BMJ Open Respiratory Research

Ten-year prediction model for post-bronchodilator airflow obstruction and early detection of COPD: development and validation in two middle-aged population-based cohorts

Jennifer L Perret ,[1,2,3] Don Vicendese,[1,4] Koen Simons,[1] Debbie L Jarvis,[5] Adrian J Lowe,[1] Caroline J Lodge,[1] Dinh S Bui,[1] Daniel Tan,[1] John A Burgess,[1] Bircan Erbas,[6] Adrian Bickerstaffe,[1] Kerry Hancock,[7] Bruce R Thompson,[8] Garun S Hamilton,[9,10] Robert Adams,[11] Geza P Benke,[12] Paul S Thomas,[13] Peter Frith ,[14] Christine F McDonald,[2,3] Tony Blakely,[1] Michael J Abramson ,[12] E Haydn Walters,[1,15] Cosetta Minelli,[5] Shyamali C Dharmage,[1] on behalf of the TAHS and ECRHS Investigator Groups

To cite: Perret JL, Vicendese D, Simons K, *et al*. Ten-year prediction model for post-bronchodilator airflow obstruction and early detection of COPD: development and validation in two middle-aged population-based cohorts. *BMJ Open Resp Res* 2021;**8**:e001138. doi:10.1136/bmjresp-2021-001138

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjresp-2021-001138).

JLP, CM and SCD are joint senior authors.

Received 1 November 2021
Accepted 15 November 2021

For numbered affiliations see end of article.

Correspondence to
Dr Jennifer L Perret; jennifer.perret@unimelb.edu.au

ABSTRACT

Background Classifying individuals at high chronic obstructive pulmonary disease (COPD)-risk creates opportunities for early COPD detection and active intervention.

Objective To develop and validate a statistical model to predict 10-year probabilities of COPD defined by post-bronchodilator airflow obstruction (post-BD-AO; forced expiratory volume in 1 s/forced vital capacity<5th percentile).

Setting General Caucasian populations from Australia and Europe, 10 and 27 centres, respectively.

Participants For the development cohort, questionnaire data on respiratory symptoms, smoking, asthma, occupation and participant sex were from the Tasmanian Longitudinal Health Study (TAHS) participants at age 41–45 years (n=5729) who did not have self-reported COPD/emphysema at baseline but had post-BD spirometry and smoking status at age 51–55 years (n=2407). The validation cohort comprised participants from the European Community Respiratory Health Survey (ECRHS) II and III (n=5970), restricted to those of age 40–49 and 50–59 with complete questionnaire and spirometry/smoking data, respectively (n=1407).

Statistical method Risk-prediction models were developed using randomForest then externally validated.

Results Area under the receiver operating characteristic curve (AUC_{ROC}) of the final model was 80.8% (95% CI 80.0% to 81.6%), sensitivity 80.3% (77.7% to 82.9%), specificity 69.1% (68.7% to 69.5%), positive predictive value (PPV) 11.1% (10.3% to 11.9%) and negative predictive value (NPV) 98.7% (98.5% to 98.9%). The external validation was fair (AUC_{ROC} 75.6%), with the PPV increasing to 17.9% and NPV still 97.5% for adults aged 40–49 years with ≥1 respiratory symptom. To illustrate the model output using hypothetical case scenarios, a 43-year-old female unskilled worker who smoked 20 cigarettes/day for 30 years had a 27% predicted probability for post-BD-AO at age 53 if she continued to smoke. The predicted risk was 42% if she had coexistent active asthma, but only 4.5% if she had quit after age 43.

Conclusion This novel and validated risk-prediction model could identify adults aged in their 40s at high 10-year COPD-risk in the general population with potential to facilitate active monitoring/intervention in predicted 'COPD cases' at a much earlier age.

Key messages

- How can we classify individuals at high chronic obstructive pulmonary disease (COPD)-risk to create opportunities for early COPD detection before too much lung damage has occurred?
- Using information that is readily accessible from patients and a machine learning methodology, we have developed and validated a COPD risk-prediction model with good discriminatory ability from Australian and European general populations aged in their 40s to predict post-bronchodilator airflow obstruction approximately 10 years later.
- This approach can classify individuals when aged from their 40s but at high or very high COPD-risk who could benefit from serial spirometry; we strengthen the rationale for smoking cessation strategies in middle-age; and advance available precision medicine.

INTRODUCTION

Chronic obstructive pulmonary disease (COPD) ranks among the highest causes of potentially preventable hospitalisations,[1 2] yet there is a lack of action to generate

high-quality evidence to support the pre-emptive identification and/or management of individuals most at-risk. A risk-prediction approach like what is used to manage modifiable risk factors for cardiovascular disease and type II diabetes,[3 4] could also be useful for COPD which is multifactorial and typically features a gradual progression of airflow obstruction that can be established by middle-age. Evaluating COPD-risk for adults aged in their 40s represents an important time window, as selected screening of high-risk individuals using spirometry could confirm disease well before they usually seek medical attention.[5] Although only one study has studied the cost-effectiveness of actively finding COPD cases and found systematic case-finding could be useful if targeting older smokers,[6] theoretically, appropriate and early individualised interventions have potential to favourably influence poorer lung function trajectories,[7 8] and thereby slow or even prevent COPD onset. In the usual clinical scenario where healthcare professionals see patients prior to testing,[9] a risk-prediction model can have both diagnostic and 'prognostic' features as it would cover current and onward risks and assist in determining both the need for further tests and prognosis.

Previous attempts to develop COPD risk-prediction models have been limited and include: administrative databases, which had inaccurate smoking and COPD information; case–control designs, which are prone to selection bias; and/or stepwise regression statistical models, which are inclined to overfitting.[10 11] To date there has been only one externally validated risk-prediction tool that used longitudinal data but this was based on several clinical test results that would generally be unavailable to treating clinicians and their patients at the time of initial assessment.[12] Furthermore, no previous risk-prediction model has incorporated changes in smoking status prior to lung function measurement to contrast continuing smokers with quitters, which would indicate the potential prospective impact of subsequent smoking behaviour.

Using data from two of the largest respiratory cohorts worldwide, the Tasmanian Longitudinal Health Study (TAHS) and European Community Respiratory Health Survey (ECRHS), we aimed to develop and validate such a COPD risk-prediction model for middle-aged adults using a 'real world' scenario in a general population setting.

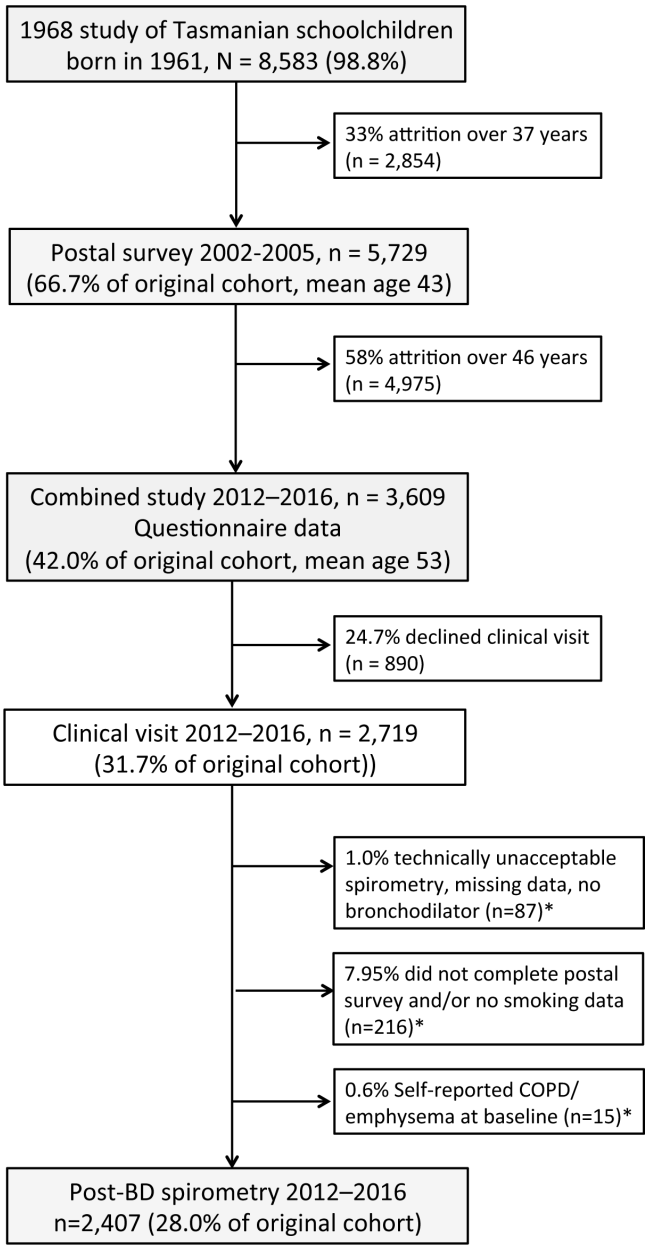

Figure 1 Study flow diagram of participation and non-participation in the development cohort, Tasmanian Longitudinal Health Study 1968–2016. Percentages for non-participation at subsequent follow-ups relate the proportion from the original 1968 survey. *Numbers may overlap. BD, bronchodilator; COPD, chronic obstructive lung disease.

METHODS

The Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis prediction model development and validation checklist,[13] and 2020 Editors' prediction framework on prediction modelling were followed.[11]

Study design: development cohort

Our sample included participants from the whole-of-population TAHS cohort, born in 1961, first studied in 1968 (n=8583) and followed into middle-age (figure 1).[14 15] At mean age 43 years, baseline questionnaire data from 5729 (67%) respondents were collected (online supplemental Methods E1). Approximately 10 years later, this original cohort was retraced and invited to participate in the 2012–2016 study (n=6128). Of 3609 respondents (58.9%), 2719 underwent pre-bronchodilator/post-bronchodilator (BD) spirometry (75.3%). Participants were those who had postal survey data *plus* 10-year smoking status/spirometry data (n=2407). Participants who reported doctor-diagnosed

COPD and/or emphysema at baseline were excluded (n=15).

Study design: validation cohort

ECRHS, a collaborative study of 29 centres within 14 mostly European countries, first recruited 17 250 20–44-year-old adults in the general community between 1992 and 1994 (ECRHS I),[16] details of which are available at https://wwwecrhsorg/. Participants of ECRHS II completed a detailed questionnaire, work history calendar (n=9645) and pre-BD spirometry (1998–2004, n=8033, age range 26–56). ECRHS III (2008–2012) was conducted in 27 centres in which participants underwent a detailed administered questionnaire and pre-BD/post-BD spirometry (n=5970, age range 38–67). The validation sample consisted of those persons aged in their 40s who participated in ECRHS II and subsequently underwent post-BD spirometry at ECRHS III in their 50s with complete predictor data (n=1407, online supplemental figure E1).

Outcome data collection and definition

Details on lung function data collection using international standards[17] and reference values[18] are outlined in online supplemental Methods E3. Post-bronchodilator airflow obstruction (post-BD-AO), referred to as spirometry consistent with COPD, was defined by forced expiratory volume in 1 s (FEV_1)/forced vital capacity (FVC)<5th percentile of normal predicted values following inhaled BD administered via spacer (ie, z-score<–1.645 SD).[18] Using this FEV_1/FVC criterion, mild-to-moderately severe post-BD-AO was defined by post-BD FEV_1 ≥50% predicted, and severe-to-very severe post-BD-AO by <50% predicted.[19]

Prediction model development and validation

Predictors

A pragmatic approach to selecting the predictor variables was adopted through using information which could be reasonably recalled in middle-age, practical to collect in primary care and feasibly harmonised with ECRHS data (online supplemental table E1,Method E2). The final input variables included: sex; current respiratory symptoms (wheezing, cough, sputum, breathlessness on exertion, chest tightness); smoking (current, duration, intensity, age-of-onset); asthma (asthma-ever, current adult asthma by age-of-onset) and socioeconomic status (occupational class, online supplemental Methods E1). Smoking at baseline and 10-year follow-up was expressed by a four-level variable: never-smoker; ex-smoker who quit before baseline; current smoker at baseline who quit before follow-up; or current smoker at follow-up. Baseline spirometry was not included as a predictor in the final model as post-BD spirometry was only collected for a subset of TAHS participants, enriched for asthma and symptoms (n=897).

Model development

Using R statistical software, we adopted randomForest,[20] a flexible, non-parametric and semi-automated machine learning method that considered all possible predictors and their interactions (online supplemental Methods E4a,table E2). The model was built on four randomly selected subsets of the data (80% of 2407 observations) and tested on a distinct fifth subset (20%, ie, remaining observations), optimally tuned and internally validated using a fivefold cross-validation scheme and this process was replicated 25 times. The final model was chosen based on the maximum area under the receiver operator characteristic curve (AUC_{ROC}, that is, its ability to discriminate between participants with and without post-BD-AO), followed by maximal sensitivity. Two thresholds were used to define a positive outcome:>50% probability of being a 'COPD case'; and the "optimal" threshold as defined by the Youden index.[21] Imputation of missing data was performed using a single imputation method integral to randomForest. More detailed statistical methods are reported in online supplemental Methods E4 (online supplemental sections E4a–g, Figures E2–4).

Hypothetical cases, individualised predictions and risk classification

Using the final model, personalised predictions were calculated from different case scenarios and recalibrated using the Platt scaling method.[22] Model calibration was assessed using the Hosmer-Lemeshow (HL) test that is, to assess the model's ability to match the predictions to the observed (or actual) COPD outcomes.

COPD-risk groups were defined based on the following approximations previously used in other clinical tools[3 4]: minimal risk if <1% predicted probability; low 1%–5% predicted probability; moderate 5%–10% predicted probability; high 10%–20% predicted probability or very high >20% predicted probability.

External calibration and discrimination

After model development, ECRHS data were used for external validation as two participant subsets: the main validation was derived from ECRHS participants with an extended age range of 40–49 years at baseline and 50–59 years when undergoing spirometry (n=1407) to broaden the model's transportability, and this was compared with ages similar to the development cohort that is, 40–44 years and 50–54 years, respectively (n=548). The final mean (SD) of model performance metrics was extracted from bootstrapped replications (n=50) and repeated 50 times to summarise uncertainty (online supplemental Methods E4h, table E3).[23]

Patient and public involvement

Patients, TAHS participants or the public were not involved in the design, or conduct, or reporting, or dissemination plans of our research.

RESULTS

TAHS and ECRHS participants

Descriptive results

Of the 2407 TAHS participants, 4.5% (n=108) fulfilled the lung function criterion for COPD at mean age 53 (table 1). Of these 108 participants, mild, moderate and moderately severe airflow obstruction was present for 106 (98%, n=91, n=11 and n=4, respectively). Post-BD-AO of any severity was present for 11.8% (n=62) of current smokers and 12.9% (n=50) of those who reported wheezing at age 43. A total of 187 (0.52%) clinical datapoints were missing in 3.8% (n=87) participants which included two cases with post-BD-AO (online supplemental table E4).

Among 1407 ECRHS participants who had complete data, post-BD-AO was present in 6.7% (n=95) and this included 10.1% (n=39) of all current smokers and 18.8% (n=47) of those who reported wheezing at baseline. Compared with TAHS, ECRHS participants were somewhat more likely to have exertional breathlessness, be current and heavier smokers, and not have current asthma (online supplemental table E4).

TAHS and ECRHS participants who had post-BD-AO in their 50s reported more current wheeze, chronic cough, sputum and chest tightness at baseline, that is, they were substantially more symptomatic than those without post-BD-AO (table 1). There were fewer current smokers in the group with complete compared with some missing data, but otherwise there were no appreciable differences in baseline characteristics (online supplemental table E5) or spirometry (online supplemental table E6).

Internal cross-validation of the final developed model

Discrimination between the risk-predictions and observed outcome was good, with an AUC_{ROC} of 80.8% (95% CI 80.0% to 81.5%) (table 2 and figure 2). Using the Youden index,[21] sensitivity was 80.3% (77.7 to 82.9) and specificity 69.1% (68.7 to 69.5). The NPV was ≥98.5% compared with a low PPV (11.1%), but this was 2.5-fold higher than the baseline prevalence of post-BD-AO (4.5%). The HL test provided reasonable evidence of calibration (p>0.13, table 2 and online supplemental figure E4). Imputing missing data did not appreciably improve the predictive model performance (AUC_{ROC} 81.1%).

External validation of the final developed model

Validation in the extended age group (ie, 1407 observations) performed similarly but with greater precision than that in the restricted age group (n=548 observations) and showed fairly good discriminatory ability, that is, AUC_{ROC} 75.6 and 74.6%, respectively (table 2 and figure 2). The PPV was not appreciably different when restricted to only current smokers aged 40–49 years but was slightly higher for adults with any current respiratory symptom/s (17.9% compared with 13.7%, table 2). This PPV was 2.7-fold higher than the baseline prevalence of post-BD airflow obstruction (6.7%).

Interactions between predictors

Of 210 potential interactions, the most frequent combination was between occupational class and smoking duration. For smoking beyond 25 years duration, the 10-year predicted probabilities for post-BD-AO were around 25% (figure 3, highlighted in yellow) which increased to around 40% for the occupational classes of labourers/cleaners, intermediate production/transport, house persons but not trade workers (highlighted in orange). The example of the single classification tree in online supplemental Methods E4c, figure E2, shows the 10-level occupational variable could be split multiple times within the same individual tree, with the averaging of predicted probabilities across thousands of classification trees plausibly explained the gradient (or blurring) of colours. The frequency of interactions is illustrated by online supplemental figure E5; 8 of the 10 most frequent interactions were between the smoking variables and occupation, 2 were between asthma and occupation, and none were between smoking and asthma. The 'multi-way importance plots' showed that occupational class, smoking duration and age-of-asthma and smoking onsets were more significant predictors in the TAHS dataset (online supplemental figures E6 and E7).

Individualised predicted probabilities and predicted occurrence

Due to the large number of potential combinations of predictors, it was not possible to present the full prediction model and predictions for all hypothetical scenarios. Selected examples of 43-year-old adults have been entered into the primary model (ie, complete cases and threshold >0.50) to predict probabilities of having the COPD outcome in their 50s. These scenarios included: an asymptomatic current smoker with varying smoking intensities/pack-years, then a current smoker with symptoms (online supplemental table E7); an ex-smoker with varying quit dates and respiratory symptoms (online supplemental table E8); a non-smoker with asthma (table 3); and comparisons between groups of quitters and continued smokers with or without active asthma at baseline (table 3).

Predictions for a current smoker

Predicted probabilities for post-BD-AO while aged 50s for a 43-year-old tradesman who currently smoked are presented in online supplemental table E7, while varying the daily cigarette intensity and age of smoking onset separately. Overall, the results suggest two smoking thresholds: (1) predicted risk-estimates that plateau beyond a smoking intensity of 20 cigarettes/day despite an increasing pack-year smoking history and (2) an acceleration of predicted risk-estimates beyond 20 years duration of smoking. Thus, the COPD-risk for a 43-year-old tradesman who smoked ≥20 cigarettes/day from age 18 was high (ie, predicted occurrence of one in every seven similar individuals), with and without respiratory symptoms typical of obstructive lung diseases. The predicted probability was very high if he started smoking from age

Table 1 Characteristics of participants with and without post-BD airflow obstruction in the development and validation samples

	Post-BD airflow obstruction aged 50s (n (%))†			
	Development cohort (TAHS, N=2407)‡		Validation cohort (ECRHS, N=1407)§	
Characteristics in middle-age*	**No (n=2299)**	**Yes (n=108)**	**No (n=1317)**	**Yes (n=95)**
Sex (% male)	1086 (49)	60 (55)	641 (49)	53 (56)
Age (mean years (SD))§				
Questionnaire	42.6 (0.5)	42.7 (0.6)	43.8 (2.5)	43.8 (2.6)
Post-BD spirometry	52.7 (0.8)	52.4 (0.7)	55.2 (2.5)	55.3 (2.5)
Post-BD spirometry at age 50s (mean (SD))				
Post-BD FEV_1 (L)	3.33 (0.7)	2.67 (0.7)	3.10 (0.7)	2.47 (0.7)
Post-BD FVC (L)	4.16 (0.9)	4.29 (1.0)	3.94 (0.9)	3.99 (1.0)
Post-BD FEV_1/FVC (ratio)	0.80 (0.05)	0.63 (0.07)	0.79 (0.05)	0.62 (0.05)
z-score (SD)	0.14 (0.7)	–2.30 (0.7)	–0.03 (0.8)	–2.33 (0.6)
Symptoms at age 40s (n (%))				
Current wheezing	327 (15)	52 (47)	203 (15)	47 (49)
Chronic cough	159 (7.1)	24 (22)	88 (7)	21 (22)
Chronic sputum	130 (5.8)	16 (15)	78 (6)	15 (16)
Breathlessness				
MRC-1 (none)	2026 (91)	78 (72)	1087 (82)	69 (73)
MRC-2	141 (6.3)	20 (18)	179 (14)	19 (20)
MRC-3/4	66 (3.0)	11 (10)	51 (4)	7 (7)
Chest tightness	343 (15)	37 (34)	197 (15)	32 (34)
Smoking (n (%); mean (SD); median (IQR); range†)				
Never smoker	1094 (49)	22 (20)	566 (43)	25 (26)
Past smoker	698 (31)	24 (22)	405 (31)	31 (33)
Pack-years	6.4 (1.7, 16)	2.0 (0.3, 17)	10.0 (4, 20)	10.0 (5, 20)
Current smoker	441 (20)	63 (58)	346 (26)	39 (41)
Cigs per day	14.0 (10)	19.8 (10)	14.1 (10)	19.7 (11)
Duration	26.0 (6)	27.4 (3)	26.1 (5)	27.1 (4)
Age of onset	16.3 (5)	15.6 (3)	17.5 (5)	16.6 (3)
Pack-years	17.4 (7, 28)	27.0 (18, 38)	21.9 (11, 30)	31.0 (22, 39)
Current at age 50s	290 (13)	53 (49)	234 (18)	36 (38)
Quit by age 50s	221 (10)	16 (15)	147 (11)	7 (7)
Asthma at age 40s (n (%))				
No asthma or wheezy breathing	1459 (65)	34 (31)	961 (73)	29 (31)
Wheezy breathing only	134 (6)	12 (11)	166 (13)	22 (23)
Self-reported asthma.				
Remitted	382 (17)	22 (20)	80 (6)	12 (13)
Active, early onset	88 (4)	15 (14)	33 (2.5)	12 (13)
Active, late onset	170 (8)	26 (24)	77 (6)	20 (21)
Employment at age 40s (n (%))				
Legislators, managers	257 (12)	12 (11)	121 (9)	10 (11)
Professionals	474 (21)	11 (10)	248 (19)	18 (19)
Technicians, associates	263 (12)	13 (12)	198 (15)	14 (15)
Trade workers	277 (12)	15 (14)	106 (8)	8 (8)
Clerks, services	534 (24)	24 (28)	240 (18)	19 (20)

Continued

Table 1 Continued

Characteristics in middle-age*	Post-BD airflow obstruction aged 50s (n (%))†			
	Development cohort (TAHS, N=2407)‡		Validation cohort (ECRHS, N=1407)§	
	No (n=2299)	Yes (n=108)	No (n=1317)	Yes (n=95)
Machine operators	130 (6)	9 (8)	46 (4)	4 (4)
Labourers, cleaners, other	147 (7)	15 (14)	62 (5)	4 (4)
House persons, other	139 (6)	10 (9)	74 (6)	5 (5)
Employed (unspecified)	6 (0.3)	0	208 (16)§	12 (13)
Non-work other	6 (0.3)	0	1 (0.1)	1 (1)

*Post-BD airflow obstruction defined by post-BD FEV1/FVC<5th percentile (z-score<–1.645).
†Summary data expressed by n (%) unless by mean (SD), for example, smoking intensity/duration/start age or median (IQR), for example, pack-years. Ranges for continuous predictors: smoking intensity 0–60; duration 0–37; age-of-onset 6–41; pack-years 0–108 (ever-smokers).
‡TAHS participant numbers refer to those aged in their 50s who underwent post-BD spirometry.
§ECRHS validation of participants aged 50 to up to 60 years (validation numbers for ages 50 up to 55 years not shown). Self-reported but unspecified employment was higher in ECRHS, as current job in the work history calendar was used with some missing (online supplemental Methods E1,E2).
BD, bronchodilator; ECRHS, European Community Respiratory Health Survey; FEV_1, forced expiratory volume in 1 s; FVC, forced vital capacity; LLN, lower limit of normal; MRC, Medical Research Council breathlessness scale; TAHS, Tasmanian Longitudinal Health Study.

Table 2 Performance metrics for the internal cross-validation and external validation of the COPD risk-prediction model, with and without imputation in the development TAHS dataset*

Model validation (n/N=COPD/total cases)	Diagnostic metrics (SE) †						HL 2 p value
	AUC_{ROC}	(Cut-off)‡	Sens	Spec	NPV	PPV	
Internal validation (TAHS)							
Complete case model (n/N=106/2320)	0.808 (0.004)	0.480	0.803 (0.013)	0.691 (0.002)	0.987 (0.001)	0.111 (0.004)	0.13
		0.50	0.779 (0.017)	0.713 (0.004)	0.985 (0.001)	0.115 (0.002)	
Imputed data model (n/N=108/2407)	0.811 (0.004)	0.450	0.816 (0.013)	0.671 (0.003)	0.987 (0.001)	0.105 (0.002)	0.30
		0.50	0.764 (0.012)	0.724 (0.003)	0.985 (0.001)	0.115 (0.002)	
External validations (ECRHS) using complete case model§							
Equivalent age group (n/N=39/548)§	0.746 (0.006)	0.483	0.745 (0.010)	0.668 (0.003)	0.972 (0.001)	0.148 (0.005)	0.95
		0.50	0.666 (0.011)	0.686 (0.003)	0.964 (0.001)	0.141 (0.005)	
Extended age group (n/N=95/1407)¶	0.756 (0.001)	0.483	0.769 (0.003)	0.659 (0.001)	0.975 (0.001)	0.140 (0.001)	0.69
		0.50	0.737 (0.003)	0.677 (0.001)	0.975 (0.001)	0.142 (0.001)	
Current smokers only (n/N=36/268)	0.639** (0.010)	0.50	0.835 (0.011)	0.173 (0.003)	0.870 (0.008)	0.137 (0.003)	
Current asthma only (n/N=32/142)	0.458** (0.006)	0.50	0.969 (0.005)	0.018 (0.002)	0.662 (0.055)	0.223 (0.005)	
Any current respiratory symptom (n/N=72/631)	0.719** (0.004)	0.50	0.905 (0.005)	0.469 (0.003)	0.975 (0.001)	0.179 (0.004)	

*In TAHS, complete case numbers (n/N=106/2320) and imputed data (n/N=108/2407 participants).
†SE=SD deviations from the mean (equivalent to SE).
‡Based on >50% predicted probably for a positive case or optimised cut-off as per the Youden index.
§Data from ECRHS II (age 40–44) and ECRHS III (age 50–55) (n/n=39/548 participants).
¶Data from ECRHS II (age 40–49) and ECRHS III (age 50–59) (n/n=95/1407 participants).
**AUC_{ROC} values based only on a subset of data are poor indicators of model performance (as not based on the entire dataset).
AUC_{ROC}, area under the receiver operator characteristic curve; COPD, chronic obstructive lung disease; ECRHS, European Community Respiratory Health Survey; HL, Hosmer-Lemeshow; n, number of COPD cases; N, total number; NPV, negative predictive values; PPV, positive predictive value; sens, sensitivity; spec, specificity; TAHS, Tasmanian Longitudinal Health Study.

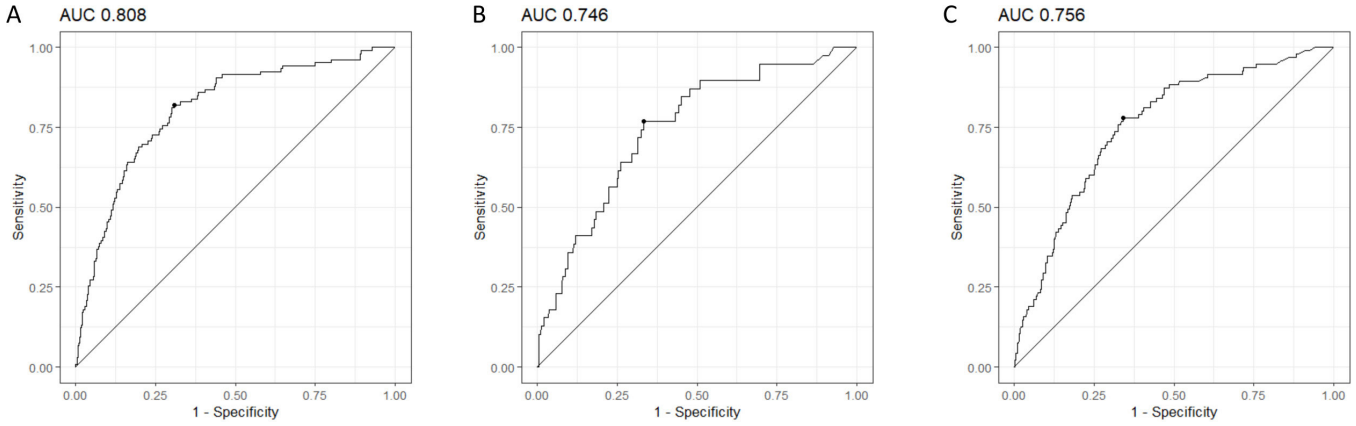

Figure 2 (A–C) Area under the receiver operator characteristic curve (AUC_{ROC}). Internal validation of the main chronic obstructive lung disease risk-prediction model using complete cases in Tasmanian Longitudinal Health Study (A). External validation using the corresponding (40–44 and 50–54 years) and extended age groups (40–49 and 50–59 years) in European Community Respiratory Health Survey (B and C, respectively). The Youden index that defines the optimal cut-off as specified in table 2 is indicated by the small black dot on the corresponding curves.

13 (1 in 3.7 persons). A twofold variation in the predicted probabilities for post-BD-AO when aged 50s was observed across the spectrum of occupations (online supplemental table E9).

Predictions for an ex-smoker

Predicted probabilities for post-BD-AO while aged 50s for a 43-year-old tradesman who had quit smoking are presented in online supplemental table E8, with varying years since quit dates (and therefore varied quit age and pack-years). These risk-estimates showed that the subgroup who quit even as recently as 12 months prior to baseline had substantially lower COPD-risk when compared with current smokers in table 3. Thus, the predicted COPD-risk for a 43-year-old ex-smoker of 25 pack-years who quit 5 years earlier, was only low-to-moderate, even in the presence of isolated respiratory symptoms typical of obstructive lung diseases. A similar 2.2-fold variation across occupational classes was also seen, however, all risk-predictions were in the low range (1.12%–2.50%) (online supplemental table E10).

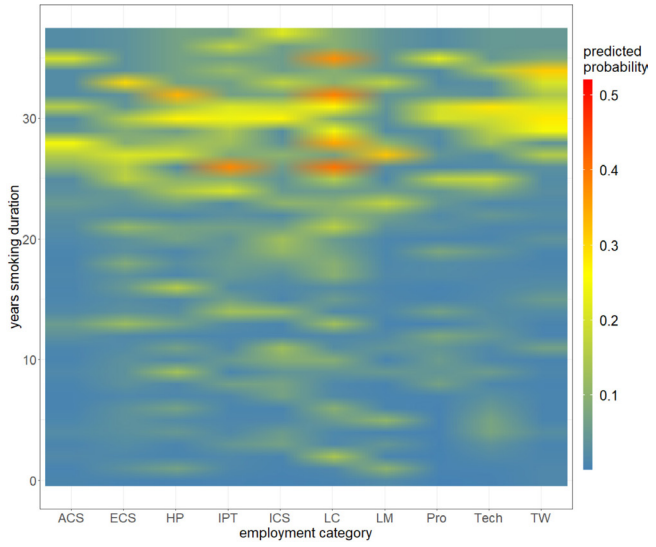

Figure 3 Interaction plot between the effects of increasing smoking duration (0–37 years) and occupation class on post-bronchodilator airflow obstruction at age 53 years. Recalibrated predicted probabilities range between <0.1 (blue) and 0.5 (red). Occupation class categories labelled from right to left: advanced clerical services (ACS), elementary clerical services (ECS), house persons (HP), intermediate production/transport (IPT), intermediate clerical services (ICS), labourer/cleaner/related workers (LC), legislator/manager (LM), professional (Pro), technicians/associate professional (Tech) and trade/related workers (TW).

Predictions for a non-smoker who has active asthma

Predicted probabilities for a 43-year-old female unskilled worker (eg, cleaner) showed that having active (current) asthma in the absence of smoking inferred moderate COPD-risk at age 50s with little variation by age-of-asthma onset (predicted probability 6%–9%, table 3) . The risk-estimate was low for remitted asthma, although the predicted occurrence was not negligible at around 1 in 38 similar persons.

Difference in COPD-risk between groups of quitters and continuing smokers

Four hypothetical examples of asymptomatic 43-year-old unskilled current heavy smokers who were partitioned into subgroups of quitters and continuing smokers over the next 10 years, with or without concurrent asthma at baseline. For current smokers *without* active asthma, the risk-difference in predicted probabilities between those who quit or continued smoking over the next 10 years was 22.5% (4.51% compared with 27.0%, respectively, table 3), which is equivalent to a one in 4.4–fold difference in COPD-risk. For similar smokers *with* active asthma, the risk-difference was 25.6% (16.4% compared

Table 3 Hypothetical examples of individualised predictions by baseline smoking and asthma status in a high-risk occupation: risk difference with and without quitting by age 50s

Predictions by recalled asthma and smoking status for an at-risk occupational group*†	Predicted probability (%)	Predicted occurrence (1/n persons)	Risk category (age in 40s)‡
Smoking status—no asthma or respiratory symptoms			
Non-smoker	0.6	166	Minimal
Past smoker	2.5	40	Low
Current smoker at mean age 43			
Quit smoking by mean age 53	4.5	22	Low
Continued smoking at 53§	27.0	3.7	Very high
Smoking status—adult-onset asthma with wheeze in the last 12 months			
Non-smoker	6.4	16	Moderate
Past smoker	10.8	9.3	High
Current smoker at mean age 43			
Quit smoking by mean age 53	16.4	6.1	High
Continued smoking at 53¶	42.0	2.4	Very high

*Based on a 30 pack-year smoking history starting at age 13 and asthma onset at age 23 years.
†Based on a female worker from an at-risk occupation (eg, labourers and related workers such as cleaners, factory workers, farm and/or kitchen hands).
‡Risk categories: minimal risk if predicted occurrence of 1 in >100 similar persons; low risk if 1 in 20–100 persons; moderate risk if 1 in 10–20 persons; high risk if 1 in 5–10 persons and very high risk if 1 in 1.5–5 persons.
§Same clinical scenario has been presented in online supplemental table E7 (30 pack-years of smoking).
¶Same clinical scenario as in online supplemental table E9 except the predicted probability was for a male worker (42.5%, labelled with ‡).

with 42.0%, respectively or one in 3.9-fold difference in COPD-risk).

DISCUSSION

Using information from questionnaires that is readily accessible from patients and clinicians in a typical clinical scenario, we developed and validated a COPD risk-prediction model from general Australian and European populations aged in their 40s, to calculate 10-year COPD-risks as determined by post-BD airflow obstruction in their 50s. The variables of the final model comprised nine stem questions on known risk factors and resembled those of a basic respiratory assessment that covered participant sex, respiratory symptoms, smoking, asthma and occupation. As indicated by figure 3, online supplemental figure E5–7, our machine learning methods were able to account for the likely interactions between these predictors, especially with regards to smoking and at-risk occupations. Our risk-predictions could potentially inform on further testing of high-risk adults aged in their 40s using spirometry to uncover ‘COPD cases’ which could create opportunities for earlier detection and active intervention. However, the predictions do not relate to actual cases of clinical COPD but to spirometrically defined COPD, with and without symptoms or risk factors.

The baseline prevalence of post-BD airflow obstruction for adults aged in their 40s is low, yet our model had good discriminatory ability. The PPV in the validation subset of symptomatic adults aged 40–49 indicated that the predictions were multiple times the baseline prevalence of post-BD airflow obstruction, and was higher than the recent *Lancet* article that presented machine learning-based predictions of non-fatal adverse effects following an acute coronary syndrome (in theonline supplemental file).[24] However, it is acknowledged that for symptomatic adults who are identified by our risk-prediction model to be at high or very high COPD-risk, approximately 5.6 spirometry tests would be performed to uncover one case of spirometrically defined COPD, with the remainder being false positive results. Using individual case scenarios as examples, our prediction model confirmed high COPD-risk smoking profiles, but also has added to the knowledge base of causal inference by challenging the assumption of dose–response associations with smoking through illustrating two threshold effects: insignificant increases in predicted probability beyond smoking >20 cigarettes/day and an escalating risk for smoking durations longer than 20 years. It also highlighted a moderate COPD-risk for active asthma in non-smokers and discovered the modest predictive ability of respiratory symptoms which in retrospect is not unexpected given active asthma and chronic bronchitis can commonly occur in the absence of airflow obstruction. The modelling also found a 2.2-fold occupational risk and these risk-predictions were surprisingly highest for unskilled workers of lower socioeconomic status rather than for trade workers, and this possibly relates to the healthy worker effect. The 10-year risk-predictions for current smokers in their 40s were substantially lower

for subsequent quitters than for continuing smokers thus supporting more intensive tobacco cessation counselling and support for this age group.

Active case-finding to identify individuals with moderate-to-severe COPD is advocated by expert bodies[25 26] to identify adults with early COPD and reduce morbidity, mortality and economic costs through early intervention, although conclusive evidence to support this initiative is lacking. Spirometry testing has generally been underused as a diagnostic test for COPD,[27 28] despite recommendations for testing to be considered in symptomatic adults with and without known risk factors.[29 30] While active case-finding in smokers is feasible and likely to be cost-effective,[6 31] there has been a lack of action among primary care physicians to pre-emptively manage individuals who have relatively few symptoms.[32] This is typified by the inclusion of early spirometry within the section, 'screening tests of unproven benefit' of the Australian primary care guidelines.[29] This recommendation was largely informed by a lack of direct evidence to determine the benefits and harms of screening in *asymptomatic* adults even when at high risk,[33 34] while comparable screening programmes for coronary heart disease and diabetes[3 4] were based on only limited data before being recommended as part of routine practice.[35 36] Historically, COPD may not be given equal priority by primary care physicians as the disease has traditionally been regarded to be self-inflicted with stigmatisation of affected people,[37] and its multifactorial nature beyond smoking has only recently been appreciated. The health system seems to place the responsibility for COPD prevention primarily with public health initiatives, and thus, establishing the cost-effectiveness of pre-emptive identification and providing integrated research support for practice change would be needed to improve the uptake of spirometry use in primary care.

Our risk-prediction model includes known risk factors as predictors of lung function consistent with COPD, that is, smoking, active asthma[38] and potential hazardous exposures from unskilled jobs, for which preventive management strategies are the cornerstone of best clinical practice. External validation in a similar population-based cohort suggests robustness in our predicted probabilities for individual case scenarios. While we acknowledge that our model alone cannot identify all individuals on an accelerated course to severe airways obstruction, this approach could help identify adults aged in their 40s who are at higher risk and may benefit from serial spirometry to detect rapidly progressive COPD as a 'targeted intervention'. Although the use of a supervised learning model requires careful interpretation of the findings, our individualised risk-predictions might be useful in refining guideline recommendations that consider spirometry testing in adults at least 40 years of age[29] who are heavy smokers,[29 39] symptomatic[29 40] and/or have recurrent chest infections.[39] Similarly, our novel approach to partition current smokers aged in their 40s into either quitters or persistent smokers over the next 10 years could motivate middle-aged smokers to change their behaviours.[41] While we did not account for the reasons underlying quitting and acknowledge that these are distinct participant subgroups, a causal interpretation is biologically plausible given smoking cessation can improve lung function trajectories,[7] and asthma control.[8]

Strengths and limitations

By design, our risk-predictions were based on information that was easily collected and relevant to an age when early COPD begins to manifest clinically and when there is some potential for reversal or at least stabilisation. Our use of randomForest methodology was advantageous over regression methods for prediction as it could inherently accommodate non-linearity, multi-collinearity and multiple interactions (figure 3, online supplemental figure E4). External validation using general population-based data from Europe extends the generalisability to different geographical regions and to a broader age group of 50–59 years old, although validation in non-Caucasian populations is still needed.

Although much larger participant numbers such as those available in administrative health databases could have improved the predictive accuracy, our study design was superior because we used objective and individualised spirometry measurements (rather than ICD-9 codes) and a detailed smoking history. Post-BD spirometry is more relevant to clinically important COPD outcomes than pre-BD measurements,[42] especially for countries with moderate-to-high asthma prevalence such as Australia. Although we did not have post-BD spirometry for the majority of participants in their 40s, we argue that this represents a usual clinical scenario when an individual is assessed for the first time.

Our selection of predictor variables could have limited our model performance as we did not have reliable data on family history of COPD/emphysema, respiratory infections and other air pollutants. Finally, this study was not designed to address causal inference and rate of lung function decline, so caution is advised when interpreting the effect size of quitting smoking on change in COPD-risk and progression to clinical COPD, respectively.

CONCLUSION

This pragmatic and validated COPD risk-prediction model could predict high or very high risk of post-BD airflow obstruction in 10 years' time in Caucasian adults aged 40–49 years. These risk-predictions are especially relevant to COPD in the presence of respiratory symptoms, and to the asthma-COPD overlap (in the presence of current asthma). We have quantified substantial differences in COPD-risk between middle-aged quitters and continuing smokers, which provide rationale to intensify tobacco cessation strategies for smokers less than 50 years of age, especially unskilled workers with a history of asthma. This work has potential to facilitate the

pre-emptive detection of COPD at a much earlier age in primary care settings.

Author affiliations
[1]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, VIC, Australia
[2]Department of Respiratory and Sleep Medicine, The Austin Hospital, Melbourne, VIC, Australia
[3]Institute for Breathing and Sleep (IBAS), Melbourne, VIC, Australia
[4]The Department of Mathematics and Statistics, La Trobe University, Bundoora, VIC, Australia
[5]National Heart and Lung Institute (NHLI), Imperial College London, London, UK
[6]School of Psychology and Public Health, La Trobe University, Melbourne, VIC, Australia
[7]Chandlers Hill Surgery, Adelaide, SA, Australia
[8]Faculty of Health, Arts and Design, Swinburne University of Technology, Hawthorn, VIC, Australia
[9]Department of Lung, Sleep, Allergy and Immunology, Monash Health, Melbourne, VIC, Australia
[10]School of Clinical Sciences, Monash University, Melbourne, VIC, Australia
[11]Adelaide Institute for Sleep Health (AISH), Flinders University, Adelaide, SA, Australia
[12]School of Public Health & Preventive Medicine, Monash University, Melbourne, VIC, Australia
[13]Faculty of Medicine, University of New South Wales, Sydney, NSW, Australia
[14]College of Medicine and Public Health, Flinders University, Adelaide, SA, Australia
[15]School of Medicine, University of Tasmania, Hobart, TAS, Australia

Acknowledgements We acknowledge the TAHS study participants and previous investigators, Dr Heather Gibson, Dr Bryan Gandevia, Dr Harold Silverstone and Dr Norelle Lickiss. We thank Professors Mark Jenkins and John Hopper (Centre for Epidemiology & Biostatistics, VIC), Dr James Markos (Launceston Hospital, TAS), Dr Richard Wood-Baker (Royal Hobart Hospital, TAS) and Dr Iain Feather (Gold Coast Hospital, Queensland) who are investigators of TAHS but not coauthors of this manuscript, for their assistance with obtaining funds and data collection. We also thank A/Professor David Johns, Dr Melanie Matheson, Professor Graham Giles, Professor Lyle Gurrin, Professor Alan James, Professor Nicholas Zwar, Professor Peter Sly and Professor Nicholas de Klerk for their input into the study design and methodology, but who are not coauthors of this manuscript. Furthermore, we recognise all the study site coordinators and respiratory scientists who collected data in the lung function laboratories of Tasmania, Victoria, Queensland and New South Wales; the research interviewers and data entry operators; and the organisational roles of Ms Cathryn Wharton and Dr Desiree Mészáros. We thank the late Stephen Morrison (University of Queensland) for his assistance with obtaining funds and collecting data and recognise the Archives Office of Tasmania for providing data from the 1968 and 1974 TAHS questionnaires and copies of the school medical records.We also formally acknowledge the Investigators of the European Community Respiratory Health Survey (ECRHS) for their role in obtaining research funding and collection of data, as well as the support of the study coordination by the European Commission (018996), Medical Research Council, and separate grant funding for the local studies as outlined below. (A) Scientific teams of ECRHS: ECRHS II Coordinating Centre (London): P Burney, S Chinn, C Luczynska†, DJ, J Knox. Project Management Group: P Burney (Project leader-UK) S Chinn (UK), C Luczynska† (UK), D Jarvis (UK), P Vermeire† (Antwerp), J Bousquet (Montpellier), JH (Erfurt), M Wjst (Munich) RdM† (Verona), JMA (Barcelona), J Sunyer (Barcelona) CJ (Uppsala), U Ackermann-Liebrich (Basel), N Kuenzli (University of Basel and University of Southern California, Los Angeles, USA); F Neukirch (Paris), ECRHS II Participating Centres: Australia: Melbourne (M Abramson, E H Walters, J Raven), Belgium: Antwerp South, Antwerp Central (P Vermeire, JW, M van Sprundel, V Nelen) Estonia: Tartu (R Jõgi, A Soon), France: Bordeaux (A Taytard, CR),) Grenoble (IP, J Ferran-Quentin), Paris (F Neukirch, BL, R Liard, M Zureik) Montpellier (J Bousquet, P J Bousquet), Germany: Erfurt (JH, M Wjst, C Frye, I Meyer) Hamburg (H Magnussen, D Nowak), Iceland: Reykjavik (TG, E Bjornsson, D Gislason, K B Jörundsdóttir) Italy: Pavia (A Marinoni, S Villani, M Ponzio, F Frigerio, M Comelli, M Grassi, I Cerveri, AC) Turin: (RB, M Bugiani, P Piccioni, E Caria, A Carosso, E Migliore, G Castiglioni) Verona: RdM†, G Verlato, E Zanolin, SA, A Poli, V Lo Cascio, M Ferrari, I Cazzoletti) Netherlands: Groningen (M Kerkhof) Norway: Bergen (A Gulsvik, E Omenaas, CS, B Laerum) Spain: Albacete (JM-MR, E Almar, M Arévalo, C Boix, G González, J M Ignacio García, J Solera, JDamián) Barcelona (JMA, J Sunyer, M Kogevinas, JPZ, X Basagaña, A Jaen, F Burgos, C Acosta) Galdakao: (N Muñozguren, J Ramos, IU, U Aguirre) Huelva: (J Maldonado, AP-V, J L Sanchez) Oviedo (F Payo, I Huerta, A de la Vega, L Palenciano, J Azofra, A Cañada) Sweden: Göteborg (K. Toren, L Lillienberg, A C Olin, B Balder, A Pfeifer-Nilsson, R Sundberg) Uppsala: (CJ, G Boman, D Norback, G Wieslander, M Gunnbjornsdottir) Umeå (E Norrman, M Soderberg, K A Franklin, B Lundback, BF, L Nystrom) Switzerland: Basel (N Küenzli, B Dibbert, M Hazenkamp, M Brutsche, U Ackermann-Liebrich) UK: Caerphilly (M Burr†, J Layzell) Ipswich (DJ, R Hall, D Seaton) Norwich (DJ, B Harrison), ECRHS III Coordinating Centre (London): D Jarvis, P Burney, M Tumilty, J Potts. Project Management Group: D Jarvis (UK), P Burney (UK), JH (Erfurt), RdM† (Verona), JMA (Barcelona) CJ (Uppsala), K Toren (Goteburg), T Gislasson (Iceland) T Rochat (Basel), B Leyneart (Paris) C Svanes (Bergen) JW (Antwerp) JPZ (Barcelona). ECRHS III Participating Centres: Australia: Melbourne (M Abramson, G Benke, S Dharmage, B Thompson, S Kaushik, M Matheson). Belgium: South Antwerp & Antwerp City (JW, H Bentouhami, V Nelen) Estonia: Tartu (R J õ gi, H Orru) France: Bordeaux (CR, P O Girodet) Grenoble (IP, V Siroux, J Ferran, J L Cracowski) Montpellier (PD, A Bourdin, I Vachier) Paris (BL, D Soussan, D Courbon, C Neukirch, L Alavoine, X Duval, I Poirier) Germany: Erfurt (JH, E Becker, G Woelke, O Manuwald) Hamburg (H Magnussen, D Nowak, A-MK), Iceland: Reykjavik (TG, B Benediktsdottir, D Gislason, E S Arnardottir, M Clausen, G Gudmundsson, L Gudmundsdottir, H Palsdottir, K Olafsdottir, S Sigmundsdottir, K Bara-Jörundsdottir), Italy: Pavia (I Cerveri, AC, A Grosso, F Albicini, E Gini, E M Di Vincenzo, V Ronzoni, S Villani, F Campanella, M Gnesi, F Manzoni, L Rossi, O Ferraro) Turin: (M Bugiani, RB, P Piccioni, R Tassinari, V Bellisario, G Trucco) Verona: (RdM†, SA, L Calciano, L Cazzoletti, M Ferrari, A M Fratta Pasini, F Locatelli, P Marchetti, A Marcon, E Montoli, G Nguyen, M Olivieri, C Papadopoulou, C Posenato, G Pesce, P Vallerio, G Verlato, E Zanolin), Netherlands: Groningen (HMB), Norway: (CS, E Omenaas, A Johannessen, T Skorge, F Gomez Real) Spain: Albacete (JM-MR, E Almar, A Mateos, S García, A Núñez, P López, R Sánchez, E Mancebo) Barcelona:(JMA, JPZ, J Garcia-Aymerich, M Kogevinas, X Basagaña, A E Carsin, F Burgos, C Sanjuas, S Guerra, B Jacquemin, P Davd and Galdakao: N Muñozguren, IU, U Aguirre, S Pascual) Huelva: (J Antonio Maldonado, AP-V, J Luis Sánchez, L Palacios, Oviedo: (F Payo, I Huerta, N Sánchez, M Fernández, B Robles) Sweden: Göteborg (K Torén, M Holm, J-L Kim, A-C Olin, A Dahlman-Höglund), Umea (BF, L Braback, L Modig, B Järvholm, H Bertilsson, K A Franklin, C Wahlgreen) Uppsala: (B Andersson, D Norback, U Spetz Nystrom, G Wieslander, G M Bodinaa Lund, KNisser), Switzerland: Basel (NMP-H, N Künzli, D Stolz, C Schindler, T Rochat, J M Gaspoz, E Zemp Stutz, M Adam, C Autenrieth, I Curjuric, J Dratva, A Di Pasquale, R Ducret-Stich, E Fischer, L Grize, A Hensel, D Keidel, A Kumar, M Imboden, N Maire, A Mehta, H Phuleria, M Ragettli, M Ritter, E Schaffner, G A Thun, A Ineichen, T Schikowski, M Tarantino, M Tsai, UK:London (PB, DJ, S Kapur, RN, J Potts,) Ipswich: (DJ, M Tumilty, N Innes) Norwich: (DJ, M Tumilty, A Wilson). (B) Local funding grants for ECRHS testing centres: ECRHS II: Australia: NHMRC grant code 980894; Belgium: Antwerp: Fund for Scientific Research (grant code, G.0402.00), University of Antwerp, Flemish Health Ministry; Estonia: Tartu Estonian Science Foundation grant no. 4350, France: (all) Programme Hospitalier de Recherche Clinique—Direction de la Recherche Clinique (DRC) de Grenoble 2000 number 2610, Ministry of Health, Ministère de l'Emploi et de la Solidarité, Direction Génerale de la Santé, Centre Hospitalier Universitaire (CHU) de Grenoble, Bordeaux: Institut Pneumologique d'Aquitaine; Grenoble: Comite des Maladies Respiratoires de l'Isere Montpellier: Aventis (France), Direction Regionale des Affaires Sanitaires et Sociales Languedoc-Roussillon; Paris: Union Chimique Belge-Pharma (France), Aventis (France), Glaxo France; Germany: Erfurt GSF—National Research Centre for Environment and Health, Deutsche Forschungsgemeinschaft (grant code, FR1526/1-1), Hamburg: GSF—National Research Centre for Environment and Health, Deutsche Forschungsgemeinschaft (grant code, MA 711/4-1), Iceland: Reykjavik, Icelandic Research Council, Icelandic University Hospital Fund; Italy: Pavia GlaxoSmithKline Italy, Italian Ministry of University and Scientific and Technological Research (MURST), Local University Funding for Research 1998 and 1999; Turin: Azienda Sanitaria Locale 4 Regione Piemonte (Italy), Azienda Ospedaliera Centro Traumatologico Ospedaliero/Centro Traumatologico Ortopedico—Istituto Clinico Ortopedico Regina Maria Adelaide Regione Piemonte Verona: Ministero dell'Universita e della Ricerca Scientifica (MURST), Glaxo Wellcome SPA, Norway: Bergen: Norwegian Research Council, Norwegian Asthma and Allergy Association, Glaxo Wellcome AS, Norway Research Fund; Spain: Fondo de Investigacion Santarias (grant codes, 97/0035-01, 99/0034-01 and 99/0034 02), Hospital Universitario de Albacete, Consejeria de Sanidad; Barcelona: Sociedad Espanola de Neumologıa y Cirugıa Toracica, Public HealthService (grant code, R01 HL62633-01), Fondo de Investigaciones Santarias (grant codes, 97/0035-01, 99/0034-01 and 99/0034-02), Consell Interdepartamentalde Recerca i Innovacio Tecnologica (grant code, 1999SGR 00241), Instituto de Salud Carlos III; Red de Centros de Epidemiologıa y Salud Publica, C03/09, Red de Bases moleculares y fisiologicas de las Enfermedades Respiratorias, C03/011, and Red de Grupos Infancia y Medio Ambiente G03/176; Huelva: Fondo de Investigaciones Santarias (grant codes, 97/0035-01, 99/0034-01 and 99/0034-02); Galdakao: Basque Health Department Oviedo: Fondo de Investigaciones Sanitaria (97/0035-02, 97/0035, 99/0034-01, 99/0034-02, 99/0034-04, 99/0034-06, 99/350, 99/0034-07), European Commission (EU-PEAL PL01237), Generalitat de Catalunya (CIRIT 1999 SGR 00214), Hospital Universitario de Albacete, Sociedad Española de Neumología y Cirugía Torácica (SEPAR R01

HL62633-01), Red de Centros de Epidemiología y Salud Pública (C03/09), Red de Bases moleculares y fisiológicas de las Enfermedades Respiratorias (C03/011) and Red de Grupos Infancia y Medio Ambiente (G03/176); 97/0035-01, 99/0034-01 and 99/0034-02); Sweden: Göteborg, Umea, Uppsala: Swedish Heart Lung Foundation, Swedish Foundation for Health Care Sciences and Allergy Research, Swedish Asthma and Allergy Foundation, Swedish Cancer and Allergy Foundation, Swedish Council for Working Life and Social Research (FAS), Switzerland: Basel Swiss National Science Foundation, Swiss Federal Office for Education and Science, Swiss National Accident Insurance Fund; UK: Ipswich and Norwich: Asthma UK (formerly known as National Asthma Campaign). ECRHS III: Australia: NHMRC (grant code 1007965), Belgium: Antwerp South, Antwerp City: Research Foundation Flanders (FWO), grant code G.0.410.08.N.10 (both sites), Estonia: Tartu-SF0180060s09 from the Estonian Ministry of Education. France: (all) Ministère de la Santé. Programme Hospitalier de Recherche Clinique (PHRC) National 2010. Bordeaux: INSERM U897 Université Bordeaux Segalen, Grenoble: Comite Scientifique AGIR adom 2011. Paris: Agence Nationale de la Santé, Région Ile de France, domaine d'intérêt majeur (DIM) Germany: Erfurt: German Research Foundation HE 3294/10-1, Hamburg: German Research Foundation MA 711/6-1, NO 262/7-1, Iceland: Reykjavik, The Landspitali University Hospital Research Fund, University of Iceland Research Fund, ResMed Foundation, California, USA, Orkuveita Reykjavikur (Geothermal plant), Vegagerðin (The Icelandic Road Administration, ICERA). Italy: all Italian centres were funded by the Italian Ministry of Health, Chiesi Farmaceutici SpA. In addition, Verona was funded by Cariverona Foundation, Education Ministry (MIUR). Norway: Norwegian Research council grant no 214123, Western Norway Regional Health Authorities grant no 911631, Bergen Medical Research Foundation. Spain: Fondo de Investigación Sanitaria (PS09/02457, PS09/00716, PS09/01511, PS09/02185, PS09/03190), Servicio Andaluz de Salud, Sociedad Española de Neumología y Cirurgía Torácica (SEPAR 1001/2010); Sweden: all centres were funded by The Swedish Heart and Lung Foundation, The Swedish Asthma and Allergy Association, The Swedish Association against Lung and Heart Disease. Fondo de Investigación Sanitaria (PS09/02457), Barcelona: Fondo de Investigación Sanitaria (FIS PS09/00716), Galdakao: Fondo de Investigación Sanitaria (FIS 09/01511), Huelva: Fondo de Investigación Sanitaria (FIS PS09/02185) and Servicio Andaluz de Salud Oviedo: Fondo de Investigación Sanitaria (FIS PS09/03190). Sweden: all centres were funded by The Swedish Heart and Lung Foundation, The Swedish Asthma and Allergy Association, The Swedish Association against Lung and Heart Disease. Swedish Research Council for Health, Working Life and Welfare (FORTE) Göteborg: also received further funding from the Swedish Council for Working Life and Social Research. Umea also received funding from Vasterbotten Country Council ALF grant. Switzerland: The Swiss National Science Foundation (grant nos 33CSCO-134276/1, 33CSCO-108796, 3247BO-104283, 3247BO-104288, 3247BO-104284, 3247-065896, 3100-059302, 3200-052720, 3200-042532, 4026-028099). The Federal Office for Forest, Environment and Landscape, The Federal Office of Public Health, The Federal Office of Roads and Transport, The Canton's Government of Aargan, Basel-Stadt, Basel-Land, Geneva, Luzern, Ticino, Valais and Zürich, the Swiss Lung League, the Canton's Lung League of Basel Stadt/Basel, Landschaft, Geneva, Ticino, Valais and Zurich, SUVA, Freiwillige Akademische Gesellschaft, UBS Wealth Foundation, Talecris Biotherapeutics GmbH, Abbott Diagnostics, European Commission 018996 (GABRIEL), Wellcome Trust WT 084703MA, UK: Medical Research Council (grant no 92091). Support was also provided by the National Institute for Health Research through the Primary Care Research Network.

Contributors Funding acquisition: SCD, EHW, MJA, DLJ. Data curation and resources: SCD, EHW, MJA, BRT, PST, DLJ. Conceptualisation: JLP, SCD, CM, DLJ, CFM, KH, PF, TB, AB, AJL, CJL, DSB, DT, JAB, BE, GSH, RA, GPB. Data access and verification: JLP, DV, SCD, DLJ. Formal analysis: JLP, DV, KS, CFM. Investigation methodology and validation: DV, KS, CFM. Manuscript writing - original draft: JLP, DV. Manuscript writing - review & editing: all authors especially CFM, TB, MJA, BE. Project administration: SCD, DLJ. Guarantor: JLP. JLP and DV had full data access and can verify the analysis. CM and SCD contributed equally as senior authors.

Funding The TAHS was supported by the National Health and Medical Research Council (NHMRC) of Australia, research grants 299901 and 1021275; the University of Melbourne; Clifford Craig Foundation; the Victorian, Queensland and Tasmanian Asthma Foundations; Royal Hobart Hospital; Helen MacPherson Smith Trust; GlaxoSmithKline; and John L Hopper. JP, AL and SCD are funded through the NHMRC of Australia. The ECRHS was supported by grants from the European Commission (018996) and Medical Research Council (ECRHS III no. 92091), and multiple local grants that supported study testing centres of ECRHS II and III which have been listed in the acknowledgement section. These sponsors of the study had no role in study design, data collection, data analysis, data interpretation or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Competing interests JP received a travel grant supported by Boehringer-Ingelheim, and together with MJA, EHW, AL, CL and SCD, holds an investigator-initiated grant from GlaxoSmithKline for unrelated research. SCD additionally holds an investigator-initiated grant from AstraZeneca for unrelated research. MJA also holds investigator-initiated grants from Pfizer, Boehringer-Ingelheim and Sanofi for unrelated research; has undertaken an unrelated consultancy for and received assistance with conference attendance from Sanofi; and received a speaker's fee from GlaxoSmithKline. BRT has received unrelated speaker and consultancy fees from Chiesi, Mundipharma and 4D medical. KH has received personal fees and non-financial support from Astra Zeneca, GlaxoSmithKline, Novartis, Chiesi, Boehringer Ingelheim and Teva outside the submitted work. AL has additionally received non-financial support from Primus Pharmaceuticals for unrelated research. No other authors reported financial disclosures.

Patient consent for publication Not applicable.

Ethics approval TAHS was approved by Human Ethics Review Committees at all participating institutions, principally The Universities of Melbourne (040375) and Tasmania (H0012710). ECRHS II and III were performed with the approval of the corresponding local/regional committees for all participating centres (refer reference 23). Written informed consent was obtained from all participants.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. TAHS is a cohort study with data that has been prospectively collected since 1968 and will be an ongoing resource for future epidemiological analyses. Data collection protocols have been detailed in the TAHS cohort profile paper published in 2016 (Matheson *et al* 2016 doi: 10.1093/ije/dyw028). The raw data have not been made widely available, but expressions of interest can be discussed with the corresponding author, Dr J Perret, and/or principal investigator, Professor S Dharmage, on an individual basis. ECRHS is a cohort study with data that has been prospectively collected since 1990 and will be an ongoing resource for future epidemiological analyses. Data collection protocols are detailed at https://www.ecrhs.org/. The raw data have not been made widely available, but expressions of interest can be discussed with the principal investigator, Professor D Jarvis, on an individual basis.

ORCID iDs
Jennifer L Perret http://orcid.org/0000-0001-7034-0615
Peter Frith http://orcid.org/0000-0003-3265-0131
Michael J Abramson http://orcid.org/0000-0002-9954-0538

REFERENCES

1 Australian Institute of Health and Welfare. *Chronic obstructive pulmonary disease (COPD). Cat. No. ACM 35*. Canberra: AIHW, 2019. https://www/aihw.gov.au/reports/chronic-respiratory-conditions/copd
2 Falster M, Jorm L. *A guide to the potentially preventable hospitalisations indicator in Australia*. Sydney: Centre for Big Data Research in Health, University of New South Wales in consultation with Australian Commission on Safety and Quality in Health Care and Australian Institute of Health and Welfare, 2017.
3 Wilson PW, D'Agostino RB, Levy D, *et al*. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97:1837–47.
4 Chen L, Magliano DJ, Balkau B, *et al*. AUSDRISK: an Australian type 2 diabetes risk assessment tool based on demographic, lifestyle and simple anthropometric measures. *Med J Aust* 2010;192:197–202.
5 Wilson JMG, Jungner G. *Principles and practice of screening for disease*. Geneva: World Health Organization, 1968.
6 Lambe T, Adab P, Jordan RE, *et al*. Model-based evaluation of the long-term cost-effectiveness of systematic case-finding for COPD in primary care. *Thorax* 2019;74:730–9.

7 Anthonisen NR, Connett JE, Murray RP. Smoking and lung function of lung health study participants after 11 years. *Am J Respir Crit Care Med* 2002;166:675–9.
8 McLeish AC, Zvolensky MJ. Asthma and cigarette smoking: a review of the empirical literature. *J Asthma* 2010;47:345–61.
9 Perret JL, Simons K, Vicendese D, *et al*. Optimizing prediction of the lung function features of COPD. *Chest* 2020;157:738.
10 Matheson MC, Bowatte G, Perret JL, *et al*. Prediction models for the development of COPD: a systematic review. *Int J Chron Obstruct Pulmon Dis* 2018;13:1927–35.
11 Leisman DE, Harhay MO, Lederer DJ, *et al*. Development and reporting of prediction models: guidance for authors from editors of respiratory, sleep, and critical care journals. *Crit Care Med* 2020;48:623–33.
12 Chen W, Sin DD, FitzGerald JM, *et al*. An individualized prediction model for long-term lung function trajectory and risk of COPD in the general population. *Chest* 2020;157:547–57.
13 Moons KGM, Altman DG, Reitsma JB, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1–73.
14 Matheson MC, Abramson MJ, Allen K, *et al*. Cohort profile: the Tasmanian longitudinal health study (TAHS). *Int J Epidemiol* 2017;46:407–8.
15 Perret JL, Lodge CJ, Lowe AJ, *et al*. Childhood pneumonia, pleurisy and lung function: a cohort study from the first to sixth decade of life. *Thorax* 2020;75:28–37.
16 Burney PG, Luczynska C, Chinn S, *et al*. The European community respiratory health survey. *Eur Respir J* 1994;7:954–60.
17 Miller MR, Hankinson J, Brusasco V, *et al*. Standardisation of spirometry. *Eur Respir J* 2005;26:319–38.
18 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;40:1324–43.
19 Pellegrino R, Viegi G, Brusasco V, *et al*. Interpretative strategies for lung function tests. *Eur Respir J* 2005;26:948–68.
20 Breiman L. Random forest. *Mach Learn* 2001;45:5–32.
21 Thiele C, Hirschfeld G. cutpointr: improved estimation and validation of optimal cutpoints in R. *J Stat Software* 2020 https://arxiv.org/abs/2002.09209
22 Bostrom H. *Calibrating random forests*. KTH Royal Institute of Technology, 2008. https://dl.acm.org/doi/10.1109/ICMLA.2008.107
23 Bédard A, Carsin A-E, Fuertes E, *et al*. Physical activity and lung function-Cause or consequence? *PLoS One* 2020;15:e0237769.
24 D'Ascenzo F, De Filippo O, Gallone G, *et al*. Machine learning-based prediction of adverse events following an acute coronary syndrome (praise): a modelling study of pooled datasets. *Lancet* 2021;397:199–207.
25 National Heart, Lung and Blood Institute. A case-finding strategy for moderate-to-severe COPD in the United States, 2008. Available: https://www.nhlbi.nih.gov/events/2008/case-finding-strategy-moderate-severe-copd-united-states [Accessed 30 Sep 19].
26 Lung Foundation Australia. Position paper. COPD case finding in community settings. Available: https://lungfoundation.com.au/wp-content/uploads/2018/11/Information-Paper-COPD-Case-Finding-position-paper-Oct2019.pdf
27 COPD. *MedicineInsight post-market surveillance report number 11*. Sydney: NPS MedicineWise, 2017.
28 O'Sullivan JW, Albasri A, Nicholson BD, *et al*. Overtesting and undertesting in primary care: a systematic review and meta-analysis. *BMJ Open* 2018;8:e018557.
29 The Royal Australian College of General Practitioners. *Guidelines for preventive activities in general practice*. 9th edn. East Melbourne, Vic: RACGP, 2018.
30 Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease: 2021 report, 2021. Available: https://goldcopd.org/wp-content/uploads/2020/11/GOLD-REPORT-2021-v1.1-25Nov20_WMV.pdf [Accessed 21 May 21].
31 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;4:720–30.
32 Haroon S, Adab P, Dickens AP, *et al*. Impact of COPD case finding on clinical care: a prospective analysis of the TargetCOPD trial. *BMJ Open* 2020;10:e038286.
33 US Preventive Services Task Force (USPSTF), Siu AL, Bibbins-Domingo K, *et al*. Screening for chronic obstructive pulmonary disease: US preventive services Task force recommendation statement. *JAMA* 2016;315:1372–7.
34 Guirguis-Blake JM, Senger CA, Webber EM, *et al*. Screening for chronic obstructive pulmonary disease: evidence report and systematic review for the US preventive services Task force. *JAMA* 2016;315:1378–93.
35 US Preventive Services Task Force, Davidson KW, Barry MJ, *et al*. Screening for prediabetes and type 2 diabetes: US preventive services Task force recommendation statement. *JAMA* 2021;326:736–43.
36 Grant RW, Gopalan A, Jaffe MG. Updated USPSTF screening recommendations for diabetes: identification of abnormal glucose metabolism in younger adults. *JAMA Intern Med* 2021;181:1284–6.
37 Australian Department of Health. National strategic action plan for lung conditions, 2019. Available: https://www.health.gov.au/resources/publications/national-strategic-action-plan-for-lung-conditions [Accessed 31 Oct 19].
38 Perret JL, Dharmage SC, Matheson MC, *et al*. The interplay between the effects of lifetime asthma, smoking, and atopy on fixed airflow obstruction in middle age. *Am J Respir Crit Care Med* 2013;187:42–8.
39 Global Initiative for Chronic Obstructive Lung Disease. *Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease*, 2020.
40 Chronic obstructive pulmonary disease in over 16s: diagnosis and management, 2019. Available: nice.org.uk
41 Bize R, Burnand B, Mueller Y, *et al*. Biomedical risk assessment as an aid for smoking cessation. *Cochrane Database Syst Rev* 2012;12:CD004705.
42 Bhatta L, Leivseth L, Carslake D, *et al*. Comparison of pre- and post-bronchodilator lung function as predictors of mortality: the HUNT study. *Respirology* 2020;25:401–9.