Development and validation of a model to predict the 10-year risk of general practitioner-recorded COPD

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters
Development and validation of a model to predict the 10-year risk of general practitioner-recorded COPD

Daniel Kotz1,2, Colin R Simpson2, Wolfgang Viechtbauer3, Onno CP van Schayck1,2 and Aziz Sheikh1,2,4

BACKGROUND: There is increasing interest in the earlier detection of, and intervention in, patients at highest risk of developing chronic obstructive pulmonary disease (COPD).

AIMS: The objective of this research was to develop and validate a risk prediction model for general practitioner (GP)-recorded diagnosis of COPD.

METHODS: We used data from 239 Scottish GP practices; two-thirds were randomly allocated to a derivation cohort and the other third to a validation cohort. We included patients aged 35–74 years at the cohort entry date, and excluded patients with a recorded diagnosis of COPD prior to the entry date and with missing data on smoking status.

RESULTS: There were 480,903 patients in the derivation cohort and 247,755 in the validation cohort. The incidence of COPD in the total cohort was 5.53/1,000 patient-years of follow-up (95% confidence interval (CI), 5.46–5.60). In the derivation cohort, the COPD risk for ever- versus never-smokers was substantially higher in women (hazard ratio (HR) = 9.61, 95% CI, 8.92–10.34) than in men (HR = 6.72, 95% CI, 6.19–7.30). Other risk factors for both sexes were level of deprivation and a previously recorded asthma diagnosis. In the validation cohort, the model discriminated well between patients who did and those who did not develop COPD: area under the receiver operating characteristics curve = 0.845 (95% CI, 0.840–0.850) for females and 0.832 (95% CI, 0.827–0.837) for males.

CONCLUSIONS: We have developed and validated the first risk prediction model for COPD, which has the major advantage of being populated entirely by routinely collected data and consequently may be used for clinical practice.

npj Primary Care Respiratory Medicine (2014) 24, Article number: 14011; doi:10.1038/npjpcrm.2014.11; published online 20 May 2014

INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is now one of the leading causes of chronic morbidity and mortality worldwide.1 The World Health Organization estimates that about 3.3 million people die from COPD across the world each year (i.e., 6% of all deaths).2 Considering the high and still increasing prevalence of COPD being seen in many parts of the world, this figure is expected to rise substantially in the coming decades.3–5 In developed countries, the health-care burden of COPD particularly affects primary care where most patients with COPD are managed.3–5 Still, even among primary care patients with known risk factors of the disease, the rate of undiagnosed COPD is high.6 With an increasing appreciation of the substantial morbidity and mortality resulting from COPD, there is growing international interest in the earlier detection of, and intervention in, patients with COPD; this more proactive approach is, however, currently hampered by the lack of clinically relevant, validated risk algorithms.

Smoking is the most important risk factor of COPD.1,8,9 Other risk factors that have been identified as playing a role in the aetiology of COPD include age, sex, socioeconomic status, childhood asthma, body mass index, acute respiratory infections, respiratory symptoms (such as cough and wheezing), occupational exposure to risk factors, exposure to biomass pollution indoors (an important risk factor in developing countries), family history of COPD, pulmonary tuberculosis, physical activity and alpha-1 anti-trypsin deficiency.1,10 However, there is still a limited understanding of these other risk factors, particularly with respect to their independent contribution to COPD risk and if there are differences in susceptibility between men and women.

There are currently no tools to predict the development of COPD in individuals free of the disease. Most available tools are only able to identify patients with already established, but undiagnosed, COPD (see e.g., refs 11,12). There is then a pressing need for an accurate and easy-to-use instrument that simultaneously takes account of a range of risk factors and accurately identifies individuals at increased risk of developing COPD, thereby offering the opportunity to target interventions in order to reduce morbidity and mortality.13 The aim of this study was to develop and validate a model for risk prediction of COPD using routinely collected data from a large national primary care data set.

MATERIALS AND METHODS

The Primary Care Clinical Informatics Unit at the University of Aberdeen collects data from General Practice Administration System for Scotland practices and includes almost 5 million patient-years of individual patient-level data.14,15 These practices have been shown to be representative of all Scottish practices.15 Furthermore, the database has high completeness and accuracy of morbidity data and recordings on patients’ tobacco use. Data extracted from 239 general practitioner (GP) practices contributing to the database were used in this analysis. Using the random sample function in

1Department of Family Medicine, CAPHRI School for Public Health and Primary Care, Maastricht University Medical Centre, Maastricht, The Netherlands; 2Allergy & Respiratory Research Group, Centre for Population Health Sciences, The University of Edinburgh, Edinburgh, UK; 3Department of Psychiatry and Psychology, MHeNS School for Mental Health and Neuroscience, Maastricht University, Maastricht, The Netherlands and 4Division of General Internal Medicine and Primary Care, Brigham and Women’s Hospital/ Harvard Medical School, Boston, MA, USA.

Correspondence: D Kotz (d.kotz@maastrichtuniversity.nl)
Received 18 November 2013; revised 10 January 2014; accepted 23 January 2014

© 2014 Primary Care Respiratory Society UK/Macmillan Publishers Limited
We performed an open cohort of patients aged 35–74 at the start date (1 April 1998), drawn from patients registered with the practices during the period from 1 April 1998 to 1 April 2008. An entry date to the cohort was defined for each patient as the latest of the following dates: 35th birthday, date of registration with the practice or start date of the cohort (1 April 1998). An exit date to the cohort was defined for each patient as the earliest of the following dates: date of first recorded diagnosis of COPD, date of deregistration with the practice, death or end date of the cohort (1 April 2008). Patients were excluded if they had a recorded diagnosis of COPD prior to the entry date or no recording of smoking status at any time. The number of patient-years of follow-up was calculated as the difference in years between each patient’s entry date and exit date into the cohort.

According to our study, smoking status, sex and the other above-mentioned risk factors with a model that in addition included the interaction term between smoking status and sex. The step to the model including the interaction term was statistically significant ($\chi^2 = 37.77$, d.f. = 1, $P < 0.001$). We therefore performed all analyses for men and women separately.

The primary analysis consisted of the following steps, conducted separately for men and women. In the derivation cohort, a multiple Cox proportional hazard regression model was used to estimate the coefficients and hazard ratios (HRs) of the potential risk factors for the primary outcome. On the basis of this model, a prognostic index (PI) was calculated for each patient from the derivation cohort as $P_{\text{der}} = \sum \hat{\beta}_i X_i$, where $\beta_i$ is the regression coefficient of the risk factor $X_i$ from the Cox model (this method was adapted from ref. 18). $P_{\text{der}}$ ranged from 0 (lowest risk) to 7.51 (highest risk) in males and from 0 to 7.48 in females and was transformed into a variable with 10 categories based on the deciles of $P_{\text{der}}$ (the values for calculating $P_{\text{der}}$ are presented in Supplementary Appendix I). We then calculated the 10-year incidence rate of COPD per interval in patients from the derivation cohort (1 = lowest incidence of COPD; 10 = highest), which we defined as the 10-year predicted incidence rate. $P_{\text{val}}$ for patients was then calculated from the validation cohort using the regression coefficients from the derivation cohort. $P_{\text{val}}$ was then again transformed into a variable with 10 intervals, but using the same cutoff points as in the derivation cohort. The 10-year observed incidence rate of COPD per interval in patients from the validation cohort was then determined. The accuracy of the risk prediction model in discriminating between patients from the validation cohort who developed COPD versus patients who did not was assessed by calculating the area under the receiver operating characteristics curve (ROCauc) for all values of $P_{\text{val}}$.

Finally, the prediction model was calibrated by plotting the predicted 10-year incidence of COPD against the observed incidence for each interval on $P_{\text{val}}$ in patients from the validation cohort.

In ancillary analyses, we deconstructed the prediction model and calculated the ROCauc, for models including only age, only smoking, and only age and smoking using the same method as described above. The purpose of this analysis was to compare the accuracy of the full prediction model (including all risk factors) with models including only the most important risk factors of COPD.

**RESULTS**

The total number of patients in the cohort was 728,658: 480,903 (66.0%) in the derivation cohort and 247,755 (34.0%) in the validation cohort. The median follow-up duration was 7.92 years (interquartile range: 3.76–10.00 years) and 7.88 years (3.77–10.00), respectively (Table 1).

During the study period there were 27,088 incident cases of COPD from 4.9 million patient-years of observation in the total cohort (Table 1), giving a crude incidence rate for COPD of 5.53 per 1,000 patient-years (95% confidence interval (CI), 5.46–5.60). The mean age at COPD diagnosis was 65.43 (s.d. = 9.73) years, with

<table>
<thead>
<tr>
<th>Table 1. Baseline characteristics of patients in the derivation and validation cohorts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Derivation cohort, N = 480,903</strong></td>
</tr>
<tr>
<td>Patient-years</td>
</tr>
<tr>
<td>Years follow-up, median (IQR)</td>
</tr>
<tr>
<td>COPD</td>
</tr>
<tr>
<td>Age, mean (s.d.)</td>
</tr>
<tr>
<td>Male sex</td>
</tr>
<tr>
<td>Ever smokers</td>
</tr>
<tr>
<td>Carstairs level of deprivation</td>
</tr>
<tr>
<td>1st Quintile (least deprived)</td>
</tr>
<tr>
<td>2nd Quintile</td>
</tr>
<tr>
<td>3rd Quintile</td>
</tr>
<tr>
<td>4th Quintile</td>
</tr>
<tr>
<td>5th Quintile (most deprived)</td>
</tr>
<tr>
<td>Prior asthma</td>
</tr>
</tbody>
</table>

Data are presented as column percentage (N), unless otherwise stated. Abbreviations: COPD, chronic obstructive pulmonary disease; IQR, interquartile range.
the risk of COPD being found to increase with age. This association was stronger in males than in females (Table 2). In both sexes, the risk of COPD increased with increasing socioeconomic deprivation and in patients who had a previous recording of asthma (Table 2). The most important modifiable risk factor was smoking. Compared with never-smokers, the risk of COPD was substantially higher in female ever-smokers (95% CI, 8.92–10.43) and 6.72 times higher in male ever-smokers (95% CI, 6.19–7.30).

The accuracy of the risk prediction model in discriminating between patients who did and those who did not develop COPD during the 10-year follow-up was ROC_AUC = 0.845 (95% CI, 0.840–0.850) for females and ROC_AUC = 0.832 (95% CI, 0.827–0.837) for males (Table 3; the ROC curves are shown in Supplementary Figure 1). The accuracy of the model was higher than that of the deconstructed models that included only age, only smoking, or only age and smoking (Table 3). Sensitivity and specificity values for the various cutoffs on the model’s PI are presented in Supplementary Appendix II.

Figure 1 shows the calibration plots for the full risk prediction model including all risk factors in the validation data set, separately for men and women. The model was well calibrated except for the highest risk category, in which the incidence of COPD was overestimated by the model.

### Table 2. Multiple Cox regression models for the association between risk factors and COPD in the derivation cohort, separately for females (N = 245,351) and males (N = 235,552)

<table>
<thead>
<tr>
<th>Years of age</th>
<th>HR*, females</th>
<th>95% CI around HR</th>
<th>HR*, males</th>
<th>95% CI around HR</th>
</tr>
</thead>
<tbody>
<tr>
<td>35–39 (reference)</td>
<td>1</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>40–44</td>
<td>2.05</td>
<td>1.75</td>
<td>2.41</td>
<td>2.06</td>
</tr>
<tr>
<td>45–49</td>
<td>3.71</td>
<td>3.19</td>
<td>4.31</td>
<td>3.87</td>
</tr>
<tr>
<td>50–54</td>
<td>5.49</td>
<td>4.75</td>
<td>6.34</td>
<td>6.02</td>
</tr>
<tr>
<td>55–59</td>
<td>8.15</td>
<td>7.07</td>
<td>9.40</td>
<td>9.66</td>
</tr>
<tr>
<td>60–64</td>
<td>10.52</td>
<td>9.13</td>
<td>12.11</td>
<td>14.01</td>
</tr>
<tr>
<td>65+</td>
<td>25.75</td>
<td>22.46</td>
<td>29.52</td>
<td>31.89</td>
</tr>
<tr>
<td>Ever-smoker (reference = never-smoker)</td>
<td>9.61</td>
<td>8.92</td>
<td>10.34</td>
<td>6.72</td>
</tr>
</tbody>
</table>

### Table 3. Accuracy of the full prediction model and deconstructed models in predicting the 10-year incidence of COPD in the validation cohort for females (N = 126,449) and males (N = 121,306)

<table>
<thead>
<tr>
<th>Model</th>
<th>ROC_AUC, females</th>
<th>95% CI around ROC_AUC</th>
<th>ROC_AUC, males</th>
<th>95% CI around ROC_AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model including only risk factor smoking</td>
<td>0.709</td>
<td>0.703</td>
<td>0.715</td>
<td>0.671</td>
</tr>
<tr>
<td>Model including only risk factor age</td>
<td>0.728</td>
<td>0.721</td>
<td>0.734</td>
<td>0.763</td>
</tr>
<tr>
<td>Model including risk factors age and smoking</td>
<td>0.829</td>
<td>0.824</td>
<td>0.834</td>
<td>0.817</td>
</tr>
<tr>
<td>Full model including all risk factors: age, smoking, level of deprivation and prior asthma</td>
<td>0.845</td>
<td>0.840</td>
<td>0.850</td>
<td>0.832</td>
</tr>
</tbody>
</table>

Abbreviations: CI, confidence interval; COPD, chronic obstructive pulmonary disease; HR, hazard ratio.

© 2014 Primary Care Respiratory Society UK/Macmillan Publishers Limited npj Primary Care Respiratory Medicine (2014) 14011
DISCUSSION

Main findings

We have developed and validated the first model for risk prediction of incident cases of COPD using routinely collected data from a very large national general practice database. In the derivation cohort, the COPD risk was 9.6 times higher in female ever-smokers compared with never-smokers and 6.7 times higher in male ever-smokers compared with never-smokers. The risk of COPD increased for both sexes with increasing level of deprivation and in patients with a previous recording of asthma. In the validation cohort, the model discriminated well between patients who did and those who did not develop COPD.

Strengths and limitations of this study

An advantage of the approach followed in this study is the inclusion of less well-established risk factors. All time variant risk factors in our model were recorded by the GPs before the patients’ entry date into the cohort. We used a longitudinal design in which we followed up patients for a median duration of 8 years. Furthermore, we were able to assess the interaction between sex and the most important risk factor, smoking, which has not been possible before. However, only risk factors that are sufficiently assessed and/or recorded in general practice could be considered. Although we included the most important risk factors known from the literature, other factors may have been overlooked, such as physical inactivity as well as occupational exposure to dust, chemical agents and fumes. Inclusion of such risk factors, in the presence of sufficiently recorded data, may have increased the accuracy of the model. Also, the use of routine data did not allow us to calculate pack-years of smoking history, which would be the desirable indicator of risk from tobacco exposure than our rather crude categorization of ‘ever-smokers’ versus ‘never-smokers’.

We used a GP-recorded diagnosis of COPD to define our primary outcome, but the diagnosis could not be formally verified through linkage to individual lung function measures. As a consequence, misclassification of COPD (over- or under-diagnosis) may have occurred in some cases. It would have been very useful to validate the GPs’ diagnosis of COPD with individual patient spirometric data, but, unfortunately, these data were not available in our database. Generally speaking, however, the validity of the GP-recorded diagnosis of COPD in Scottish primary care can be considered accurate. Since the publication of the National Institute for Health and Clinical Excellence guideline for the management of COPD in 2004 and the introduction of the Quality Outcomes Framework, which provides GP practices additional payment for high levels of clinical care, the recording of spirometry values in COPD patients has markedly increased. A recent study undertaken in Scottish GP practices showed that 88% of COPD patients had a recording of forced expiratory volume in 1s in the previous 15 months. Results from a Dutch study indicate that the validity and quality of spirometry performed in general practice is satisfactory compared with spirometry performed in a pulmonary function laboratory. Furthermore, a Canadian study showed that individuals with COPD can be accurately identified in health administration data. Nevertheless, it should be noted that COPD is a complex disease that consists of several clinical phenotypes, but these were not distinguished by our model.

Interpretation of findings in relation to previously published work

Our work is novel in its use of a longitudinal primary care database to predict future COPD in a population of individuals registered who have no previous recording of the disease. The incidence rate of 5.53 per 1,000 patient-years was higher than that found in our earlier analysis of English QRResearch practices (2.0 per 1,000 patient-years); however, data from QRResearch included the entire population and the estimates are therefore likely to substantially underestimate incidence in the age groups most at risk for COPD. Our incidence rate was also higher than that found in a Dutch study using a GP database (2.9 per 1,000 patient-years) but comparable to that found in a Canadian study using population-based health administrative data (5.9 per 1,000 person-years). Differences between studies in reported rates may be explained by differences in source population, definition of outcome and duration of follow-up.

As expected, smoking was found to be the most important modifiable risk factor for COPD. The risk was substantially higher in female smokers than in male smokers. This important finding indicates that smoking is likely to lead to higher rates of newly diagnosed COPD among women than in men. There is some evidence from previous research that females who smoke are more susceptible to developing COPD than men who smoke. Explanations for this phenomenon include gender differences in which cigarette smoke is inhaled and metabolised, genetic predisposition for smoking-related lung damage and differences in airway anatomy. It may also be possible, however, that this gender difference is a result of bias. If, for example, men are more likely to underreport their smoking behaviour or are less likely to be recorded as a smoker, there would be a higher misclassification rate among men, causing bias towards a smaller HR. Furthermore, if some important but unmeasured risk factor would exist primarily in men (e.g., occupational exposure to dust, chemicals or fumes) the relative risk of another risk factor (smoking) would be lower in men than in women.

We also found that increasing socioeconomic deprivation was a risk factor for COPD diagnosis, independent of smoking status. The association between COPD and socioeconomic status has been found previously and is likely due to a number of exposures, including environmental or occupational exposure to smoke or to other pollutants.

Similar to our cohort, previous studies reported that patients with physician-diagnosed asthma were at increased risk of developing COPD. It has been hypothesised that asthma and COPD share a common background and that airway inflammation in those with increased airway hyperresponsiveness may lead to lung remodelling with resulting airflow obstruction. The model discriminated well between patients who did and those who did not develop COPD during the 10 years of follow-up, indicated by ROC_{AUCS} of 0.85 for females and 0.83 for males with very small CIs. These figures were higher than for the deconstructed model including only age and smoking (ROC_{AUCS} of 0.83 for females and 0.82 for males). Thus, inclusion of the risk factors, level of deprivation and previous recording of asthma, increased the accuracy of predicting future COPD over and above the most important and well-known risk factors smoking and age. The model’s calibration was also good, except for the highest risk category in which the incidence of COPD was slightly overestimated by the model. This may have been the result of the very high HRs in the oldest-age category (HR = 25.75 for females and HR = 31.89 for males aged 65+ years relative to the 35–39 age category).

Implications for future research, policy and practice

Our risk prediction model has the potential to be used in routine clinical practice to identify those at highest risk and thereby offers the opportunity for better and more efficient targeting of interventions aiming to reduce the risk of developing COPD, in particular smoking cessation interventions. For this use, we have developed a simple ‘COPD risk calculator’ (see Supplementary Appendix III). Our model is therefore complementary to the various existing risk models concerned with the early detection of patients with existing, but still undiagnosed, COPD (see e.g., refs 11, 12).
Conclusions
In summary, we have developed and validated the first risk prediction model for the development of COPD, which has the major advantage of being populated entirely by routinely collected data held in electronic health records.

ACKNOWLEDGEMENTS
The authors thank staff at The Primary Care Clinical Informatics Unit and the general practitioners that contributed data to the study.

CONTRIBUTIONS
I (DK) can certify that the manuscript represents valid work and that neither this manuscript nor one with substantially similar content under their authorship has been published or is being considered for publication elsewhere. All co-authors have contributed to the conception of the design, drafting of the article and approval of its final version.

COMPETING INTERESTS
OCP received grant money from Pfizer and Boehringer Ingelheim. AS received grant money from NAPP. DK received an unrestricted research grant from Pfizer for a smoking cessation trial. OCP is an Assistant editor of, and AS is Joint Editor-in-Chief of, the PCRJ. Neither were involved in the editorial review of, nor the decision to publish, this article. CRS and WV have nothing to declare.

FUNDING
DK was supported by a short-term research fellowship from the Dutch Asthma Foundation. CRS was supported by a national post-doctoral fellowship from the Commonwealth Fund, its directors, or staff. These organisations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript. DK had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

REFERENCES
20 de Wet C, McKay J, Bowie P. Combining QOF data with the care bundle approach may provide a more meaningful measure of quality in general practice. BMC Health Serv Res 2012; 12: 351.