.

Since this is a new algorithm, there are no established thresholds for defining high-risk groups. A range of centiles of predicted risk were calculated from the validation population, to define a high-risk group (that is, the top 0.5%, 1%, 5%, and 10% at highest risk) for males and females combined. The numbers and proportion of incident cases in the validation cohort that fell within each category of risk were then determined.

The 90th centile defined a high-risk group with a 2-year risk score of >0.37% (Table 5). There were 1697 new cases of lung cancer within this group, out of 2196 new cases identified in the validation cohort over 2 years, which accounted for 77.3% of all new cases of lung cancer (sensitivity). The positive predictive value (PPV) with this threshold was 1.3%. Alternatively, using a threshold based on the top 0.5% of risk had a sensitivity of 27.4% and a PPV of 9.5%. In contrast, only 18.4% of lung cancers occurred in patients aged 40 years and over presenting with a first onset of haemoptysis, who were current or ex-smokers (in other words, the sensitivity of this approach is low and approximately 82% of cases of lung cancer cases would be missed). The PPV in this group was 9.7%. Only 23.0% of lung cancer cases occurred in patients with haemoptysis, and the PPV for haemoptysis was 6.4%.

**DISCUSSION**

*Summary*

This research has developed and validated a
Box 1. Clinical examples

- A 78-year-old female who is an ex-smoker with a BMI of 25.7 kg/m² and has a history of chronic obstructive airways disease, who presents to the GP with haemoptysis and has had a cough and a Hb<11 g/dl recorded in the last 12 months, has an estimated risk of 37% of having existing lung cancer as yet undiagnosed. If the patient also has loss of appetite and weight loss, the estimated risk increases to 76%. Although this patient is an ex-smoker, she is at particularly high estimated risk of having lung cancer and therefore would warrant an urgent referral for further investigation.

- A 67-year-old male who is a heavy smoker who presents with haemoptysis, loss of appetite, and weight loss but who has not presented to the GP with a cough or haemoptysis, has a 29% estimated risk of having existing lung cancer as yet undiagnosed. While this patient does not have the ‘red-flag’ symptom of haemoptysis, the other factors that are present place him into a high-risk category likely to need urgent referral or investigation.

- A 40-year-old male who has a BMI of 27.5 kg/m² who is a heavy smoker who presents with haemoptysis but no other symptoms and no evidence of anaemia, has a 0.2% estimated risk of having existing lung cancer as yet undiagnosed.

- A 50-year-old male who is a non-smoker and presents to the GP with haemoptysis, loss of appetite, and weight loss, and has had a cough and Hb<11 g/dl recorded in the last 12 months, has a 28% estimated risk of having existing lung cancer as yet undiagnosed.

The study has good face validity, since it has been conducted in the setting where the majority of patients in the UK are assessed, treated, and followed-up.

The algorithms have been developed in one cohort and validated in a separate cohort that is representative of the patients likely to be considered for preventative measures. While other risk-prediction models for lung cancer have been developed, none can be directly compared since none include symptoms. Limitations of the study include lack of formally adjudicated outcomes, information bias, and missing data. The database has linked cause of death from the UK ONS and the study is therefore likely to have picked up the majority of cases of lung cancer, thereby minimising ascertainment bias. Patients who die of lung cancer in hospital will be included on the linked cause-of-death data. Patients diagnosed with lung cancer in hospital will have the information recorded in hospital discharge letters, which are sent to the GP and this information is then entered into the patient’s electronic record. The incidence rate in the study population was close to published national data, with similar patterns by age and sex.31 While the study was reliant on the accuracy of information recorded by primary care physicians, the quality of information is likely to be good since previous studies have validated similar outcomes and exposures using questionnaire data, and found levels of completeness and accuracy in similar GP databases to be good.32,33 For example, one systematic review reported that on average 89% of diagnoses recorded on the GP electronic record are confirmed from other data sources.33,35 However, one significant limitation of this study is that the stage of lung cancer at diagnosis is not recorded in either the GP record or the linked cause-of-death record. Additional data from cancer registries would need to be linked to the GP

| Table 5. Comparison of strategies to identify patients at risk of having a diagnosis of lung cancer in the next 2 years based on the validation cohort |
|-----------------------------------------------|
| Risk threshold at 2 years, % | Number with criterion | Number of patients with criterion AND lung cancer | Total number of new diagnoses of lung cancer | Sensitivity, % (95% CI) | Positive predictive value, % (95% CI) |
|-----------------------------------------------|
| Haemoptysis | n/a | 7861 | 504 | 2196 | 23.0 (21.1 to 24.8) | 8.4 (5.9 to 10.7) |
| ≥40 years AND haemoptysis AND current or ex-smoker | n/a | 4144 | 404 | 2196 | 18.4 (16.8 to 20.1) | 13.7 (11.7 to 15.7) |
| Top 10% risk score | 0.97 | 126,672 | 174 | 2196 | 77.3 (75.5 to 79.0) | 3.4 (2.6 to 2.4) |
| Top 5% risk score | 0.68 | 63,336 | 1377 | 2196 | 62.7 (61.6 to 64.7) | 2.2 (2.1 to 2.3) |
| Top 1% risk score | 0.21 | 12,667 | 796 | 2196 | 36.2 (34.2 to 38.2) | 6.3 (5.9 to 6.7) |
| Top 0.5% risk score | 0.47 | 6333 | 602 | 2196 | 27.4 (25.6 to 29.3) | 9.5 (8.8 to 10.3) |

n/a = not applicable.
record. This is not currently available, although work is in progress to undertake this linkage so it will be available for future versions of this tool. Also, there is no evidence from the present study about whether use of this symptom-based tool is likely to lead to earlier identification of lung cancer at a stage when curative treatment (that is, surgery) is more likely to be possible. A cluster randomised clinical controlled trial comparing use of this tool in intervention general practices against usual practice in control practices could help answer such a question.

Another limitation of the study is that recording of symptoms may be less complete or accurate than diagnostic codes, since patients might not visit their GP with mild symptoms, and may not report all symptoms to their GP when they do consult, or GPs might not record all the symptoms in the electronic health record. The effect of this information or recording bias would be to overinflate the hazard ratios if they relate to more severe symptoms (for example, major loss of appetite) or underestimate the hazard ratio if patients with the symptoms do not have them recorded. Similarly, family history of lung cancer might be underestimated, since it is not routinely assessed and recorded in GP records. Lastly, it is possible that some patients might misreport their smoking habits to their GP. For example, smoking status was defined on the basis of self-report, and the definition of an ex-smoker is a patient whose last recorded smoking status was as an ex-smoker, regardless of when they stopped smoking. Some ex-smokers may consider themselves as a never-smoker after many years have elapsed. If this were to occur, then it would tend to bias the hazard ratios for ex-smoker towards one.

Comparison with existing literature
The study is based on a large representative primary care population. While other studies have examined chronic risk factors, symptoms separately,12,13,36 to the authors’ knowledge, this is the first study to produce a measure of absolute risk of current lung cancer based on a combination of symptoms (haemoptysis, appetite loss, weight loss, and cough) as well as demographic information, anaemia, BMI, smoking status, chronic obstructive airways disease, and prior cancer (in females). The significance of prior cancer as a risk factor in females but not males is of interest and deserves further study. The direction and magnitude of the hazard ratios in the present study for smoking status and history of another cancer are comparable to those reported in other studies.27,38 The algorithm performed well in a separate validation sample, with good discrimination and calibration. It can identify the 10% of the population in which approximately 76% of all new lung cancer cases are likely to be diagnosed over the next 2 years.

Comparison of published discrimination statistics suggests the new model performs well. The ROC values were 0.92 in males and females, which is substantially higher than for the model by Spitz and coworkers, with biomarkers [ROC of 0.73],39 and the Liverpool Lung Project [ROC value of 0.71].9 The Bach et al model is based on a person’s age, sex, and smoking history but only applies for individuals aged 50–75 years who have smoked 10–60 cigarettes/day for 25–35 years.22 The expanded Spitz et al model includes more variables — environmental tobacco smoke, family history of cancer, dust exposure, prior respiratory disease, and smoking history variables — but requires genetic testing, which is unavailable in the dataset for the present study, and unlikely to be available for routine clinical use.15

Implications for research and practice
One practical mechanism to help improve clinical recording of family history and symptoms for future studies would be to introduce electronic templates into GP clinical systems, which are displayed when a ‘red-flag’ symptom is recorded in the patient’s record. The template would then help structured data entry of other related symptoms including significant negative findings. Over time, this would improve the accuracy and completeness of the electronic record and hence the underlying data used for future versions of this algorithm.

The algorithm has a number of potential clinical applications. First, it could be used to help inform the revision of NICE guidance on the investigation and referral of patients with suspected cancer.20 For example, current NICE guidance recommends an urgent referral for a chest X-ray in patients with persistent symptoms such as haemoptysis, chest pain, dyspnoea, cough, or weight loss, but not for appetite loss, although this study has demonstrated a four- to fivefold increase in risk of cancer with this symptom, independently of other factors. Urgent referral is recommended by NICE for persistent haemoptysis in smokers or ex-smokers who are aged 40 years and older, or those whose chest X-ray is suggestive of lung
cancer. An approach based on haemoptysis in smokers or ex-smokers aged 40 years and older alone is likely to miss over 80% of lung cancers. Alternatively, an approach could be developed based on a risk estimate derived from the new algorithm, which might include the possibility of spiral CT or referral for a high-risk patient even in the presence of a normal chest X-ray. Another possible application is to automatically calculate risk scores for every patient aged 30–84 years registered with a practice, by running a programme within the clinical computer system. This could then generate a rank-ordered list of high-risk patients who need to be recalled for further assessment or investigation. This process might also identify patients with ‘red-flag’ symptoms who have not been investigated already or for whom no diagnosis has been found. The risk score might also be useful for providing patients with a realistic estimation of their risk of lung cancer pending the result of the investigations, which for many patients may actually be reassuring if the absolute risk is low. A modified version of the web calculator could be developed for use by patients themselves, which would prompt symptomatic or high-risk patients to attend their GP for further assessment.
REFERENCES

1. Ferlay J, Autier P, Boniol M, et al. Estimates of the cancer incidence and mortality in Europe in 2006. Ann Oncol 2007; 18(3): 581–592.

2. The Information Centre. National Lung Cancer Audit. Key findings about the quality of care for people with lung cancer in England incorporating headline and completeness data from Wales. Report for the audit period 2006. Leeds: The Information Centre, 2006. http://www.ic.nhs.uk/webfiles/Services/NCASP/Cancer/Lung%20cancer%2006.pdf (accessed 13 Jul 2011).

3. van Rens MTM, Brutel de la Rivière AB, Elbers HJ, van den Bosch JMM. Prognostic assessment of 2361 patients who underwent pulmonary resection for non-small cell lung cancer, stage I, II, and IIIA. Chest 2000; 117(2): 374–379.

4. Field JK, Raji OY. The potential for using risk models in future lung cancer screening trials. F1000 Med Rep 2010; 2: pii: 38.

5. Silvestri GA, Alberg AJ, Ravenel J. The changing epidemiology of lung cancer with a focus on screening. BMJ 2009; 339: b3053.

6. US National Cancer Institute. Lung cancer trial results show mortality benefit with low-dose CT. Bethesda: US National Cancer Institute, 2010.

7. Peto R, Darby S, Deo H, et al. Smoking, smoking cessation, and lung cancer risk: modeling total exposure and intensity. Am J Epidemiol 2007; 166(4): 479–489.

8. Cassidy A, Myles JP, van Tongeren M, et al. The LLP risk model: an individual risk prediction model for lung cancer. Br J Cancer 2007; 98(2): 270–276.

9. Crispo A, Breneman P, Jezek KH, et al. The cumulative risk of lung cancer among current, ex- and never-smokers in European men. Br J Cancer 2004; 91(7): 1280–1286.

10. Subramanian J, Gowdian R. Lung cancer in never smokers. J Clin Oncol 2007; 25(5): 9.

11. Lubin JH, Alavanja MCR, Caporaso N, et al. Cigarette smoking and cancer risk: modeling total exposure and intensity. Ann J Epidemiol 2007; 166(4): 479–489.

12. Hamilton W, Peters TJ, Round A, Sharp D. What are the clinical features of lung cancer before the diagnosis is made? A population based case-control study. Thorax 2005; 60(12): 1059–1065.

13. Jones R, Lavinio C, Charlton J, Guilfoil MC. Alarm symptoms in early diagnosis of cancer in primary care: cohort study using General Practice Research Database. BMJ 2007; 334(7602): 1040.

14. Spitz MR, Hong WK, Ames CP, et al. A risk model for prediction of lung cancer. J Natl Cancer Inst 2007; 99(19): 715–726.

15. Spitz MR, Etzel CJ, Dong G, et al. Expanded risk prediction model for lung cancer. Cancer Prev Res 2008; 1(4): 250–254.

16. Hippisley-Cox J, Coupland C, Robson J, et al. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. BMJ 2009; 338: b680.

17. Hippisley-Cox J, Coupland C. Predicting risk of osteoporotic fracture in men and women in England and Wales: prospective derivation and validation of QFRactureScores. BMJ 2009; 339: b4229.

18. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ 2008; 336(7659): 1475–1482.

19. Hippisley-Cox J, Coupland C. Predicting the risk of chronic kidney disease in men and women in England and Wales: prospective derivation and external validation of the QKidney® scores. BMC Fam Pract 2010; 11: 49.

20. Hippisley-Cox J, Coupland C. Individualising the risks of statins in men and women in England and Wales: population based cohort study. Heart 2010; 96(11): 939–947.

21. Jones R, Charlton J, Lavinio C, Guilfoil MC. Alarm symptoms and identification of non-cancer diagnoses in primary care: cohort study. BMJ 2009; 339: b3094.

22. Bach PB, Kattan MW, Thornquist MD, et al. Variations in lung cancer risk among smokers. J Natl Cancer Inst 2003; 95(4): 470–478.

23. National Cancer Intelligence Network. Lung cancer incidence by deprivation in England, 1995–2004. London: National Cancer Intelligence Network, 2008.

24. De Vos Irene H, Lamont DW, et al. Asbestos and lung cancer in Glasgow and the west of Scotland. BMJ 1993; 306(6891): 1503–1506.

25. Royston P. Multiple imputation of missing values. Stata J 2004; 4(3): 227–241.

26. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. Int J Epidemiol 1999; 28(5): 964–974.

27. Hippisley-Cox J, Coupland C, Robson J, Brindle P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. BMJ 2010; 341: c6624.

28. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Stat Med 2006; 25(5): 723–748.

29. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat Med 1998; 17(8): 857–872.

30. National Institute for Health and Clinical Excellence. Referral guidelines for suspected cancer. London: National Institute for Health and Clinical Excellence, 2005.

31. Cancer Research UK. Lung Cancer — UK incidence statistics. London: Cancer Research UK, 2010.

32. Pick H, Jick LS, Derby LE. Validation of information recorded on general practitioner based computerised data resource in the United Kingdom. BMJ 1991; 302(6779): 766–768.

33. Herrett E, Thomas SL, Schoonen WM, et al. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. Br J Clin Pharmacol 2010; 69(11): 4–14.

34. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. Br J Gen Pract 2010; 60(572): e126–136.

35. Pick H, Jick LS, Derby LE, et al. Calcium-channel blockers and risk of cancer. Lancet 1997; 349(9066): 525–528.

36. Hamilton W. The CAPER studies: five case-control studies aimed at identifying and quantifying the risk of cancer in symptomatic primary care patients. Br J Cancer 2009; 101(Suppl 2): S80–86.

37. Kabat GC. Previous cancer and radiotherapy as risk factors for lung cancer in lifetime non-smokers. Cancer Causes Control 1993; 4(5): 489–495.

38. Mery CM, Pappas AN, Bueno R, et al. Relationship between a history of antecedent cancer and the probability of malignancy for a solitary pulmonary nodule. Chest 2004; 125(4): 2175–2181.