npj | Primary Care Respiratory Medicine

www.nature.com/npjpcrm

ARTICLE OPEN

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

Daniel Kotz[1,2], Colin R Simpson[2], Wolfgang Viechtbauer[3], Onno CP van Schayck[1,2] and Aziz Sheikh[1,2,4]

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.[6] Still, even among primary care patients with known risk factors of the disease, the rate of undiagnosed COPD is high.[7] 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.[16] 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

[1]Department of Family Medicine, CAPHRI School for Public Health and Primary Care, Maastricht University Medical Centre, Maastricht, The Netherlands; [2]Allergy & Respiratory Research Group, Centre for Population Health Sciences, The University of Edinburgh, Edinburgh, UK; [3]Department of Psychiatry and Psychology, MHeNS School for Mental Health and Neuroscience, Maastricht University, Maastricht, The Netherlands and [4]Division 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

SPSS Version 17.0 (SPSS, Chicago, IL, USA), 66% of practices ($N = 159$) were allocated to a derivation cohort and 34% ($N = 80$) to a validation cohort.

We defined 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.

Coding of primary outcome and risk factors

The primary outcome was the first recorded diagnosis of COPD during the period between a patient's entry date and exit date into the cohort. The definition of COPD was based on codes from the Read Clinical Classification System, which was produced for clinicians in primary care and is used by the majority of primary care electronic patient record systems (read codes H3, H31 and below (excluding H3101, H31y0, H3122), H32 and below, and H36 to H3z). For a complete list of Read codes used to define outcome and risk factors see Supplementary Appendix I.

A range of potential risk factors for COPD, which have been described in the literature and which were sufficiently recorded in the database, were assessed. Age was categorised into 35–39, 40–44, 45–49, 50–54, 55–59, 60–64 and 65+ years. Smoking status was categorised into 'ever-smoker' or 'never-smoker'. We defined 'ever-smokers' as patients recorded as 'smoker' or 'ex-smoker' at any time, and 'never-smokers' as patients recorded as 'non-smoker' at any time and no codings as 'smoker' or 'ex-smoker' at any other time. Asthma diagnosis was based on Read codes (see Supplementary Appendix I) and regarded as a risk factor if it had been recorded prior to the patient's entry date into the cohort, with no recording as a reference. Sex and socioeconomic status were regarded as time invariant potential risk factors, with the latter being measured using the Carstairs Index of Deprivation (coded 1 = least deprived to 5 = most deprived).[17]

There were too few data entries at baseline for the potential risk factors acute respiratory infections, respiratory symptoms, asthma, physical inactivity, ethnicity, occupational exposure to risk factors, family history of respiratory disease, pulmonary tuberculosis, and prescription of adrenoreceptor agonists, bronchodilators, theophylline and inhaled corticosteroids, and hence these were discarded from the prediction analyses (all variables were present in fewer than 3% of patients).

Statistical analyses

We performed an *a priori* test to determine whether an association between COPD and the most important risk factor, smoking status, was modified by sex by comparing a logistic regression model including 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 $PI_{der} = \Sigma \beta_i X_i$, where β_i is the regression coefficient of the risk factor X_i from the Cox model (this method was adapted from ref. 18). PI_{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 PI_{der} (the values for calculating PI_{der} are presented in Supplementary Appendix II). 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. PI_{val} for patients was then calculated from the validation cohort using the regression coefficients from the derivation cohort. PI_{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 (ROC_{AUC}) for all values of PI_{val}. Finally, the prediction model was calibrated by plotting the predicted 10-year incidence of COPD against the observed incidence for each interval on PI_{val} in patients from the validation cohort.

In ancillary analyses, we deconstructed the prediction model and calculated the ROC_{AUCs} 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 1. Baseline characteristics of patients in the derivation and validation cohorts

	Derivation cohort, N = *480,903*	*Validation cohort,* N = *247,755*	*Total,* N = *728,658*
Patient-years	3,229,285	1,669,471	4,898,756
Years follow-up, median (IQR)	7.92 (3.76–10.00)	7.88 (3.77–10.00)	7.92 (3.76–10.00)
COPD	3.81 (18,342)	3.53 (8,746)	3.7 (27,088)
Age, mean (s.d.)	50.5 (11.3)	50.7 (11.3)	50.6 (11.3)
Male sex	49.9 (235,552)	49.0 (121,306)	49.0 (356,858)
Ever smokers	55.6 (267,533)	54.5 (134,986)	55.2 (402,519)
Carstairs level of deprivation			
1st Quintile (least deprived)	19.3 (92,946)	17.1 (42,330)	18.6 (135,276)
2nd Quintile	18.3 (87,957)	23.0 (57,035)	19.9 (144,992)
3rd Quintile	24.1 (115,976)	24.6 (61,014)	24.3 (176,990)
4th Quintile	23.0 (110,712)	20.0 (49,549)	22.0 (160,261)
5th Quintile (most deprived)	15.2 (73,312)	15.3 (37,827)	15.3 (111,139)
Prior asthma	3.2 (15,584)	3.4 (8,535)	3.3 (24,119)

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 smokers when compared with male smokers: 9.61 times 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 ($N = 235,552$)

	HR[a], females	*95% CI around HR*		*HR[a], males*	*95% CI around HR*	
		Lower	*Upper*		*Lower*	*Upper*
Years of age						
35–39 (reference)	—	—	—	—	—	—
40–44	2.05	1.75	2.41	2.06	1.72	2.46
45–49	3.71	3.19	4.31	3.87	3.29	4.56
50–54	5.49	4.75	6.34	6.02	5.14	7.04
55–59	8.15	7.07	9.40	9.66	8.28	11.27
60–64	10.52	9.13	12.11	14.01	12.04	16.32
65+	25.75	22.46	29.52	31.89	27.48	37.01
Ever-smoker (reference = never-smoker)	9.61	8.92	10.34	6.72	6.19	7.30
Carstairs level of deprivation						
1st Quintile (least deprived: reference)	—	—	—	—	—	—
2nd Quintile	1.25	1.14	1.37	1.36	1.25	1.48
3rd Quintile	1.65	1.52	1.78	1.60	1.48	1.73
4th Quintile	1.95	1.80	2.11	1.91	1.77	2.06
5th Quintile (most deprived)	2.58	2.39	2.79	2.52	2.34	2.73
Prior asthma (reference = no recording)	2.79	2.62	2.97	3.37	3.15	3.60

Abbreviations: CI, confidence interval; COPD, chronic obstructive pulmonary disease; HR, hazard ratio.
[a]HR of COPD diagnosis for a risk factor (or risk factor category) compared with its reference, adjusted for all other risk factors.

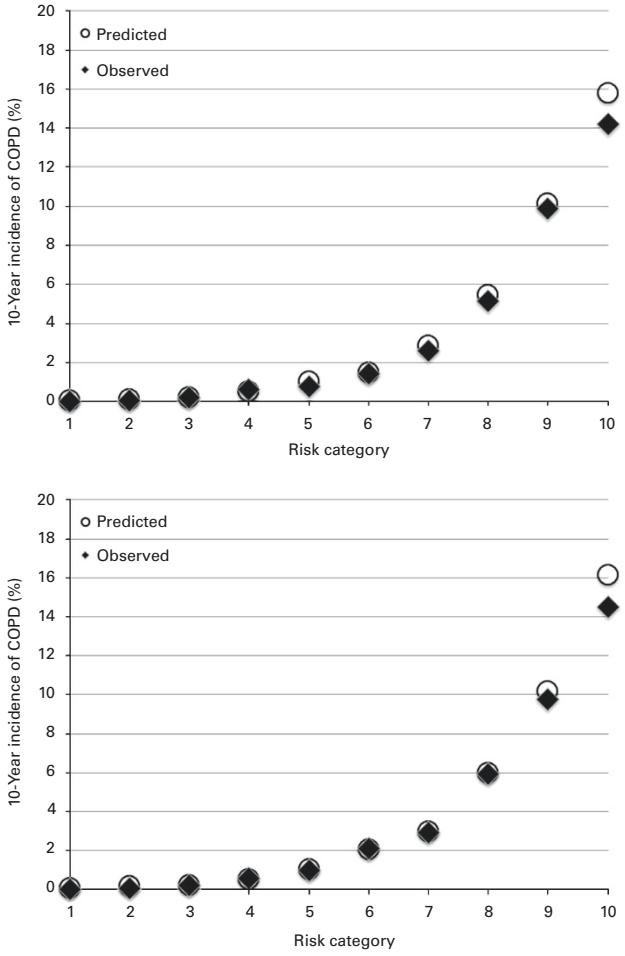

Figure 1. Calibration plot of the full risk prediction model including all risk factors showing the predicted and observed 10-year incidence of chronic obstructive pulmonary disease (COPD) per risk category in the validation cohort for females (upper) and males (lower).

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 ($N = 121,306$)

	ROC_{AUC}, females	*95% CI around ROC_{AUC}*		*ROC_{AUC}, males*	*95% CI around ROC_{AUC}*	
		Lower	*Upper*		*Lower*	*Upper*
Model including only risk factor smoking	0.709	0.703	0.715	0.671	0.665	0.678
Model including only risk factor age	0.728	0.721	0.734	0.763	0.757	0.769
Model including risk factors age and smoking	0.829	0.824	0.834	0.817	0.812	0.822
Full model including all risk factors: age, smoking, level of deprivation and prior asthma	0.845	0.840	0.850	0.832	0.827	0.837

Abbreviations: CI, confidence interval; COPD, chronic obstructive pulmonary disease; ROC_{AUC}, area under the receiver operating characteristics curve.

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.[19] 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.[20] 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.[21] Furthermore, a Canadian study showed that individuals with COPD can be accurately identified in health administration data.[22] Nevertheless, it should be noted that COPD is a complex disease that consists of several clinical phenotypes,[23] 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 QResearch practices (2.0 per 1,000 patient-years); however, data from QResearch included the entire population and the estimates are therefore likely to substantially underestimate incidence in the age groups most at risk for COPD.[24] Our incidence rate was also higher than that found in a Dutch study[25] using a GP database (2.9 per 1,000 patient-years) but comparable to that found in a Canadian study[26] 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.[27] 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.[28,29] 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[26,30] 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.[25,31] It has been hypothesised that asthma and COPD share a common background[32] 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 practices 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

OCPS 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. OCPS 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 Chief Scientist's Office of the Scottish Government. AS is supported by The Commonwealth Fund, a private independent foundation based in New York City. The views presented here are those of the author and not necessarily those of The Commonwealth Fund, its directors, officers, 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

1 Global Initiative for Chronic Obstructive Lung Disease. Global strategy for the diagnosis, management, and prevention of chronic pulmonary disease. Revised 2011, http://www.goldcopd.com (accessed 24 April 2012). NHLBI/WHO.
2 World Health Organisation (WHO). Global burden of disease http://www.who.int/mediacentre/factsheets/fs310/en/index.html (accessed 24 April 2012).
3 Menezes AMB, Perez-Padilla R, Jardim JB, Muiño A, Lopez MV, Valdivia G *et al.* Chronic obstructive pulmonary disease in five Latin American cities (the PLATINO study): a prevalence study. *Lancet* 2005; **366**: 1875–1881.
4 Buist AS, McBurnie MA, Vollmer WM, Gillespie S, Burney P, Mannino DM *et al.* International variation in the prevalence of COPD (the BOLD Study): a population-based prevalence study. *Lancet* 2007; **370**: 741–750.
5 Gershon AS, Wang C, Wilton AS, Raut R, To T. Trends in chronic obstructive pulmonary disease prevalence, incidence, and mortality in Ontario, Canada, 1996 to 2007: a Population-Based Study. *Arch Intern Med* 2010; **170**: 560–565.
6 Bischoff EWMA Schermer TRJ, Bor H, Brown P, van Weel C, van den Bosch WJ. Trends in COPD prevalence and exacerbation rates in Dutch primary care. *Br J Gen Pract* 2009; **59**: 927–933.
7 Hill K, Goldstein RS, Guyatt GH, Blouin M, Tan WC, Davis LL *et al.* Prevalence and underdiagnosis of chronic obstructive pulmonary disease among patients at risk in primary care. *CMAJ* 2010; **182**: 673–678.
8 Soriano JB, Zielinski J, Price D. Screening for and early detection of chronic obstructive pulmonary disease. *Lancet* 2009; **374**: 721–732.
9 Sin DD, McAlister FA, Man SF, Anthonisen NR. Contemporary management of chronic obstructive pulmonary disease: scientific review. *JAMA* 2003; **290**: 2301–2312.
10 Mannino DM, Buist AS. Global burden of COPD: risk factors, prevalence, and future trends. *Lancet* 2007; **370**: 765–773.
11 Price DB, Tinkelman DG, Halbert RJ, Nordyke RJ, Isonaka S, Nonikov D *et al.* Symptom-based questionnaire for identifying chronic obstructive pulmonary disease in smokers. *Respiration* 2006; **73**: 285–295.
12 Martinez FJ, Raczek AE, Seifer FD, Conoscenti CS, Curtice TG, D'Eletto T *et al.* Development and initial validation of a self-scored COPD Population Screener Questionnaire (COPD-PS). *COPD* 2008; **5**: 85–95.
13 Taylor DR. Risk assessment in asthma and COPD: a potential role for biomarkers? *Thorax* 2009; **64**: 261–264.
14 The Primary Care Clinical Informatics Unit—Research. http://www.abdn.ac.uk/pcciu/index.htm (accessed 6 July 2012).
15 Anandan C, Simpson CR, Fischbacher C, Sheikh A. Exploiting the potential of routine data to better understand the disease burden posed by allergic disorders. *Clin Exp Allergy* 2006; **36**: 866–871.
16 Helms PJ, Ekins Daukes S, Taylor MW, Simpson CR, McLay JS. Utility of routinely acquired primary care data for paediatric disease epidemiology and pharmacoepidemiology. *Br J Clin Pharmacol* 2005; **59**: 684–690.
17 Morris R, Carstairs V. *Deprivation and health in Scotland*. Aberdeen University Press: Abderdeen, 1991.
18 Terwee CB, Nieveen Van Dijkum EJ, Gouma DJ, Bakkevold KE, Klinkenbijl JH, Wade TP *et al.* Pooling of prognostic studies in cancer of the pancreatic head and periampullary region: the Triple-P study. Triple-P study group. *Eur J Surg* 2000; **166**: 706–712.
19 Smith CJP, Gribbin J, Challen KB, Hubbard RB. The impact of the 2004 NICE guideline and 2003 General Medical Services contract on COPD in primary care in the UK. *QJM* 2008; **101**: 145–153.
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.
21 Schermer TR, Jacobs JE, Chavannes NH, Hartman J, Folgering HT, Bottema BJ *et al.* Validity of spirometric testing in a general practice population of patients with chronic obstructive pulmonary disease (COPD). *Thorax* 2003; **58**: 861–866.
22 Gershon AS, Wang C, Guan J, Vasilevska-Ristovska J, Cicutto L, To T *et al.* Identifying individuals with physician diagnosed COPD in health administrative databases. *COPD* 2009; **6**: 388–394.
23 Rabe KF, Wedzicha JA. Controversies in treatment of chronic obstructive pulmonary disease. *Lancet* 2011; **378**: 1038–1047.
24 Simpson CR, Hippisley-Cox J, Sheikh A. Trends in the epidemiology of chronic obstructive pulmonary disease in England: a national study of 51 804 patients. *Br J Gen Pract* 2010; **60**: e277–e284.
25 Afonso ASM, Verhamme KMC, MCJM Sturkenboom, Brusselle GG. COPD in the general population: prevalence, incidence and survival. *Respir Med* 2011; **105**: 1872–1884.
26 Gershon AS, Warner L, Cascagnette P, Victor JC, To T. Lifetime risk of developing chronic obstructive pulmonary disease: a longitudinal population study. *Lancet* 2011; **378**: 991–996.
27 Sorheim I-C, Johannessen A, Gulsvik A, Bakke PS, Silverman EK, DeMeo DL. Gender differences in COPD: are women more susceptible to smoking effects than men? *Thorax* 2010; **65**: 480–485.
28 Ben-Zaken Cohen S, Paré PD, Man SFP, Sin DD. The Growing burden of chronic obstructive pulmonary disease and lung cancer in women. *Am J Respir Crit Care Med* 2007; **176**: 113–120.
29 Han MK, Postma D, Mannino DM, Giardino ND, Buist S, Curtis JL *et al.* Gender and chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2007; **176**: 1179–1184.
30 Nacul LC, Soljak M, Meade T. Model for estimating the population prevalence of chronic obstructive pulmonary disease: cross sectional data from the Health Survey for England. *Popul Health Metr* 2007; **5**: 8.
31 Silva GE, Sherrill DL, Guerra S, Barbee RA. Asthma as a risk factor for COPD in a longitudinal study. *Chest* 2004; **126**: 59–65.
32 Orie NGM, Sluiter HJ, De Vries K, Tammeling GJ, Witkop JI. The host factor in bronchitis. In: Orie NGM, Sluiter HJ (eds). *Bronchitis Assen: Royal Van Gorcum* 1961, 43–59.

Supplementary Information accompanies this paper on the *npj Primary Care Respiratory Medicine* website (http://www.nature.com/npjpcrm)