Predicting risk of undiagnosed COPD: development and validation of the TargetCOPD score

Shamil Haroon[1], Peymane Adab[1], Richard D. Riley[2], David Fitzmaurice[3] and Rachel E. Jordan[1]

Affiliations: [1]Institute of Applied Health Research, University of Birmingham, Birmingham, UK. [2]Research Institute for Primary Care and Health Sciences, Primary Care Sciences, Keele University, Keele, UK. [3]Warwick Medical School, Coventry, UK.

Correspondence: Rachel E. Jordan, Institute of Applied Health Research, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK. E-mail: r.e.jordan@bham.ac.uk

@ERSpublications
Patients at high risk of undiagnosed symptomatic COPD can be identified using electronic primary care health records http://ow.ly/pAc4309MBl5

Cite this article as: Haroon S, Adab P, Riley RD, *et al.* Predicting risk of undiagnosed COPD: development and validation of the TargetCOPD score. *Eur Respir J* 2017; 49: 1602191 [https://doi.org/10.1183/13993003.02191-2016].

ABSTRACT Chronic obstructive pulmonary disease (COPD) is greatly underdiagnosed worldwide and more efficient methods of case-finding are required. We developed and externally validated a risk score to identify undiagnosed COPD using primary care records.

We conducted a retrospective cohort analysis of a pragmatic cluster randomised controlled case-finding trial in the West Midlands, UK. Participants aged 40–79 years with no prior diagnosis of COPD received a postal or opportunistic screening questionnaire. Those reporting chronic respiratory symptoms were assessed with spirometry. COPD was defined as presence of relevant symptoms with a post-bronchodilator forced expiratory volume in 1 s/forced vital capacity ratio below the lower limit of normal. A risk score was developed using logistic regression with variables available from electronic health records for 2398 participants who returned a postal questionnaire. This was externally validated among 1097 participants who returned an opportunistic questionnaire to derive the c-statistic, and the sensitivity and specificity of cut-points.

A risk score containing age, smoking status, dyspnoea, prescriptions of salbutamol and prescriptions of antibiotics discriminated between patients with and without undiagnosed COPD (c-statistic 0.74, 95% CI 0.68–0.80). A cut-point of ⩾7.5% predicted risk had a sensitivity of 68.8% (95% CI 57.3–78.9%) and a specificity of 68.8% (95% CI 65.8.1–71.6%).

A novel risk score using routine data from primary care electronic health records can identify patients at high risk for undiagnosed symptomatic COPD. This score could be integrated with clinical information systems to help primary care clinicians target patients for case-finding.

This article has supplementary material available from erj.ersjournals.com

Received: Nov 07 2016 | Accepted after revision: March 02 2017

Support statement: This paper presents independent research funded by the National Institute for Health Research (NIHR). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Dept of Health. S. Haroon was funded by a NIHR doctoral fellowship (DRF-2011-04-064). R.E. Jordan was funded by an NIHR post-doctoral fellowship (pdf/01/2008/023). P. Adab and R.E. Jordan are both principal investigators on an NIHR programme grant for investigating COPD in primary care (RP-PG-0109-10061). Funding information for this article has been deposited with the Crossref Funder Registry.

Conflict of interest: Disclosures can be found alongside this article at erj.ersjournals.com

Introduction

Chronic obstructive pulmonary disease (COPD) is the third leading cause of mortality worldwide [1], but 50–90% of the disease burden remains undiagnosed. Patients with undiagnosed COPD have been shown to have significant morbidity and burden to health services from exacerbations many years prior to their diagnosis, therefore contributing to a large drive worldwide to improve early diagnosis [2, 3]. While mass screening with spirometry among asymptomatic individuals is not recommended [4], earlier identification of patients with clinically significant but unreported symptoms (case-finding) could improve access to care and prevent disease progression [5].

A systematic approach to case-finding using an initial screening questionnaire mailed to ever-smokers (current and ex-smokers) followed by invitation to spirometry among those reporting relevant symptoms was recently evaluated in primary care [6, 7]. This proved to be twice as effective, and was more cost-effective than opportunistic case-finding and identified a substantial proportion of patients with potential to benefit from effective interventions. However, this method targeted a broad population (all ever-smokers aged 40–79 years) and was also reliant on patient response [8]. A more efficient approach is therefore needed.

A number of risk scores have been proposed, including one developed by our team, to help identify patients at high risk of undiagnosed COPD using routine clinical records [9–11]. However, their case definitions included patients with a new record of COPD diagnosed through usual care. Estimates in England suggest approximately two-thirds of COPD cases are undiagnosed [12–14]. Given this extent, the characteristics of patients diagnosed through routine clinical care may differ from those detected through active case-finding. Risk scores should therefore ideally be derived using case-found populations.

We report the development and validation of a new clinical score for identifying patients at high risk of undiagnosed COPD in primary care using data from TargetCOPD, a large cluster randomised controlled case-finding trial [6, 7].

Methods

This report has been written in accordance with the TRIPOD statement [15].

Study design

This is a retrospective cohort analysis of the intervention (case-finding) arm of the TargetCOPD cluster randomised controlled trial (RCT) [6], to develop and validate a risk score for identifying undiagnosed COPD. General practices in the TargetCOPD trial were randomised to either targeted case-finding or routine care. Eligible participants were recruited from August 2012 to June 2014. Those in practices that were allocated to the case-finding arm were individually randomised to either receive a screening questionnaire only when attending routine clinical appointments or to additionally receive a screening questionnaire by post. Participants reporting relevant respiratory symptoms (chronic cough or phlegm for ⩾3 months of the year for ⩾2 years, wheeze in the previous 12 months or dyspnoea of Medical Research Council grade ⩾2) were offered a diagnostic assessment with post-bronchodilator spirometry. We used data from their primary care electronic health records (EHRs) and spirometry assessment to develop and validate a risk score for undiagnosed COPD.

Population

Participants were aged 40–79 years with no prior diagnosis of COPD (supplementary table S1 provides clinical codes used for exclusion). Subjects were further excluded at the discretion of their general practitioner (*e.g.* terminal illness, recent bereavement, learning difficulties or pregnancy). This analysis was restricted to a subset of participants from 13 of the participating 27 practices allocated to the case-finding arm for whom data from their EHRs were available.

Setting

The TargetCOPD trial was based in primary care practices in the West Midlands, UK [6]. Participating practices broadly reflected the diversity of the population in terms of age, ethnicity, socioeconomic status (SES) and practice characteristics.

Outcome

COPD was defined as the presence of at least one chronic respiratory symptom (as described in the study design) together with airflow limitation measured by post-bronchodilator spirometry. Spirometry was performed to American Thoracic Society/European Respiratory Society standards [16] by trained research assistants using EasyOne spirometers (ndd Medical Technologies, Zurich, Switzerland) 20 min after the inhalation of 400 µg of salbutamol delivered through a metered dose inhaler and Volumatic spacer. Spirometers were calibrated on a daily basis and all research assistants underwent supervised training over a period of 3–6 months. All spirometry traces were reviewed by a lung function specialist. For this analysis airflow limitation was defined as a forced expiratory volume in 1 s (FEV_1)/forced vital capacity (FVC) ratio less than the lower limit of normal (less than the fifth percentile) adjusted for age, sex, height and ethnic group using the Global Lung Initiative 2012 equations, which provide the most recent and most representative global estimates [17]. This conservative definition of airflow limitation is less likely to overdiagnose COPD in older patients compared with using a fixed-ratio definition [18].

Data extraction

Data (clinical codes; see supplementary table S2) were extracted from EHRs based on predictors identified as potentially important in our previous analysis [10], including demographic characteristics, smoking status, respiratory symptoms, comorbidities, lower respiratory tract infections (LRTIs), respiratory medication prescriptions and selected antibiotic use indicated for the treatment of LRTIs. Data from residential postcodes were used to estimate SES using the Index of Multiple Deprivation (a measure of SES based on participants' residential postcodes; higher scores indicate higher levels of socioeconomic deprivation) [19]. All data were stored on an encrypted database.

Sample size

Subjects with missing outcome (COPD) status (predominantly those invited but who did not attend a spirometry assessment) were excluded from the analysis (n=755). Data from 2398 subjects who returned a postal questionnaire were used for model development (development sample) and from 1097 subjects from the same set of practices who returned an opportunistic questionnaire for external validation (external validation sample) (figure 1). This nonrandom splitting of the data ensured the developed risk score could be validated in new data from a different part of the intended population [20]. 7.9% of all subjects were newly diagnosed with COPD through the trial (198 in the development sample and 77 in the external validation sample). At least 10 outcome events are recommended per candidate predictor considered for inclusion in a logistic regression model [21]. There was therefore sufficient power to consider up to 19 candidate predictors in the developed model.

Model development

The model was developed using multivariable logistic regression considering the following candidate predictors for inclusion: age, sex, most recent smoking status, history of asthma and LRTIs, complaints of cough, dyspnoea, wheeze and sputum, and prescriptions of salbutamol, prednisolone and antibiotics, within the previous 3 years. Since there were very little (<1%) missing data for these candidate predictors, a complete-case analysis was performed [22]. We tested for interactions with a particular focus on age, sex and smoking status. The best-fitting terms for continuous variables were determined using fractional polynomial regression [23]. Predictors not statistically significant at the $p<0.05$ level were removed from the model (although age and smoking status were forced in because of their known clinical importance). The fit of the reduced model was then compared with the full model using a likelihood ratio test.

To improve the calibration of the model predictions and adjust for overfitting, the model's calibration slope coefficient was estimated in 1000 bootstrap samples to determine the shrinkage factor (the average calibration slope). This was multiplied against predictor coefficients in the developed model to produce the final model equation [24].

Internal validation performance

The sensitivity and specificity of the predicted probabilities from the final risk score were plotted on a receiver operator characteristic (ROC) curve to examine the discrimination performance. The risk score was internally validated using bootstrap resampling (with 1000 replications) to estimate the c-statistic (area under the ROC curve) corrected for overfitting [25]. Calibration was assessed by grouping subjects into deciles of predicted risk and comparing the observed with the expected number diagnosed with COPD.

External validation performance

The c-statistic and calibration of the final risk score were then assessed in the external validation sample. As a comparator, we also assessed the discrimination performance of our previously developed clinical score [10] in the external validation sample. This model included smoking status, history of asthma, LRTIs and prescriptions of salbutamol as predictors of undiagnosed COPD.

Sensitivity analysis

The final risk score was additionally validated in the external validation sample using a case definition that also included the presence of at least one chronic respiratory symptom, but required an alternative definition of airflow obstruction commonly used in clinical practice (FEV_1/FVC <0.7) [26].

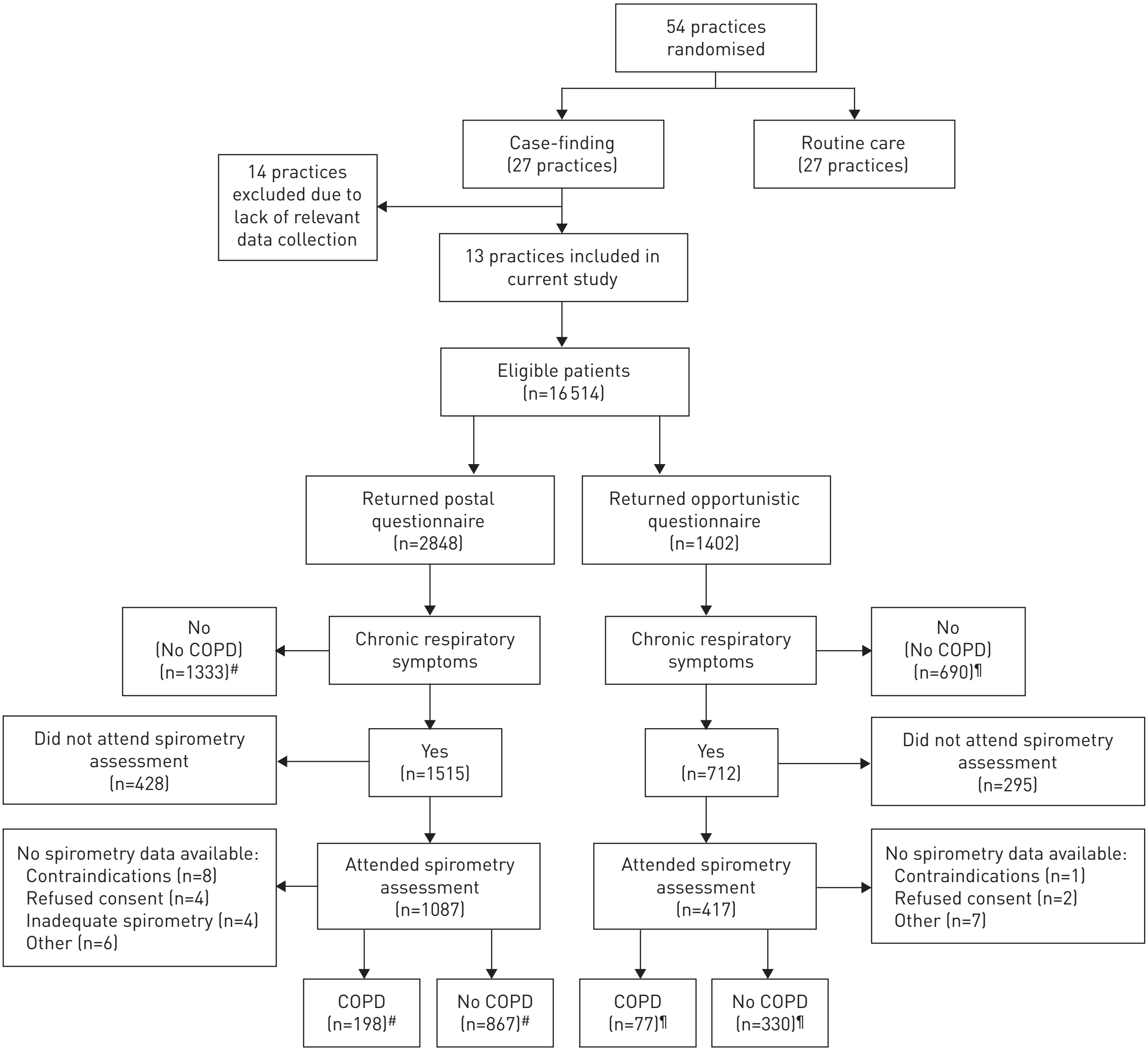
54 practices randomised
Case-finding (27 practices)
Routine care (27 practices)
14 practices excluded due to lack of relevant data collection
13 practices included in current study
Eligible patients (n=16 514)
Returned postal questionnaire (n=2848)
Returned opportunistic questionnaire (n=1402)
No (No COPD) (n=1333)#
Chronic respiratory symptoms
Chronic respiratory symptoms
No (No COPD) (n=690)¶
Did not attend spirometry assessment (n=428)
Yes (n=1515)
Yes (n=712)
Did not attend spirometry assessment (n=295)
No spirometry data available: Contraindications (n=8) Refused consent (n=4) Inadequate spirometry (n=4) Other (n=6)
Attended spirometry assessment (n=1087)
Attended spirometry assessment (n=417)
No spirometry data available: Contraindications (n=1) Refused consent (n=2) Other (n=7)
COPD (n=198)#
No COPD (n=867)#
COPD (n=77)¶
No COPD (n=330)¶

FIGURE 1 Participant selection. COPD: chronic obstructive pulmonary disease. #: development sample; ¶: external validation sample.

Preparing the risk score for clinical practice

To prepare the risk score for use as a screening tool, we evaluated cut-points for dichotomising the predicted probabilities into low and high risk. The sensitivity and specificity were calculated in the external validation sample across a range of cut-points, alongside the positive and negative predictive values, likelihood ratios, and number of diagnostic assessments needed to identify one individual with undiagnosed COPD. All analyses were performed using Stata version 13.1 (StataCorp, College Station, TX, USA).

Ethical approval

Ethical approval for the TargetCOPD trial was received from the Solihull Ethics Committee (Integrated Research Application System reference 11/WM/0403).

Results

Practice characteristics

Practice size varied, with the majority having a list size below 10 000 (supplementary table S3). Most practices served populations in socioeconomically deprived areas with a diverse range of ethnicities. The mean (range) prevalence of diagnosed COPD prior to the trial was 1.3% (0.8–2.9%).

Development sample: population characteristics

The development sample included 2398 individuals, of whom 198 (8.3%) were diagnosed with COPD during the study (figure 1). The mean age was 59.6 years, 51.6% were male and the majority (85.0%) were of white ethnicity. The majority (77.7%) of newly diagnosed COPD was mild (FEV1 % pred ⩾80%), with 21.1% moderate (FEV1 % pred 50–79%), 1.0% severe (FEV1 % pred 30–49%) and 0.2% very severe (FEV1 % pred <30%).

Based on data extracted from EHRs (table 1), current smoking was significantly more common among participants with COPD than those without (32.8% *versus* 14.1%). There was also a higher prevalence of asthma, and a slightly higher prevalence of anxiety and depression, among those with COPD. However, the prevalence of other chronic conditions was similar in both groups. Documented cough, dyspnoea, sputum production, LRTIs and respiratory prescriptions were all also more common among individuals with COPD.

Individuals with unknown COPD status (predominantly those who did not attend an assessment) differed from those in the development sample across a number of demographic characteristics (supplementary table S4): they were generally younger (mean age 55.8 *versus* 59.6 years), and higher proportions were female (52.5% *versus* 48.4%) and current smokers (33.5% *versus* 15.8%).

Model results

Complete data for candidate predictors were available for 2380 patients (99.2%) in the development sample (table 2). The final model of EHR-recorded factors included smoking status, age, dyspnoea, prescriptions of

TABLE 1 Characteristics extracted from electronic health records (EHRs): development sample

	COPD	Non-COPD	Missing data
Subjects	198 (8.3)	2200 (91.7)	
Age years	60.8±9.6	59.5±10.7	7 (0.3)
40–49	30 (15.2)	528 (24.0)	
50–59	54 (27.3)	621 (28.2)	
60–69	74 (37.4)	595 (27.0)	
70–79	40 (20.2)	456 (20.7)	
Male	107 (54.0)	1128 (51.3)	4 (0.2)
Smoking status			
Never-smoker	37 (18.7)	744 (33.8)	11 (0.5)
Ex-smoker	95 (48.0)	1135 (51.6)	
Current smoker	65 (32.8)	311 (14.1)	
IMD score	37.4 (19.8–41.3)	23.3 (19.8–41.3)	0 (0.0)
Comorbidities			
Asthma	10 (5.1)	29 (1.3)	Unknown[¶]
Ischaemic heart disease	11 (5.6)	146 (6.6)	
Heart failure	3 (1.5)	20 (0.9)	
Diabetes	17 (8.6)	192 (8.7)	
Stroke	2 (1.0)	18 (0.8)	
Tuberculosis	3 (1.5)	11 (0.5)	
Osteoporosis	4 (2.0)	37 (1.7)	
Depression/anxiety[#]	36 (18.2)	335 (15.2)	
LRTIs[#]	41 (20.7)	233 (10.6)	
Symptoms[#]			
Cough	61 (30.8)	385 (17.5)	Unknown[¶]
Dyspnoea	23 (11.6)	78 (3.5)	
Wheeze	30 (15.2)	362 (16.5)	
Sputum	11 (5.6)	43 (2.0)	
Unintended weight loss	2 (1.0)	9 (0.4)	
Prescriptions[#]			
Salbutamol	74 (37.4)	251 (11.4)	Unknown[¶]
Prednisolone	40 (20.2)	138 (6.3)	
Antibiotics[+]	116 (58.6)	783 (35.6)	

Data are presented as n (%), mean±SD or median (interquartile range). COPD: chronic obstructive pulmonary disease; IMD: Index of Multiple Deprivation (higher scores indicate higher levels of socioeconomic deprivation); LRTI: lower respiratory tract infection. [#]: recorded within previous 3 years of commencing case-finding at the registered practice; [¶]: it was unknown whether absence of a record of comorbidities, symptoms and prescriptions in EHRs was due to true absence of those factors or due to underrecording; [+]: amoxicillin, clarithromycin, co-amoxiclav, erythromycin, doxycycline and cephalexin.

TABLE 2 Candidate predictors evaluated in the multivariable logistic regression model

	Unadjusted		Adjusted	
	OR (95% CI)	**p-value**	**OR (95% CI)**	**p-value**
Age years				
40–49	Reference category		Reference category	
50–59	1.53 (0.97–2.43)	0.070	1.66 (1.02–2.70)	0.043*
60–69	2.19 (1.41–3.40)	<0.001*	2.63 (1.64–4.23)	<0.001*
70–79	1.54 (0.95–2.52)	0.082	1.72 (1.01–2.93)	0.044*
Male	1.11 (0.83–1.49)	0.471	1.04 (0.76–1.43)	0.800
Smoking status				
Never-smoker	Reference category		Reference category	
Ex-smoker	1.68 (1.14–2.49)	0.009*	1.71 (1.13–2.59)	0.012*
Current smoker	4.20 (2.75–6.43)	<0.001*	5.58 (3.50–8.89)	<0.001*
Asthma ever	3.98 (1.91–8.30)	<0.001*	1.44 (0.62–3.37)	0.400
LRTIs#	2.20 (1.52–3.19)	<0.001*	1.00 (0.84–1.19)	0.977
Symptoms#				
Cough	2.10 (1.52–2.89)	<0.001*	1.00 (0.68–1.47)	0.986
Dyspnoea	3.58 (2.19–5.84)	<0.001*	2.19 (1.27–3.76)	0.005*
Wheeze	0.91 (0.61–1.36)	0.635	1.14 (0.73–1.76)	0.564
Sputum	2.95 (1.50–5.82)	0.002*	1.55 (0.71–3.37)	0.270
Prescriptions#				
Salbutamol	4.63 (3.38–6.36)	<0.001*	3.05 (2.01–4.62)	<0.001*
Prednisolone	3.78 (2.57–5.57)	<0.001*	1.76 (1.09–2.84)	0.020*
Antibiotics¶	2.56 (1.90–3.44)	<0.001*	1.52 (1.06–2.18)	0.023*

Candidate predictors are presented as binary variables unless specified otherwise. Based on data extracted from electronic health records for 2380 subjects in the development sample. OR: odds ratio; LRTI: lower respiratory tract infection. #: recorded within previous 3 years; ¶: amoxicillin, clarithromycin, co-amoxiclav, erythromycin, doxycycline and cephalexin. *: p<0.05, statistically significant.

salbutamol and prescriptions of antibiotics (table 3). Age was included as two fractional polynomial terms since it was not linear in the logit scale. The final model fitted as well as the full model (likelihood ratio test p=0.185) and no significant interactions were found. The final model equation was: predicted probability of undiagnosed COPD=$e^x/(1+e^x)$, where $x=(1.43\times10^{-4}\times\text{age}^3)-(3.18\times10^{-5}\times\text{age}^3\times\ln[\text{age}])+(0.51\times\text{ex-smoker [Y/N]})+(1.60\times\text{current smoker [Y/N]})+(0.72\times\text{dyspnoea [Y/N]})+(0.045\times\text{number of salbutamol prescriptions})+(0.99\times\text{salbutamol prescriptions [Y/N]})+(0.47\times\text{antibiotic prescriptions [Y/N]})-6.16$, with Y=yes (value=1) or N=no (value=0).

Internal validation

When applied to the development sample the apparent c-statistic was 0.76 (95% CI 0.73–0.80) and after correcting for overfitting using bootstrapping was 0.76 (95% 0.72–0.79). Although smoking status and age were the most important predictors in the risk score, restricting it to just these variables reduced the c-statistic to 0.65 (95% CI 0.60–0.69).

TABLE 3 Final risk score

Predictor	**β# (95% CI)**	**p-value**
Age^3	1.43×10^{-4} (6.11×10^{-5}–2.26×10^{-4})	0.001
Age^3×ln[age]	-3.18×10^{-5} (-5.02×10^{-5}– -1.34×10^{-5})	0.001
Ex-smoker	0.51 (0.10–0.91)	0.015
Current smoker	1.60 (1.14–2.05)	<0.001
Dyspnoea¶	0.72 (0.18–1.26)	0.010
Number of salbutamol prescriptions¶	0.045 (0.015–0.075)	0.003
One or more salbutamol prescription¶	0.99 (0.56–1.42)	<0.001
One or more antibiotic prescription¶	0.47 (0.13–0.80)	0.007
Constant	−6.16 (−7.63– −4.70)	<0.001

#: regression coefficient; ¶: recorded within the previous 3 years. Predicted probability of undiagnosed chronic obstructive pulmonary disease=$e^x/(1+e^x)$ (see main text for details). Note that the shrinkage factor was 1, which indicates that there was no evidence of overfitting in the final model.

External validation sample: population characteristics

Among 1097 subjects in the external validation population, 77 (7.0%) were newly diagnosed with COPD (supplementary table S5). The mean age was 60.1 years and 51.6% were male, similar to the development sample. Again, a significantly greater proportion of subjects with COPD were current smokers (31.2% *versus* 17.1%). However, participants in the external validation sample had a slightly higher SES. 1083 subjects (98.7%) had complete data on all candidate predictors and were included in the external validation.

External validation: risk score performance

The developed risk score demonstrated similar discrimination characteristics when applied to the external validation sample (c-statistic 0.74, 95% CI 0.68–0.80) and performed better than our previously developed clinical score (c-statistic 0.70, 95% CI 0.64–0.76) in the external validation sample (figure 2) [10]. The final risk score showed excellent calibration of observed to predicted COPD risk up to 10%, but slightly overestimated the predicted risk from 10% to 30%, beyond which comparisons were unreliable due to small sample sizes (table 4). When using the fixed-ratio definition of airflow limitation ($FEV_1/FVC < 0.7$), the c-statistic for the final risk score remained at 0.74 (95% CI 0.70–0.78).

Implementation in clinical practice

Increasing the cut-point to define high risk reduces the number of assessments needed for each new diagnosis of COPD, although accompanied by a reduction in sensitivity (table 5). The optimum cut-point should balance both sensitivity and specificity, taking into consideration costs and resource availability. At a cut-point of 7.5% (*i.e.* classing subjects with a predicted risk ⩾7.5% as high risk), which would represent 33.9% of the target population, the risk score is estimated in the external validation sample to have a sensitivity of 68.8% (95% CI 57.3–78.9%) and a specificity of 68.8% (95% CI 65.8–71.6%), and would require 7 (95% CI 6–10) patients to undergo a diagnostic assessment to identify one with COPD.

Discussion

Principal findings

We have developed and externally validated the TargetCOPD score from a large case-finding trial in primary care [6, 7] to predict the risk of undiagnosed COPD using routine data from EHRs. The risk score incorporates five factors commonly recorded in health records: age, smoking status, presence of dyspnoea, prescriptions of salbutamol and prescriptions of antibiotics commonly prescribed for LRTIs. When externally validated, the risk score discriminated between patients with and without COPD, and performed better than our previously developed score [10], which relied on incident COPD from routine records rather than actively case-found patients. The risk score also performed similarly when using the fixed-ratio definition of airflow limitation. In our newly developed risk score, a cut-point of ⩾7.5% would expect to identify >70% of patients with undiagnosed COPD, needing seven diagnostic assessments for each new diagnosis. Use of higher cut-points could reduce this number at the expense of reducing sensitivity.

Comparison with existing literature

Several other risk scores have previously been developed for undiagnosed COPD, although the TargetCOPD score is the only one to use case-found COPD patients (table 6). As with other scores, our own previous risk score used newly diagnosed COPD patients, identified through routine care [10]. Its final predictors differed from the TargetCOPD score, including LRTIs and history of asthma but not

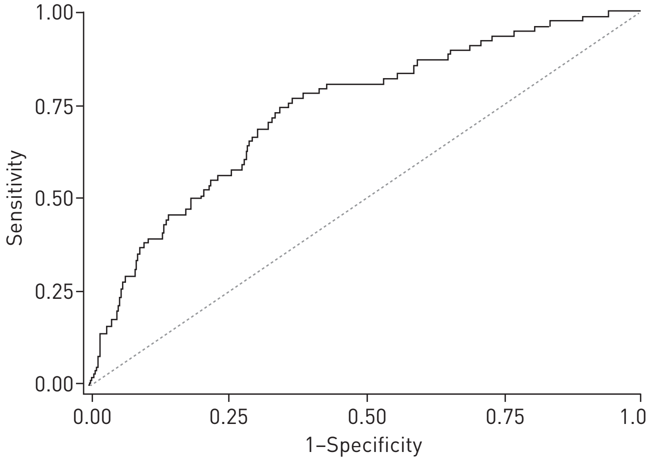

FIGURE 2 Receiver operator characteristic curve for the TargetCOPD score in the external validation sample (c-statistic 0.74, 95% CI 0.68–0.80).

TABLE 4 Model calibration

Predicted risk %	Development sample#			External validation sample¶		
	COPD n	Non-COPD n	Observed risk % (95% CI)	COPD n	Non-COPD n	Observed risk % (95% CI)
0–9	84	1781	4.5 (3.6–5.5)	35	780	4.3 (3.0–5.9)
10–19	53	271	16.4 (12.5–20.8)	20	158	11.2 (7.0–16.8)
20–29	33	89	27.0 (19.4–35.8)	11	51	17.7 (9.2–29.5)
30–39	11	17	39.3 (21.5–59.4)	6	9	40.0 (16.3–67.7)
40–49	8	15	34.8 (34.9–90.1)	3	4	42.9 (9.9–81.6)
50–59	4	6	40.0 (12.2–73.8)	0	2	0 (0–84.2)
60–69	1	3	25.0 (0.6–80.6)	0	0	0
70–79	2	0	100 (15.8–100)	1	2	33.3 (0.8–90.6)
80–89	1	0	100 (2.5–100)	0	0	0
90–100	0	1	0.0 (0–97.5)	1	0	100.0 (2.5–100)
Total	197	2183	8.3 (7.2–9.5)	77	1006	7.1 (5.7–8.8)

COPD: chronic obstructive pulmonary disease. #: n=2380; ¶: n=1083.

history of dyspnoea or prescriptions of antibiotics as predictors. Furthermore, the TargetCOPD score overcomes an important limitation of our previous risk score, where we could not include the effect of age as it was a matching factor (although it is well established that risk of COPD increases with age) [27]. A history of asthma and LRTIs did not remain statistically significant in the full multivariable model in the current analysis. However, prescriptions of salbutamol and antibiotics are closely associated with asthma and LRTIs, respectively, and are possibly better documented in EHRs; therefore, they may be reflecting similar clinical features.

Kotz *et al.* [9] also recently developed and internally validated a prediction model for COPD using routine longitudinal data from general practices in Scotland. Their model included age, smoking status, history of asthma and also socioeconomic deprivation, but only considered a limited range of risk factors and was not externally validated. Their model, like our previous clinical score, was developed on incident cases of COPD diagnosed through routine care, the disease status of which may have been misclassified because of underdiagnosis [2] and misdiagnosis [28]. Other risk scores have also been developed for COPD using routine primary and secondary healthcare data [29–31], but are unlikely to be applicable in primary care due to the predictors included, many of which are not routinely recorded solely in primary care records (table 6).

A number of other case-finding tools have also been developed and evaluated including screening questionnaires and handheld flowmeters [32–34]. However, these require additional resources and patient interactions, and are likely to be less efficient than the use of automated risk prediction scores.

Strengths

We investigated a range of risk factors and developed and validated our risk score on a population with no prior diagnosis of COPD that were actively case-found in a wide range of general practices. We employed a robust case definition which is likely to be representative of clinically significant, undiagnosed COPD and confirmed with quality-assured spirometry. The developed risk score was externally validated, increasing the likelihood of its validity in other primary care populations, although further external validation is needed on populations from a different location. The final risk score incorporates a small number of commonly recorded factors from EHRs that should ensure its applicability in routine primary care in the UK and similar health systems. However, it would be more challenging to implement in health systems that use paper-based health records or where electronic records are less detailed.

Limitations

We used a smaller sample size than several other studies reporting the development of COPD risk scores from routine healthcare data [9, 10, 29]. Although the study was adequately powered for the number of risk factors considered for the model selection, a larger sample size would have enabled estimation of the parameters with greater precision. Ideally a larger sample size would have been used for external validation (simulation-based estimates suggest at least 100 outcome events are required [35]) and would have improved our ability to evaluate the score calibration.

The case definition of COPD used in this study was the presence of relevant self-reported symptoms in addition to airflow limitation and patients who did not report symptoms were not assessed with spirometry.

TABLE 5 Diagnostic accuracy of the model in the external validation sample[#]

Cut-point	Patients at/above cut-point[¶]	Sensitivity	Specificity	Correctly classified	Positive likelihood ratio	Negative likelihood ratio	Positive predictive value	Negative predictive value	NND
⩾2.5	88.2	97.4 (90.9–99.7)	12.5 (10.5–14.7)	18.6	1.11 (1.07–1.16)	0.21 (0.05–0.82)	7.9 (6.2–9.8)	98.4 (94.5–99.8)	13 (11–17)
⩾5.0	53.7	80.5 (69.9–88.7)	48.4 (45.3–51.5))	50.7	1.56 (1.38–1.77)	0.40 (0.25–0.64)	10.7 (8.3–13.5)	97.0 (95.1–98.3)	10 (8–13)
⩾7.5	33.9	68.8 (57.3–78.9)	68.8 (65.8–71.6)	68.8	2.21 (1.85–2.63)	0.45 (0.32–0.63)	14.4 (11.0–18.5)	96.6 (95.1–97.8)	7 (6–10)
⩾10.0	24.8	54.5 (42.8–65.9)	77.5 (74.8–80.1)	75.9	2.43 (1.92–3.07)	0.59 (0.46–0.75)	15.7 (11.5–20.6)	95.7 (94.1–97.0)	7 (5–9)
⩾12.5	19.8	46.8 (35.3–58.5)	82.3 (79.8–84.6)	79.8	2.64 (2.01–3.47)	0.65 (0.52–0.80)	16.8 (12.1–22.5)	95.3 (93.7–96.6)	6 (5–9)
⩾15.0	15.0	41.6 (30.4–53.4)	87.0 (84.7–89.0)	83.8	3.19 (2.34–4.35)	0.67 (0.56–0.81)	19.6 (13.8–26.6)	95.1 (93.5–96.4)	6 (4–8)
⩾17.5	9.9	33.8 (23.4–45.4)	91.9 (90.0–93.5)	87.7	4.14 (2.85–6.03)	0.72 (0.61–0.85)	24.1 (16.4–33.3)	94.8 (93.2–96.1)	5 (3–7)
⩾20.0	8.3	28.6 (18.8–40.0)	93.2 (91.5–94.7)	88.6	4.23 (2.77–6.44)	0.77 (0.66–0.88)	24.4 (16.0–34.6)	94.5 (92.9–95.8)	5 (3–6)
⩾22.5	5.9	20.8 (12.4–31.5)	95.2 (93.7–96.5)	89.9	4.36 (2.60–7.30)	0.83 (0.74–0.93)	25.0 (15.0–37.4)	94.0 (92.4–95.4)	4 (3–7)
⩾25.0	3.7	14.3 (7.4–24.1)	97.1 (95.9–98.1)	91.2	4.96 (2.58–9.53)	0.88 (0.81–0.97)	27.5 (14.6–43.9)	93.7 (92.0–95.1)	4 (3–7)
⩾30.0	2.6	14.3 (7.4–24.1)	98.3 (97.3–99.0)	92.3	8.45 (4.11–17.4)	0.87 (0.80–0.96)	39.3 (21.5–59.4)	93.7 (92.1–95.1)	3 (2–5)
⩾35.0	1.9	10.4 (4.6–19.4)	98.8 (97.9–99.4)	92.5	8.71 (3.67–20.7)	0.91 (0.84–0.98)	40.0 (19.1–63.9)	93.5 (91.9–94.9)	3 (2–6)
⩾40.0	1.2	6.5 (2.1–14.5)	99.2 (98.4–99.7)	92.6	8.17 (2.74–24.4)	0.94 (0.89–1.00)	38.5 (13.9–68.4)	93.3 (91.6–94.7)	3 (2–8)
⩾45.0	0.8	2.6 (0.3–9.1)	99.3 (98.6–99.7)	92.4	3.73 (0.79–17.7)	0.98 (0.95–1.02)	22.2 (2.8–60.0)	93.0 (91.3–94.5)	5 (2–36)
⩾50.0	0.6	2.6 (0.3–9.1)	99.6 (99.0–99.9)	92.7	6.53 (1.22–35.1)	0.98 (0.94–1.01)	33.3 (4.3–77.7)	93.0 (91.3–94.5)	3 (2–23)

Data are presented as % or n (95% CI). NND: number of diagnostic assessments needed per case detected. [#]: n=1083; [¶]: percentage of subjects with a predicted risk score at or above the cut-point.

TABLE 6 Comparison of existing risk prediction models for chronic obstructive pulmonary disease (COPD)

Model/clinical score	Development	Validation	Predictors	c-statistic (95% CI)	Strengths	Limitations
TargetCOPD#	Retrospective cohort analysis of a case-finding cluster RCT and routine data from 13 general practices	Internal and external validation using data from subjects who completed a screening questionnaire and performed spirometry	Age; smoking; dyspnoea; salbutamol; antibiotics	External: 0.74 (0.68–0.80)	Developed and validated on subjects with previously undiagnosed COPD confirmed by quality-controlled spirometry; can be integrated with clinical information systems; good discrimination performance	Dependent on quality of clinical coding
Haroon *et al.* [10]#	Case–control study using routine data from 360 general practices	Internal and external validation using routine data	Smoking; salbutamol; asthma; LRTIs	External: 0.85 (0.83–0.86) in original study; 0.70 (0.64–0.76) in current study	Developed on large sample size; can be integrated with clinical information systems; high discrimination performance; considered wide range of risk factors	Predicts physician-diagnosed COPD¶; excluded age and sex as predictors; dependent on quality of clinical coding
Kotz *et al.* [9]#	Retrospective cohort study using routine data from 239 general practices	Internal validation using routine data	Age; smoking; SES; asthma	Internal: 0.85 (0.84–0.85) in males; 0.83 (0.83–0.84) in females	Developed on large sample size; can be integrated with clinical information systems; high discrimination performance; estimates 10-year risk of incident COPD	Predicts physician-diagnosed COPD¶; limited range of risk factors explored; includes a UK-specific index of socioeconomic deprivation (limiting applicability to other health systems); dependent on quality of clinical coding
Smidth *et al.* [31]	Cross-sectional analysis of routine data from seven general practices, secondary care registers and an RCT	Internal and external validation using routine data and data from an RCT	Chronic lung disease; respiratory medication; previous spirometry	Not reported	High positive predictive value	Predicts physician-diagnosed COPD¶; requires prior diagnosis of chronic lung disease; requires data linkage between primary and secondary care; difficult to administer
Mapel *et al.* [30]	Case–control study using routine data from four hospitals and 18 general practices	Internal and external validation using routine data	Antibiotics; respiratory and cardiovascular medications	Not reported	Only used data on medication prescriptions, which are likely to be well recorded; developed on large sample size	Predicts physician-diagnosed COPD¶;
Mapel *et al.* [29]	Case–control study using routine data from secondary care	Internal and external validation using routine data	19 healthcare utilisation characteristics including cor pulmonale and asthma	Not reported	Developed on large sample size; can be integrated with clinical information systems	Predicts physician-diagnosed COPD¶; model includes large number of predictors; includes predictors unlikely to be routinely recorded in primary care; excluded smoking status as a predictor

RCT: randomised controlled trial; LRTI: lower respiratory tract infection; SES: socioeconomic status. #: likely to be readily applicable in primary care; ¶: potential misclassification of disease (COPD) status during model development and validation.

However, patients may underreport symptoms and compensate for them by limiting their activities. This could have introduced misclassification bias. Furthermore, only 25.7% of all eligible patients responded to the screening questionnaire, which could have introduced response bias and may limit the generalisability of the score. However, this response rate is similar to the average response rate to questionnaires seen in other case-finding studies [33] and because of the pragmatic nature of the trial is likely to represent patients who might respond to screening invitations in real clinical practice.

Finally, the validity of our risk score among all potential subjects could not be determined because we were not able to include those with unknown COPD status and their characteristics differed from those included in our analysis across a number of demographic characteristics. However, our risk score is applicable to populations of individuals that are likely to respond to questionnaire surveys and are willing to attend subsequent clinical assessment.

Implications for clinicians, policy makers and research

The TargetCOPD score has been developed to help primary care services stratify patients according to their risk of undiagnosed COPD for targeted systematic case-finding (supplementary figure S1). The US Preventive Services Task Force recently recommended against screening for asymptomatic COPD on the basis that there was no evidence that it improves health-related quality of life, morbidity or mortality [4]. By contrast, the TargetCOPD score has been developed from patients with symptomatic and spirometry-verified disease who are more likely to benefit from treatment.

The score's ability to estimate the probability of undiagnosed COPD could be used to risk-stratify patients and could be used to help prioritise referral for diagnostic assessment, including spirometry, or for further screening (*e.g.* using handheld flowmeters). General practitioners could decide on a cut-point which reflects the resources available to them for conducting high-quality spirometry, balancing sensitivity and specificity. Since it relies entirely on routinely recorded data from EHRs, it could be integrated with clinical information systems by programming the model into these digital platforms. This would be applicable in countries with primary care clinical information systems similar to the UK, such as in a number of Western European countries, Israel, the USA, New Zealand, Australia and Canada [36, 37].

Finally, the TargetCOPD score should be externally validated in other primary care populations to better assess its generalisability, and its effectiveness in practice evaluated in RCTs, where the impact of using the risk score on patient outcomes can be evaluated as well as the associated costs [38]. This could include a cluster RCT comparing clinical outcomes (*e.g.* quality of life, hospitalisation and mortality) in practices that use the risk score to actively case find patients with undiagnosed COPD against practices that continue with alternative approaches to case-finding and usual care.

Conclusions

We have developed and externally validated the TargetCOPD score for assessing the risk of undiagnosed COPD among patients in primary care using routine data from EHRs. This is the first risk score for COPD that has been derived from patients identified through systematic case-finding and uses routine healthcare data readily available in many primary care settings. It could be used to help identify patients at high risk of COPD to provide appropriate clinical care, including earlier testing and treatment. The risk score should be externally validated in further populations, and its impact on clinical care and outcomes evaluated in RCTs.

Acknowledgements

We would like to thank Nicola Adderley (University of Birmingham, Birmingham, UK) for greatly assisting with data collection, Alexandra Enocson (University of Birmingham, Birmingham, UK), Andy Dickens, (University of Birmingham, Birmingham, UK), Joanne O'Beirne-Elliman (University of Warwick, Coventry, UK) and the Birmingham Lung Improvement Studies (BLISS) Team (University of Birmingham, Birmingham, UK) for the successful execution of the TargetCOPD trial, Martin Miller (University of Birmingham, Birmingham, UK) for reviewing all spirometry readings and providing expert guidance on lung function assessment, and Kym Snell (Keele University, Keele, UK) for advice on bootstrap validation. Most of all we would like to thank all the patients, including those in the BLISS patient advisory group, who participated in the TargetCOPD trial.

Author contributions: S. Haroon wrote the final protocol, collated the data, conducted the analysis and wrote the report. R.E. Jordan, P. Adab and R.D. Riley conceived the study, and wrote the initial protocol, advised on the conduct of the study, analysis and drafting of the manuscript. R.D. Riley provided specialist statistical support. R.E. Jordan, P. Adab and D. Fitzmaurice advised on the drafting of the manuscript. R.E. Jordan, P. Adab and D. Fitzmaurice are principal investigators on the TargetCOPD trial. All authors commented on and approved the final manuscript. P. Adab is the guarantor. P. Adab affirms that the manuscript is an honest, accurate and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

References

1 Lozano R, Naghavi M, Foreman K, *et al.* Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 2012; 380: 2095–2128.

2 Soriano J, Zielinski J, Price D. Screening for and early detection of chronic obstructive pulmonary disease. *Lancet* 2009; 374: 721–732.

3 Lamprecht B, Soriano JB, Studnicka M, *et al.* Determinants of underdiagnosis of COPD in national and international surveys. *Chest* 2015; 148: 971–985.

4 US Preventive Services Task Force. Screening for Chronic Obstructive Pulmonary Disease: US Preventive Services Task Force Recommendation Statement. *JAMA* 2016; 315: 1372–1377.

5 Bakke PS, Rönmark E, Eagan T, *et al.* Recommendations for epidemiological studies on COPD. *Eur Respir J* 2011; 38: 1261–1277.

6 Jordan R, Adab P, Jowett S, *et al.* TargetCOPD: a pragmatic randomised controlled trial of targeted case finding for COPD versus routine practice in primary care: protocol. *BMC Pulm Med* 2014; 14: 157.

7 Jordan RE, Adab P, Sitch A, *et al.* Targeted case finding for chronic obstructive pulmonary disease versus routine practice in primary care (TargetCOPD): a cluster-randomised controlled trial. *Lancet Respir Med* 2016; 4: 720–730.

8 Jordan RE, Lam KB, Cheng KK, *et al.* Case finding for chronic obstructive pulmonary disease: a model for optimising a targeted approach. *Thorax* 2010; 65: 492–498.

9 Kotz D, Simpson CR, Viechtbauer W, *et al.* Development and validation of a model to predict the 10-year risk of general practitioner-recorded COPD. *NPJ Prim Care Respir Med* 2014; 24: 14011.

10 Haroon S, Adab P, Riley RD, *et al.* Predicting risk of COPD in primary care: development and validation of a clinical risk score. *BMJ Open Respir Res* 2015; 2: e000060.

11 Mapel D, Frost F, Hurly J, *et al.* An algorithm for the identification of undiagnosed COPD cases using administrative claims data. *J Manag Care Pharm* 2006; 12: 458–465.

12 NHS Digital. Quality and Outcomes Framework – 2011–12, England level: clinical domain, chronic obstructive pulmonary disease data tables. 2013. http://content.digital.nhs.uk/qof Date last accessed: September 23, 2014.

13 Office for National Statistics. 2011 Census, population estimates by single year of age and sex for local authorities in the United Kingdom. 2013. www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/2011censuspopulationestimatesbysingleyearofageandsexforlocalauthoritiesintheunitedkingdom Date last accessed: September 23, 2014.

14 Jordan R, Lam K-B, Cheng K, *et al.* Case finding for chronic obstructive pulmonary disease: a model for optimizing a targeted approach. *Thorax* 2010; 65: 492–498.

15 Collins GS, Reitsma JB, Altman DG, *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med* 2015; 13: 1.

16 Miller MR, Hankinson J, Brusasco V, *et al.* Standardisation of spirometry. *Eur Respir J* 2005; 26: 319–338.

17 Quanjer PH, Stanojevic S, Cole TJ, *et al.* Multi-ethnic reference values for spirometry for the 3–95-yr age range: the global lung function 2012 equations. *Eur Respir J* 2012; 40: 1324–1343.

18 Miller MR, Levy ML. Chronic obstructive pulmonary disease: missed diagnosis versus misdiagnosis. *BMJ* 2015; 351: h3021.

19 McLennan D, Barnes H, Noble M, *et al.* The English Indices of Deprivation 2010. London, Crown, 2011.

20 Collins GS, de Groot JA, Dutton S, *et al.* External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 2014; 14: 40.

21 Peduzzi P, Concato J, Kemper E, *et al.* A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996; 49: 1373–1379.

22 White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 2011; 30: 377–399.

23 Royston P, Altman DG. Approximating statistical functions by using fractional polynomial regression. *Statistician* 1997; 46: 411–422.

24 Steyerberg EW, Eijkemans MJC, Habbema JDF. Application of shrinkage techniques in logistic regression analysis: a case study. *Stat Neerl* 2001; 55: 76–88.

25 Efron B, Tibshirani R. Improvements on cross-validation: the .632+ bootstrap method. *J Am Stat Assoc* 1997; 92: 548–560.

26 Global Initiative for Chronic Obstructive Lung Disease. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. 2013. www.goldcopd.org Date last accessed: September 23, 2014.

27 Fletcher C, Peto R. The natural history of chronic airflow obstruction. *Br Med J* 1977; 1: 1645–1648.

28 Walters JA, Walters EH, Nelson M, *et al.* Factors associated with misdiagnosis of COPD in primary care. *Prim Care Respir J* 2011; 20: 396–402.

29 Mapel DW, Frost J, Hurley JS, *et al.* An algorithm for the identification of undiagnosed COPD cases using administrative claims data. *J Manag Care Pharm* 2006; 12: 458–465.

30 Mapel DW, Petersen H, Roberts MH, *et al.* Can outpatient pharmacy data identify persons with undiagnosed COPD? *Am J Manag Care* 2010; 16: 505–512.

31 Smidth M, Sokolowski I, Kaersvang L, *et al.* Developing an algorithm to identify people with Chronic Obstructive Pulmonary Disease (COPD) using administrative data. *BMC Med Inform Decis* 2012; 12.

32 Haroon S, Jordan R, Takwoingi Y, *et al.* Diagnostic accuracy of screening tests for COPD: a systematic review and meta-analysis. *BMJ Open* 2015; 5: e008133.

33 Haroon SM, Jordan RE, O'Beirne-Elliman J, *et al.* Effectiveness of case finding strategies for COPD in primary care: a systematic review and meta-analysis. *NPJ Prim Care Respir Med* 2015; 25: 15056.

34 Tsukuya G, Samukawa T, Matsumoto K, *et al.* Comparison of the COPD Population Screener and International Primary Care Airway Group questionnaires in a general Japanese population: the Hisayama study. *Int J Chron Obstruct Pulmon Dis* 2016; 11: 1903–1909.

35 Vergouwe Y, Steyerberg EW, Eijkemans MJ, *et al.* Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol* 2005; 58: 475–483.

36 van Lieshout J, Goldfracht M, Campbell S, *et al.* Primary care characteristics and population-orientated health care across Europe: an observational study. *Br J Gen Pract* 2011; 61: e22–e30.

37 Ludwick DA, Doucette J. Adopting electronic medical records in primary care: lessons learned from health information systems implementation experience in seven countries. *Int J Med Inform* 2009; 78: 22–31.

38 Steyerberg EW, Moons KG, van der Windt DA, *et al.* Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013; 10: e1001381.