Validation of artificial intelligence spirometry diagnostic support software in primary care : a blinded diagnostic accuracy study.

Validation of artificial intelligence spirometry diagnostic support software in primary care: a blinded diagnostic accuracy study

Anthony Sunjaya[1,2], George D. Edwards[2], Jennifer Harvey[2], Karl Sylvester[3], Joanna Purvis[4], Matthew Rutter[3], Joanna Shakespeare[5], Vicky Moore[6], Ethaar El-Emir[2], Gillian Doe[7], Karolien Van Orshoven[8], Suhani Patel[2], Maarten de Vos[8,9], Ahmed Elmahy[8], Benoit Cuyvers[8], Paul Desbordes[8], Satesh Sehdev[10], Rachael A. Evans[7], Michael D. Morgan[11], Richard Russell[12], Ian Jarrold[13], Nannette Spain[14], Stephanie Taylor[15], David A. Scott[16], A. Toby Prevost[17], Nicholas S. Hopkinson[18], Samantha Kon[2], Marko Topalovic[8] and William D-C. Man[2,17,18]

[1]The George Institute for Global Health, UNSW Sydney and Imperial College London, Sydney, NSW, Australia. [2]Harefield Respiratory Research Group, Royal Brompton and Harefield Hospitals, Guy's and St Thomas' NHS Foundation Trust, London, UK. [3]Respiratory Physiology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [4]Cardio-Respiratory Department, George Eliot Hospital NHS Trust, Nuneaton, UK. [5]Department of Respiratory Medicine, University Hospitals Coventry and Warwickshire NHS Trust, Coventry, UK. [6]Respiratory and Sleep Sciences, University Hospitals Coventry and Warwickshire NHS Trust, Coventry, UK. [7]Department of Respiratory Sciences, University of Leicester, NIHR Biomedical Research Centre Leicester – Respiratory, Leicester, UK. [8]ArtiQ NV, Leuven, Belgium. [9]ESAT-STADIUS, Department of Electrical Engineering and Department of Development and Regeneration, KU Leuven, Leuven, Belgium. [10]The Confederation, Hillingdon Community Interest Company, London, UK. [11]Wolfson Palliative Care Research Centre, Hull and York Medical School, University of Hull, Cottingham, UK. [12]King's Centre for Lung Health, Peter Gorer Department of Immunobiology, King's College London, London, UK. [13]Asthma+Lung UK, London, UK. [14]Patient and public involvement representative. [15]Wolfson Institution of Population Health, Queen Mary University, London, UK. [16]Southampton Health Technology Assessments Centre, University of Southampton, Southampton, UK. [17]Cicely Saunders Institute, King's College London, London, UK. [18]National Heart and Lung Institute, Imperial College London, London, UK.

Corresponding author: William D-C. Man (w.man@nhs.net)

Shareable abstract (@ERSpublications)
AI spirometry interpretation software achieved high sensitivity and specificity in identifying COPD from spirometry and basic demographic data, and may support the accurate diagnosis of COPD in primary care https://bit.ly/4bDwN7s

Cite this article as: Sunjaya A, Edwards GD, Harvey J, *et al*. Validation of artificial intelligence spirometry diagnostic support software in primary care: a blinded diagnostic accuracy study. *ERJ Open Res* 2025; 11: 00116-2025 [DOI: 10.1183/23120541.00116-2025].

This article has an editorial commentary: https://doi.org/10.1183/23120541.00353-2025

Received: 25 Jan 2025
Accepted: 21 Feb 2025

Abstract

Objective and design The objective of the present study was to assess the discriminative accuracy of artificial intelligence (AI) software to identify COPD and other chronic respiratory diseases from primary care spirometry. This was a diagnostic study with blinded analysis.

Methods Retrospective hand-held spirometry data from consecutive patients attending primary care clinics in Hillingdon (London, UK) between September 2015 and March 2019 were used. The index diagnosis was the "preferred" diagnosis determined by AI software (highest probability) using supervised random-forest machine learning to interpret raw spirometry data and basic demographics. The reference diagnosis was based on the consensus of expert pulmonologists with access to primary and secondary care medical notes and results of relevant investigations. Cross-tabulation of the index test results by the results of the reference standard for COPD and other respiratory disease categories provided the main outcome measures.

Results In this primary care spirometry dataset from 1113 patients, 543 (48.8%) had a reference diagnosis of COPD. AI preferred diagnosis detected 456, achieving a sensitivity of 84.0% (95% CI 80.6–87.0%), specificity of 86.8% (83.8–89.5%), accuracy of 85.4% (83.2–87.5%) with area under curve (AUC) of 0.914 (0.896–0.930). AI preferred diagnosis identified 187 out of 249 patients with reference diagnosis of interstitial lung disease and 59 out of 107 patients with asthma, with AUCs of 0.900 (0.880–0.916) and 0.814 (0.790–0.836), respectively.

Conclusion AI software achieved high sensitivity and specificity in identifying COPD using spirometry and basic demographic data and may support accurate diagnosis of COPD in primary care. AI software performed less well for other chronic respiratory disease categories.

Introduction

COPD is a leading cause of mortality and morbidity globally [1]. Spirometry is recommended to diagnose COPD [2]. However, spirometry provision in primary care is suboptimal, particularly following the coronavirus disease 2019 pandemic [3]. Only 13.4% of spirometry performed in primary care fully meets international technical criteria [4], with 40% failing at least one quality criterion [5]. In addition, primary care clinicians have low levels of confidence in identifying technical errors or interpreting spirometry [6] and there are poor levels of agreement in spirometry interpretation between primary care staff and pulmonologists [7]. Important consequences include underdiagnosis, misdiagnosis and unnecessary referral to secondary care [8, 9].

Recent studies have evaluated an artificial intelligence (AI) pulmonary function interpretation software, developed using random-forest machine learning. The AI software estimates the probability of respiratory disease categories from pulmonary function tests (PFTs), comprising spirometry, gas transfer and lung volume data. This AI software outperformed trained and trainee pulmonologists in the interpretation of hospital-based PFTs [10].

The AI PFT interpretation software was developed and validated in hospitals across Belgium and against pulmonologists (including trainees) from 16 European countries [10], but has not been validated in the primary care setting. In addition, the performance of the software has not been tested on spirometry data alone; spirometry often being the only point-of-care lung function test available in most healthcare settings including primary care. Spirometry performed in primary care poses specific problems for AI software, as the technical quality may be suboptimal [4].

The aim of the current study was to assess the diagnostic performance of an AI diagnostic support software in the identification of COPD and other chronic respiratory diseases when provided with primary care spirometry data against a clinical reference diagnosis (consensus of pulmonologists with access to medical records and investigations). COPD was selected as the primary outcome as it is the only respiratory condition where spirometry is part of the core diagnostic criteria.

Methods

This was a retrospective, blinded diagnostic validation study. Reporting was done in accordance with the Standards for Reporting Diagnostic Accuracy Studies (STARD) guidelines [11]. The study was approved by the United Kingdom Health Research Authority (IRAS project identifier 314058) and pre-registered in Clinicaltrials.gov (identifier NCT05648227).

Patient and public involvement

The study was supported by representatives from Asthma+Lung UK (I. Jarrold and N.S. Hopkinson) and patient and public involvement member (N. Spain), who were invited to and regularly attended steering group meetings throughout the design, conduct and data interpretation of the study.

Primary care validation dataset

The retrospective spirometry data used in this study was collected between September 2015 and March 2019 in primary care clinics in Hillingdon borough, Northwest London. The clinics provided a variety of clinical services including spirometry, annual respiratory review, pulmonary rehabilitation assessment and home oxygen assessments, and were accessed through direct referral by general practitioner practices in Hillingdon.

Inclusion criteria for this study were adults aged ⩾18 years; referred by a general practitioner working in Hillingdon borough; presenting with at least one respiratory symptom (cough, wheeze, shortness of breath, reduced exercise tolerance); undergoing supervised spirometry performed as part of routine clinical assessment. Cases where spirometry was conducted as part of pre-operative assessment, or performed at home without supervision were excluded.

Spirometry (without bronchodilator testing) was performed by nonrespiratory physiologist health practitioners. All had completed an internal training and competency programme, but none were accredited by the Association for Respiratory Technology & Physiology (ARTP). All spirometry was performed using a portable, hand-held Easy One World Spirometer 2001 model using EasyOne connect software version 3.9.2.4 (ndd Medical Technologies, Zurich, Switzerland). Global Lung Function Initiative 2012 reference equations were used [12]. The community respiratory service was not commissioned or equipped to provide before and after bronchodilator responsiveness studies or fractional exhaled nitric oxide measurements.

Index test: machine-learning model

Anonymised raw spirometry data (time, flow, volume curves, demographic information (age, sex, ethnicity, height, weight, smoking pack-years)) was provided to a nonclinical software engineer who analysed the data using ArtiQ.Spiro developed by ArtiQ (Leuven, Belgium; www.artiq.eu). On average, <1 s was needed to generate the analysis for each case and ∼5 s to generate the example Portable Document Format output (supplementary figure S1).

The machine-learning model for interpretation of lung function tests was built using data from 1430 subjects attending hospital pulmonary function tests [13, 14]. In these datasets, a final diagnosis was established by a clinician using all additional tests deemed necessary, the patient's history and PFTs and was validated by an *ad hoc* installed expert panel or by the clinical expert panel taking care of the patients in follow-up.

Internal 10-fold cross-validation tuned the machine-learning model, with the best model resulting in a diagnostic accuracy of 74%. To obtain an unbiased estimate of accuracy and validate findings, the model was run at the Leuven pulmonary service on a randomly selected sample of 136 subjects. The model demonstrated a consistent diagnostic accuracy of 76% [10]. This software was further validated in an independent multicentre study, demonstrating that AI outperformed human interpretation and assigned a correct diagnosis in 82% of all cases [4].

The existing model [10] was adapted for primary care using spirometry data from a representative dataset (UK Biobank (UKBB)) [15]. To achieve this, a balanced bagging classifier model (a machine-learning technique useful to train imbalanced datasets) was trained on 1609 subjects who underwent lung function testing in a hospital, complemented with 500 healthy individuals from UKBB. The model was then tested on an independent dataset extracted from UKBB. Patients were included with a clinician diagnosis of asthma, bronchiectasis, COPD and interstitial lung disease (ILD) with acceptable-quality spirometry. The AI model achieved overall accuracy of 76%.

The index test was defined before the commencement of the study. No clinical history or reference standard results were known to the AI software company or software engineer before, during or subsequent to data analysis. The software engineer did not communicate with the clinical team and had no access to medical records, and only communicated with an independent researcher (A. Sunjaya) based in Sydney, Australia. Only specific deidentified data was provided for the software engineer (information readily available on the spirometry report: age, height, weight, sex, smoking status (current, never) and ethnicity) to feed into the AI software.

Reference standard

No index test results were known to the reference test adjudicators. All reference tests were defined before commencement of the trial. Real-world primary and secondary care medical records (up to 24 months from spirometry), and the results of relevant investigations for all participants were reviewed independently by two pulmonologists from the Royal Brompton and Harefield hospitals, United Kingdom. Within this cohort, participants' care was supervised by 48 general practices supported by four secondary care organisations. There was no one unified diagnostic pathway, so patients had a combination of tests as requested by their clinicians. The pulmonologists were asked to attribute the participant's main respiratory condition to one of six categories (COPD, asthma, ILD, other obstructive disease, normal, other). In participants with multiple respiratory diagnoses (*e.g.* COPD and ILD), the pulmonologist was asked to choose, at their discretion and best judgement, the predominant category. After initial independent scoring, the two pulmonologists met to discuss participants for whom there was no consensus. If consensus could not be agreed after discussion, these cases and their medical records were reviewed by a third pulmonologist to adjudicate independently, without access to index test results or to the previous scoring of the pulmonologists. All pulmonologists (S. Kon, W.D-C. Man, N.S. Hopkinson) were specialists in respiratory medicine with a minimum of 8 years as a consultant in the United Kingdom National Health Service, with expertise in the diagnosis and management of COPD and other chronic respiratory diseases. Pulmonologists had no access to the index test results (AI software reports), nor had they any communication with the software engineer or software company. Cases where the experts decided they had inadequate information to determine an appropriate diagnosis category were excluded from final analysis. The time taken to request and retrieve primary and secondary medical notes was not measured. On average, the experts took 15 min to review each case, though this was variable depending on the extensiveness of the prior patient workup.

Outcomes

The index test presented results in two ways: 1) an AI software preferred diagnosis (*i.e.* the category with the highest probability score) and 2) probabilities for all the six categories (totalling 100%). In addition, we evaluated AI software differential diagnosis (*i.e.* the top two categories with highest probability scores).

Model performance was evaluated by calculating the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) against all other categories than the one being assessed. For AUC, nonparametric analysis was conducted with 95% confidence intervals calculated as recommended by Hanley and McNeil [16]. For sensitivity, specificity, PPV and NPV, "exact" Clopper–Pearson 95% confidence intervals were reported.

Algorithmic audit

An algorithmic audit adapted from methods by Oakden-Rayner *et al.* [17] was conducted to explore misclassifications including confusion matrices and comparison of the probability cut-offs of cases classified correctly *versus* those classified incorrectly. Subgroup analyses were also conducted to audit the AI model performance in different demographic groups by sex, age, ethnicity, body mass index (BMI), smoking history and severity (for COPD) to explore possible biases or discrimination of model performance across these groups.

In a random subset of 200 cases, three expert respiratory physiologists from ARTP determined the technical quality of spirometry independently, which was then classified into high-quality *versus* suboptimal quality cases by majority consensus. High quality was defined as both forced expiratory volume in 1 s (FEV_1) and forced vital capacity (FVC) scored as grade A or B according to the American Thoracic Society (ATS)/European Respiratory Society (ERS) 2019 technical standard [18]. The 2019 standard was selected, as it is the currently accepted international standard and what the AI software was trained on. Model performance was then compared between high-quality *versus* suboptimal-quality cases.

All analyses were conducted using Stata BE version 18 and Microsoft Excel based on a pre-planned ethics approved analysis protocol designed with advice from the United Kingdom National Institute for Health and Care Excellence (supplementary material). Probability cut-offs were obtained using the Stata roctab command.

Results

1121 consecutive cases were obtained, with 1113 analysed in this validation study, as eight cases were determined by the expert pulmonologists to have insufficient medical record data for a reference diagnosis to be made. Experts independently achieved direct consensus on the reference diagnosis in 973 (87.4%) cases and achieved consensus in a further 125 (11.3%) cases following discussion; the remaining 15 (1.3%) cases required a third expert to determine the diagnosis (supplementary figure S2).

The most frequent diagnostic category was COPD (48.8%) followed by ILD (22.4%) (table 1). Almost all cases of the other obstructive disease category (86 out of 89) comprised people with bronchiectasis, whereas the unidentified category comprised primarily of people with heart failure and extrathoracic restriction (chest wall disorders, diaphragm disorders).

Identification of COPD

Out of 543 participants with a reference diagnosis of COPD, the AI preferred diagnosis had a sensitivity of 84.0% (95% CI 80.6–87.0%), specificity of 86.8% (95% CI 83.8–89.5%), PPV of 85.9% (95% CI 83.1–88.3%), negative predictive value (NPV) of 85.1% (95% CI 82.4–87.4%) and AUC of 0.914 (95% CI 0.896–0.930) (table 2, figure 1). When AI differential diagnosis was also taken into account, the software had a slight improvement in sensitivity to 90.6% (95% CI 87.8–92.9%), but with a larger reduction in specificity to 75.6% (95% CI 71.9–79.1%) (supplementary table S1).

Compared with AI software, using FEV_1/FVC <0.70 alone to identify COPD revealed increased sensitivity (90.6%, 95% CI 87.8–92.9%) and NPV (88.3%, 95% CI 85.3–90.8%), but worse specificity (67.5%, 95% CI 63.5–71.4%) and PPV (72.7%, 95% CI 70.2–75.0%).

Identification of non-COPD categories

The receiver operating characteristic curves for the AI software are illustrated in figure 1, with the AUC values reported in table 2. In non-COPD cases, AI-preferred diagnosis performance was best for ILD (AUC 0.900), which had the second highest number of cases (after COPD) evaluated, but less so for other categories. Specificity was high across all diagnoses with variable sensitivity and high negative predictive values, but low-to-moderate positive predictive values (range 0–60.5%). Probability cut-offs showed that for all conditions when classified as the highest category their probability was >30% out of a maximum 100% (table 2). AI differential diagnosis performance for all categories is shown in supplementary table S1.

TABLE 1 Demographic and clinical characteristics of patient cases analysed

	COPD	Asthma	ILD	Normal	OBD	Unidentified	Total
Patients	543	107	249	30	89	95	1113
Sex							
Male	55.8 (303)	38.3 (41)	58.6 (146)	53.3 (16)	39.3 (35)	50.5 (48)	52.9 (589)
Female	44.2 (240)	61.7 (66)	41.4 (103)	46.7 (14)	60.7 (54)	49.5 (47)	47.1 (524)
Age years	69.8±9.6	64.6±13.0	71.8±10.2	69.3±12.0	70.7±10.4	71.1±10.0	69.9±10.4
⩽50	4.2 (23)	12.1 (13)	5.6 (14)	6.7 (2)	4.5 (4)	3.2 (3)	5.3 (59)
51–60	12.7 (69)	25.2 (27)	8.8 (22)	20.0 (6)	10.1 (9)	11.6 (11)	12.9 (144)
61–70	30.6 (166)	25.2 (27)	21.3 (53)	26.7 (8)	28.1 (25)	28.4 (27)	27.5 (306)
71–80	40.5 (220)	28.0 (30)	47.0 (117)	30.0 (9)	48.3 (43)	40.0 (38)	41.1 (457)
>80	12.0 (65)	9.3 (10)	17.3 (43)	16.7 (5)	9.0 (8)	16.8 (16)	13.2 (147)
Ethnicity							
White	91.5 (497)	63.6 (68)	56.6 (141)	80.0 (24)	75.3 (67)	61.1 (58)	76.8 (855)
Other	8.5 (46)	36.4 (39)	43.4 (108)	20.0 (6)	24.7 (22)	38.9 (37)	23.2 (258)
Smoking							
Current	93.7 (509)	42.1 (45)	34.9 (87)	56.7 (17)	40.4 (36)	51.6 (49)	66.8 (743)
Nonsmoker	6.3 (34)	57.9 (62)	65.1 (162)	43.3 (13)	59.6 (53)	48.4 (46)	33.2 (370)
BMI kg·m^{-2}	28.1±7.1	31.4±7.6	27.9±6.3	31.6±5.3	28.1 7.9	30.7±8.7	28.7±7.3
<25	36.6 (199)	20.6 (22)	32.5 (81)	6.7 (2)	42.7 (38)	26.3 (25)	33.0 (367)
25–29	26.0 (141)	29.0 (31)	40.6 (101)	36.7 (11)	23.6 (21)	27.4 (26)	29.7 (331)
⩾30	37.4 (203)	50.5 (54)	26.9 (67)	56.7 (17)	33.7 (30)	46.3 (44)	37.3 (415)
Mean FEV_1 L	1.22±0.63	1.54±0.71	1.60±0.62	2.40±0.76	1.47±0.56	1.50±0.62	1.42±0.68
FEV_1 z-score	−3.00±1.22	−2.38±1.19	−1.86±1.28	−0.62±1.14	−2.13±1.12	−2.04±1.25	−2.47±1.35
Mean FVC L	2.46±0.90	2.38±0.86	2.03±0.80	3.20±0.99	2.27±0.75	2.04±0.82	2.32±0.89
FVC z-score	−1.66±1.37	−1.68±1.18	−2.13±1.38	−0.51±1.31	−1.53±1.20	−2.01±1.38	−1.76±1.37
Mean FEV_1 % predicted (GLI 2012)	48.12±20.58	60.47±19.91	67.86±21.20	89.90±21.00	62.90±19.40	64.22±21.82	57.40±22.99
Mean FVC % predicted (GLI 2012)	74.31±20.44	74.38±17.17	66.42±20.16	92.59±22.23	75.39±18.34	67.73±21.57	72.57±20.62
Mean FEV_1/FVC ratio	0.50±0.15	0.64±0.13	0.80±0.88	0.75±0.60	0.65±0.13	0.74±0.85	0.62±0.18
Consensus							
Direct consensus	95.6 (519)	83.2 (89)	91.6 (228)	53.3 (16)	77.5 (69)	54.7 (52)	87.4 (973)
Post-discussion	3.3 (18)	15.9 (17)	7.6 (19)	36.7 (11)	20.2 (18)	44.2 (42)	11.2 (125)
Third rater	1.1 (6)	0.93 (1)	0.8 (2)	10.0 (3)	2.3 (2)	1.1 (1)	1.4 (15)

Data are presented as n, % (n) or mean±SD. ILD: interstitial lung disease; OBD: other obstructive disease; BMI: body mass index; FEV_1: forced expiratory volume in 1 s; FVC: forced vital capacity; GLI: Global Lung Function Initiative.

TABLE 2 Area under the receiver operating characteristic curve (AUC) and model performance of artificial intelligence (AI) software by diagnosis

	COPD	Asthma	ILD	Normal	OBD	Unidentified[#]
Patients n	543	107	249	30	89	95
Probability cut-off	36.30%	33.90%	35.40%	33.50%	32.90%	37.50%
AUC (95% CI)	0.914 (0.896–0.930)	0.814 (0.790–0.836)	0.900 (0.880–0.916)	0.871 (0.850–0.891)	0.580 (0.551–0.610)	0.744 (0.717–0.769)
Accuracy (95% CI)[¶]	85.4 (83.2–87.5)	83.8 (81.5–85.9)	83.5 (81.2–85.6)	94.3 (92.8–95.6)	90.8 (88.9–92.4)	90.5 (88.6–92.1)
Sensitivity (95% CI)	84.0 (80.6–87.0)	55.1 (45.2–64.8)	75.1 (69.3–80.3)	33.3 (17.3–52.8)	0[+]	2.1 (0.3–7.4)
Specificity (95% CI)	86.8 (83.8–89.5)	86.9 (84.6–88.9)	85.9 (83.4–88.1)	96.0 (94.7–97.1)	98.6 (97.7–99.3)	98.7 (97.8–99.3)
PPV (95% CI)	85.9 (83.1–88.3)	30.9 (26.1–36.1)	60.5 (56.2–64.7)	18.9 (11.5–29.4)	0[+]	13.3 (3.4–40.2)
NPV (95% CI)	85.1 (82.4–87.4)	94.8 (93.7–95.8)	92.3 (90.6–93.7)	98.1 (97.6–98.5)	91.9 (91.9–92.0)	91.5 (91.3–91.8)

All cases excluding the category analysed were classified as noncases. ILD: interstitial lung disease; OBD: other obstructive disease; PPV: positive predictive value; NPV: negative predictive value. [#]: includes cases where the AI software reports the outcome as uncertain; [¶]: taking into account the prevalence of the condition in the whole cohort. Results for categories with lower case numbers would be driven by the ability to identify negative cases correctly; [+]: no positive cases detected by AI from the 89 classified by experts into the category.

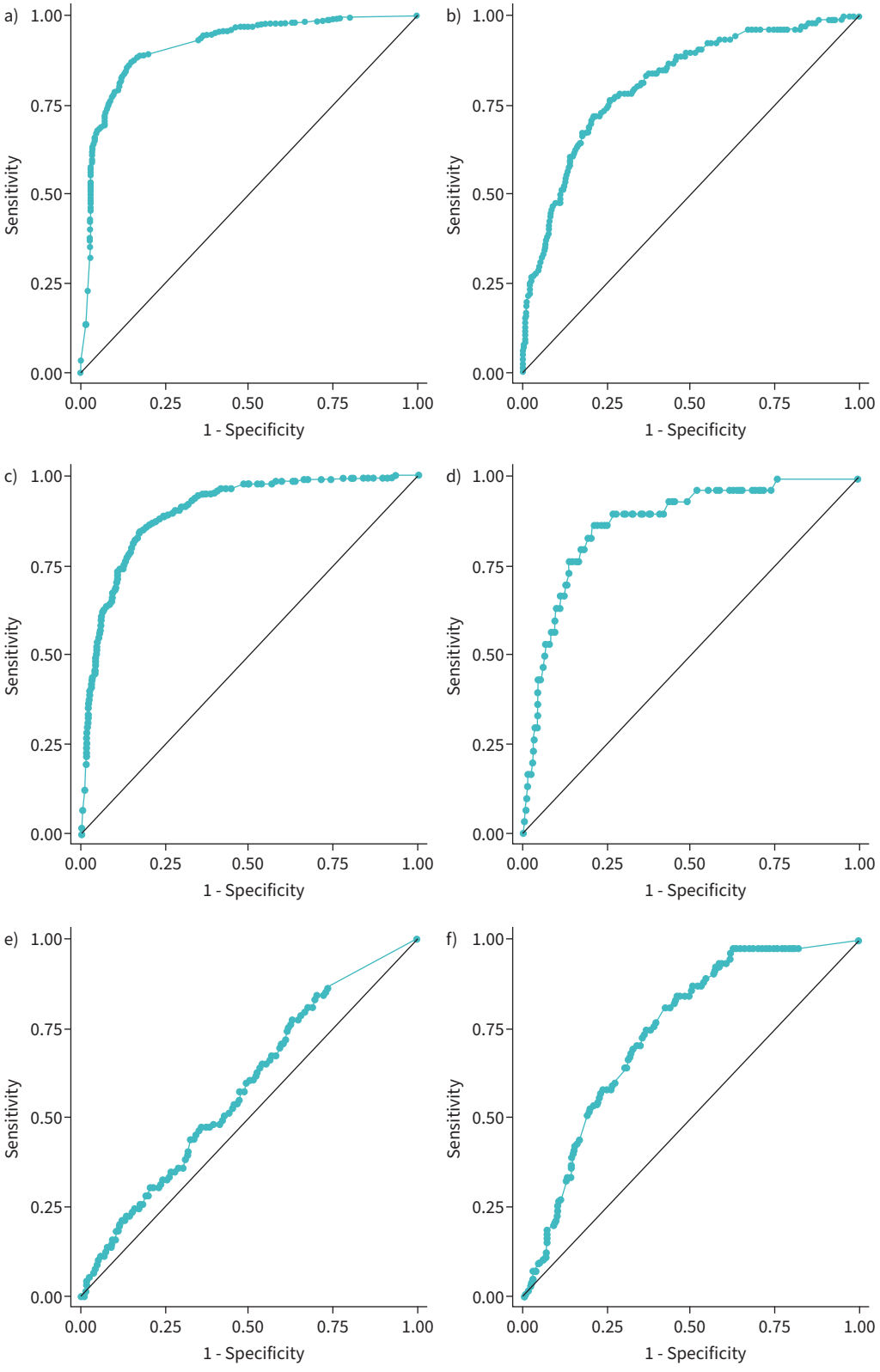

FIGURE 1 Receiver operating characteristic curves for a) COPD, b) asthma, c) interstitial lung disease, d) normal, e) other obstructive disease and f) unidentified cases.

Misclassification

Misclassifications for AI software preferred diagnoses are shown in the confusion matrix (table 3), with an overall Cohen's κ agreement coefficient of 0.477. For reference COPD cases, the most common misclassification by AI software was asthma (8.3%), whereas for reference asthma cases, COPD was the most common misclassification (16.8%). Very few reference COPD cases were misclassified as normal (1.5%), whereas across the 30 normal cases in the study, 66.7% were misclassified mainly to be ILD or asthma cases. Other obstructive disease as a category performed worst, with all cases in the category misclassified into the other groups, mainly asthma and COPD.

For the disease groups, when incorrectly classified, the median rank of the correct category was either second or third (out of six categories) (supplementary table S2). For example, on average, misclassified COPD cases had COPD ranked as the third (of six) category. A comparison of the mean probability of cases correctly *versus* incorrectly classified is shown in supplementary table S2.

Subgroup analyses

For COPD cases, subgroup analyses demonstrated that the AI software preferred diagnosis performed better in current/ex-smokers, in those with BMI <30 $kg \cdot m^{-2}$, and in cases where there was direct consensus of experts (table 4).

For all cases, the AI software preferred diagnosis classified a higher proportion of all cases correctly in smokers, those with BMI <30 $kg \cdot m^{-2}$, and in cases where the experts had direct consensus compared with those requiring discussion or the need for a third adjudicator (sensitivity 69.68% *versus* 24.00% *versus* 40.00%). Most cases requiring discussion were due to the presence of multiple respiratory diagnoses and identifying the predominant pathology. There was a statistically significant higher proportion of correct classifications in cases with White ethnicity as compared to other ethnicities (difference of 10.35%, p<0.002), although this was not found when analysis was limited to those with COPD. No significant differences were found between sex and age groups, though accuracy was found to be lower in those aged <50 years (COPD <50 years 65% *versus* ⩾50 years >80%). Subanalysis based on COPD severity according to the Global Initiative for Chronic Obstructive Lung Disease (GOLD) stages found that the AI software performed better in more severe cases (GOLD stage 3 and 4) compared to GOLD 2 and especially GOLD 1.

In the subanalysis by spirometry quality (n=200), the correct classification of overall cases by the AI software was significantly better (p=0.029) in those with optimal-quality FEV_1 compared with suboptimal FEV_1 (61.68% *versus* 46.24%), although not for reference COPD cases. Neither FVC quality nor the presence of both FEV_1 and FVC optimal quality were found to impact the accuracy in classification of cases.

Discussion

To our knowledge, this is the first study to assess the validity of AI in predicting the presence of lung disease using spirometry in primary care. In real-world data external to the training set comprising >1000 people with respiratory symptoms undergoing point-of-care hand-held spirometry in primary care, the AI software showed high sensitivity (84.0%) and specificity (86.8%) in identifying COPD from basic

TABLE 3 Confusion matrix with proportions of correctly and incorrectly artificial intelligence (AI)-classified cases by diagnosis

	Reference diagnosis					
	COPD	**Asthma**	**ILD**	**Normal**	**OBD**	**Unidentified**
Patients n	543	107	249	30	89	95
AI-preferred diagnosis						
COPD	**83.98**	16.82	4.82	6.67	30.34	16.84
Asthma	8.29	**55.14**	7.63	26.67	42.70	23.16
ILD	5.16	14.02	**75.10**	33.33	22.47	51.58
Normal	1.47	5.61	7.63	**33.33**	4.49	6.32
OBD	0.92	5.61	1.20	0	**0**	0
Unidentified[#]	0.18	2.8	3.61	0	0	**2.11**
Total	100	100	100	100	100	100

Data are presented as %, unless otherwise stated. Bold type represents percentages where both AI and reference diagnoses were concordant. ILD: interstitial lung disease; OBD: other obstructive disease. [#]: includes uncertain output.

TABLE 4 Subgroup analyses reporting proportions of correctly classified cases of COPD and correctly classified cases overall

	COPD			Overall		
	Cases n	Correctly classified %	p-value	Cases n	Correctly classified %	p-value
Sex						
Male	303	82.84	0.416	589	65.87	0.204
Female	240	85.42		524	62.21	
Age group years						
⩽50	23	65.22	0.138	59	49.15	0.162
51–60	69	82.61		144	62.50	
61–70	166	84.34		306	65.36	
71–80	220	85.00		457	65.43	
>80	65	87.69		147	65.31	
Ethnicity						
White	497	83.50	0.319	855	66.55	0.002
Other	46	89.13		258	56.20	
Smoking						
Current/ex-smoker	509	89.59	<0.001	743	72.54	<0.001
Nonsmoker	34	0		370	47.30	
BMI						
Normal	199	89.45	<0.001	367	69.21	<0.001
Overweight	141	87.94		331	68.88	
Obese	203	75.86		415	55.90	
COPD GOLD						
Stage 1	38	60.53	<0.001			
Stage 2	195	76.92				
Stage 3	189	91.01				
Stage 4	121	91.74				
Overall spirometry quality[#]	90			200		
Optimal (grade A/B)	19	84.21	0.799	50	54.00	0.935
Suboptimal	71	81.69		150	54.67	
FEV_1 spirometry quality						
Optimal (grade A/B)	63	82.54	0.904	107	61.68	0.029
Suboptimal	27	81.48		93	46.24	
FVC spirometry quality						
Optimal (grade A/B)	19	84.21	0.799	50	54.00	0.935
Suboptimal	71	81.69		150	54.67	
Adjudication[¶]						
Direct consensus	519	85.36	<0.001	973	69.68	<0.001
Consensus post-discussion	18	61.11		125	24.00	
Third adjudicator	6	33.33		15	40.00	

BMI: body mass index; GOLD: Global Initiative for Chronic Obstructive Lung Disease; FEV_1: forced expiratory volume in 1 s; FVC: forced vital capacity. [#]: spirometry quality was based on a subset of 200 random cases which were assessed for quality as per the American Thoracic Society/European Respiratory Society technical standards by three expert respiratory physiologists (>10 years' experience), with the most common grade selected determining the quality of the spirometry session. 50 cases were of optimal quality, and 150 of suboptimal quality; [¶]: most cases requiring discussion were due to the presence of multiple respiratory diagnoses (asthma+COPD, asthma+bronchiectasis, COPD+bronchiectasis, combined pulmonary fibrosis and emphysema) and identifying the predominant pathology.

demographic and spirometry data alone with an AUC of 0.914. This suggests that, with limited demographic data and spirometry (of variable quality), AI spirometry software can support identification and prediction of COPD in primary care, and potentially facilitate faster and more accurate diagnosis.

Between 50.0% and 98.3% [8, 9] of COPD cases remain undiagnosed globally. Delayed diagnosis represents a missed opportunity to initiate disease-modifying interventions [19], and is associated with increased exacerbations and healthcare costs [20]. The diagnosis of COPD requires confirmation of airflow obstruction by spirometry. Although noninvasive and cheap, there is significant inequity in the provision of spirometry in primary care [3, 21]. Furthermore, even when spirometry is available, previous studies have demonstrated low confidence in primary care practitioners, even those who have received spirometry training [6, 22–24]. In a web-based survey of 630 general practitioners in Norway, <50% correctly identified the spirometry parameters used for the diagnosis and grading of COPD [25].

Primary care physicians are open to the use of validated AI to support the accurate and differential diagnosis of chronic respiratory disease [3]. The AI software that served as the index test in the current study was previously developed for the automated reading of hospital PFTs [10] and was reported to assign a correct diagnosis in 41 (82%) out of 50 patients attending routine clinical tests in a subsequent evaluation, outperforming pulmonologists (including trainees) from 16 European countries (n=120) who made the correct diagnoses in only 44.6% of cases [10]. Furthermore, the combination of a pulmonologist and the AI software was better at interpreting PFTs than pulmonologist or AI alone [26]. A similar decision-support role is expected for primary care.

Compared to previous evaluations of hospital PFTs, there were several aspects of our study dataset that could have negatively impacted the AI software's performance. First, our cohort comprised consecutive, unselected, real-world patients undergoing hand-held spirometry in primary care settings supervised by nonphysiologists without comprehensive training in spirometry. The technical quality of spirometry was suboptimal and would be expected to be considerably poorer than spirometry conducted in hospital PFT laboratories supervised by dedicated respiratory physiologists. Second, whereas the original training set for the AI software comprised entirely of patients of White origin [10], 23.2% of the patients included in the dataset for this study were of other ethnicity. This may explain the significantly less accurate performance of the AI software in non-White cases by ∼10% and suggests benefits from possible retraining or tuning with data from a more diverse set of cases to the AI software training set. Finally, since our dataset originated from primary care, respiratory conditions for which primary-care spirometry plays a more crucial role in diagnosis were more frequently represented. For instance, COPD accounted for nearly 50% of the sample. This contrasts with the training cohort for the AI software, where rarer respiratory conditions seen in secondary and tertiary care, such as neuromuscular disease, chest wall disease, post-pleurectomy/lobectomy and cystic fibrosis were over-represented [10]. Despite these barriers, and the availability of only baseline spirometry, the AI software preferred diagnosis identified COPD cases correctly using minimal demographic data and raw spirometry data alone with a high sensitivity, specificity and AUC. This compares favourably with the correct identification of COPD by screening questionnaires in primary care (sensitivity 34.8–64.2%) [27, 28] and primary care physicians with access to clinical records and spirometry (sensitivity 41%) [29]. This is despite the variable technical quality of the spirometry, with our subsample analysis showing that only 25% had optimal quality FEV_1 and FVC measures according to international standards (table 4), in line with previous observations from multiple primary care settings [4].

Interestingly, the correct classification of COPD by AI was not significantly influenced by the quality of FVC nor by age or ethnicity (table 4). However, COPD severity did have a significant impact with better identification of cases across more severe cases (GOLD stages 3 and 4) compared to less severe ones by a margin of 31% and 15% for GOLD stages 1 and 2, respectively. Furthermore, quality of FEV_1 did affect overall correct classification of all cases with ∼15% difference in correct classifications between those from spirometry with optimal FEV_1 quality as compared with suboptimal quality. From a safety perspective, this suggests a possible need to limit the AI output to only spirometry that is at least of optimal FEV_1 quality to reduce the risk of misclassifications which may negatively affect clinical decision-making. This also emphasises the need even with AI support to undertake a systematic approach to improving the quality of spirometry conducted in primary care, including thorough provider training, quality control programmes, new models of care such as high-volume community diagnostic centres, and clinical decision-support tools. Further work is needed to understand whether AI algorithms that incorporate assessment of spirometry quality can further improve correct identification of cases and safety overall for implementation of such AI algorithms in practice.

Other than COPD, asthma is the major respiratory condition seen in primary care where spirometry is recommended as part of the diagnostic pathway. However, unlike COPD where the presence of airway obstruction identification by spirometry is mandatory for diagnosis, the role of spirometry is to primarily support a clinical diagnosis of asthma. The discriminative ability of the AI software to identify asthma from spirometry was reasonable with an AUC of 0.814 (driven by the small number of asthma cases across the whole cohort); however, sensitivity was only 55.1%, considerably lower than observed with COPD, and there was a low positive predictive value (30.9%) and correct classifications (55%). The dataset included a single point of access spirometry session for each patient, without bronchodilator responsiveness studies or fractional exhaled nitric oxide levels, and therefore data on variable airway obstruction, significant response to bronchodilator or eosinophilic airway inflammation were unavailable to AI. These lung function tests contribute to asthma diagnostic algorithms [30]. Furthermore, normal spirometry can be found in those with well-controlled or treated asthma, though this does not fully explain the findings in our study as only 5.61% of asthma cases were misclassified as normal by the AI software.

An unexpected finding was the ability of the AI software to identify ILD with a sensitivity of 75.1% and AUC of 0.900, although with a moderate PPV of 61% despite not being an essential diagnostic tool for the condition, unlike for COPD. Although restrictive spirometry is the hallmark physiological abnormality in patients with established fibrotic ILDs, FEV_1 and FVC can be well preserved in a significant proportion of patients with ILD at time of diagnosis [31] and accurate diagnosis typically requires computed tomography (CT) imaging. AI software can potentially support improvements in interpretation of spirometry findings through its ability for in-depth pattern recognition. For example, apart from the identification of typical restrictive spirometry, we speculate that the AI algorithm may have also identified early ILD in those with well-preserved FEV_1 and FVC through the identification of high FEV_1/FVC or peak flow/FVC ratios [31]; however, this requires further investigation. A caveat to this is that an incorrect AI-preferred diagnosis of ILD may lead to unnecessary investigations, patient anxiety or referral to secondary care if not interpreted with consideration of a patient's pre-test probability for the condition.

The AI software performed poorly in the classification of the other obstructive disease and other unidentified categories. The vast majority of those in the obstructive disease category had bronchiectasis, which often co-exists with other airways diseases, so it was unsurprising that 73% of the misclassifications were for asthma or COPD. The diagnosis of bronchiectasis is primarily through clinical history and high-resolution CT scanning.

In addition, the AI performed poorly in identifying the "other unidentified" group, which primarily comprised patients with heart failure (where spirometry plays a limited role in the diagnostic pathway), or extrathoracic restrictive conditions, such as chest wall disease or diaphragm disorders, which are difficult to distinguish from ILDs on the basis of spirometry alone. One reason maybe the lower direct consensus across experts for obstructive disease and unidentified cases as compared to other categories such as COPD, suggesting potentially greater ambiguity or complexity of these cases. This is in line with our subgroup analysis finding that the AI software performed better across cases where the experts had a direct consensus with each other. This is true for both classification of COPD and overall classification across the disease groups. More broadly, the results from this study emphasise the limited ability of spirometry alone to identify non-COPD conditions.

Decramer *et al.* [14] showed that spirometry, together with a detailed medical history, led to the correct identification of diagnosis in only 61% of patients presenting to a pulmonologist with respiratory symptoms. The main purpose of spirometry in primary care is to identify cases of COPD, or provide evidence to support a clinical diagnosis of asthma, and therefore further iterations of the AI software should consider simplifying the number of disease categories presented. There is also room to explore the impact of integrating other inputs, such as expanded history, physical exam results, further lung function tests (such as fractional exhaled nitric oxide, or pre–post bronchodilator studies), auscultation sounds or results from imaging [32, 33] to enhance the diagnostic prediction performance of AI spirometry software.

To our knowledge this is the first study to evaluate the performance of AI software using spirometry data to identify lung disease in primary care, the setting where most benefit can be derived from early and accurate diagnosis. The dataset was a consecutive convenience sample, with a sample size large enough to detect cases of COPD, ILD and asthma, at the observed frequency, to detect an AUC of 0.8 against a null hypothesis AUC value of 0.5 with 95% power. While we note that the patients were sourced from a single region in London, the region had a greater population diversity (especially with regards to ethnicity) and a different disease profile compared to the training cohort for the AI software. The spirometry quality themselves in the substudy was shown to be mostly suboptimal and reflective of real-world primary care practice.

To reduce bias, care was taken to keep the AI engineer blinded to the reference standard; similarly, the expert pulmonologist adjudicators were blinded to the index test results. The expert pulmonologists were provided with access to both primary and secondary care medical records including relevant investigations such as full PFTs, CT scans and echocardiograms performed after the index spirometry. This was supported by the high level of agreement around reference diagnosis when experts were scoring independently (87.4%) or following discussion (98.7%).

Although the study was conducted as a diagnostic accuracy study, it is important to point out that the primary care respiratory clinics provided multiple functions other than diagnostic spirometry and so a proportion of the analysed spirometry was not conducted for diagnostic purposes. There was a high prevalence of respiratory disease and very few people deemed "normal". As a result, it is not known whether the observed study results are generalisable to spirometry performed in other primary care settings,

such as diagnostic pathways, or pre-operative assessment screening as well as other countries, *e.g.* the United States or those from the Global South with different primary care models and spirometry conduct.

Practical implications

Future work should include whether the implementation of AI software into spirometry pathways influences the performance of primary care practitioners in the identification of COPD and other respiratory diseases, makes the diagnostic pathway for those presenting with respiratory symptoms more efficient, or affects secondary care referrals and other healthcare usage. There is a clear imperative for early diagnosis of COPD and asthma as this is associated with improved clinical outcomes and lower healthcare utilisation [34]. Our study suggests that while AI spirometry software can support this, for wider adoption, further work is required to adapt the software output, extend model representation across various demographics especially ethnicity through model retraining or tuning, and cost-effectiveness studies to ascertain the value of the software to health systems. Furthermore, this study suggests the need for the output of AI software to be contextualised to the primary care setting, for example by simplifying outputs to only report conditions such as COPD and asthma, which are commonly managed in primary care, and lung function patterns where there is a need to refer to secondary care (*e.g.* restrictive disease).

Conclusion

AI interpretation software achieved high sensitivity and specificity in identifying COPD from primary care spirometry (many suboptimal in quality) and basic demographic data. AI software performed less well for other chronic respiratory disease categories. Routine use of AI interpretation software could reduce barriers to conducting spirometry and support accurate early diagnosis of COPD, which is commonly misdiagnosed or underdiagnosed in practice.

Provenance: Submitted article, peer reviewed.

Ethics statement: The study was approved by the UK Health Research Authority (IRAS Project ID 314058) and pre-registered in Clinicaltrials.gov (NCT05648227).

Author contributions: W.D-C. Man and M. Topalovic conceived the wider research plan. All authors developed the theory and plan for this study. W.D-C. Man completed the statistical analysis plan and A. Sunjaya conducted the statistical analysis with support from A.T. Prevost. A. Sunjaya and W.D-C. Man drafted the initial manuscript. All authors reviewed, commented, and approved the manuscript. W.D-C. Man and A. Sunjaya are guarantors, and attest that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Conflict of interest: All authors have completed the ICMJE uniform disclosure form at https://www.icmje.org/disclosure-of-interest/ and declare the following. W.D-C. Man is Honorary President of the Association for Respiratory Technology and Physiology and is an associate editor of this journal. M. Topalovic is CEO of ArtiQ, a company that produces artificial intelligence-enabled lung function support software. M. de Vos, A. Elmahy, B. Cuyvers, P. Desbordes and K. Van Orshoven are employees of ArtiQ. There are no other relationships or activities that could appear to have influenced the submitted work.

Support statement: This study is funded by the National Health Service Transformation Directorate and the National Institute for Health and Care Research (NIHR) through an AI Award in Health and Care (Phase 3 – Application: grant number AI_ AWARD02204). A. Sunjaya is supported by the Australian Medical Research Future Fund Chronic Respiratory Diseases grant. S. Taylor is supported by the National Institute for Health Research ARC North Thames. R. Russell is supported by the NIHR Oxford Biomedical Research Centre – Respiratory. R.A. Evans is supported by an NIHR Clinical Scientist fellowship. The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The software company did not provide any funding for the conduct or analysis of the study. Funding information for this article has been deposited with the Open Funder Registry.

References

1 GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* 2020; 396: 1204–1222.

2 Stolz D, Mkorombindo T, Schumann DM, *et al.* Towards the elimination of chronic obstructive pulmonary disease: a Lancet Commission. *Lancet* 2022; 400: 921–972.

3 Doe G, Taylor SJ, Topalovic M, *et al.* Spirometry services in England post-pandemic and the potential role of AI support software: a qualitative study of challenges and opportunities. *Br J Gen Pract* 2023; 73: e915–e923.
4 van de Hei SJ, Flokstra-de Blok BMJ, Baretta HJ, *et al.* Quality of spirometry and related diagnosis in primary care with a focus on clinical use. *NPJ Prim Care Respir Med* 2020; 30: 22.
5 Hegewald MJ, Gallo HM, Wilson EL. Accuracy and quality of spirometry in primary care offices. *Ann Am Thorac Soc* 2016; 13: 2119–2124.
6 Dennis S, Reddel HK, Middleton S, *et al.* Barriers and outcomes of an evidence-based approach to diagnosis and management of chronic obstructive pulmonary disease (COPD) in Australia: a qualitative study. *Fam Pract* 2017; 34: 485–490.
7 White P, Wong W, Fleming T, *et al.* Primary care spirometry: test quality and the feasibility and usefulness of specialist reporting. *Br J Gen Pract* 2007; 57: 701–705.
8 Lamprecht B, Soriano JB, Studnicka M, *et al.* Determinants of underdiagnosis of COPD in national and international surveys. *Chest* 2015; 148: 971–985.
9 Axelsson M, Backman H, Nwaru BI, *et al.* Underdiagnosis and misclassification of COPD in Sweden – a Nordic Epilung study. *Respir Med* 2023; 217: 107347.
10 Topalovic M, Das N, Burgel PR, *et al.* Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests. *Eur Respir J* 2019; 53: 1801660.
11 Cohen JF, Korevaar DA, Altman DG, *et al.* STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* 2016; 6: e012799.
12 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.
13 Topalovic M, Laval S, Aerts JM, *et al.* Automated interpretation of pulmonary function tests in adults with respiratory complaints. *Respiration* 2017; 93: 170–178.
14 Decramer M, Janssens W, Derom E, *et al.* Contribution of four common pulmonary function tests to diagnosis of patients with respiratory symptoms: a prospective cohort study. *Lancet Respir Med* 2013; 1: 705–713.
15 Elmahy A, Maes J, Desbordes P, *et al.* AI models for respiratory diseases from spirometry alone: validation based on UK Biobank. *Eur Respir J* 2023; 62: Suppl. 67, PA3508.
16 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143: 29–36.
17 Oakden-Rayner L, Gale W, Bonham TA, *et al.* Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. *Lancet Digit Health* 2022; 4: e351–e358.
18 Graham BL, Steenbruggen I, Miller MR, *et al.* Standardization of spirometry 2019 update. An official American Thoracic Society and European Respiratory Society technical statement. *Am J Respir Crit Care Med* 2019; 200: e70–e88.
19 Jones RC, Price D, Ryan D, *et al.* Opportunities to diagnose chronic obstructive pulmonary disease in routine care in the UK: a retrospective study of a clinical cohort. *Lancet Respir Med* 2014; 2: 267–276.
20 Larsson K, Janson C, Ställberg B, *et al.* Impact of COPD diagnosis timing on clinical and economic outcomes: the ARCTIC observational cohort study. *Int J Chron Obstruct Pulmon Dis* 2019; 14: 995–1008.
21 Howard S. "Silent scandal" of missing lung diagnostics in England's most deprived areas – where respiratory disease is most prevalent. *BMJ* 2023; 382: 2140.
22 Walters JA, Hansen EC, Johns DP, *et al.* A mixed methods study to compare models of spirometry delivery in primary care for patients at risk of COPD. *Thorax* 2008; 63: 408–414.
23 Bunker J, Hermiz O, Zwar N, *et al.* Feasibility and efficacy of COPD case finding by practice nurses. *Aust Fam Physician* 2009; 38: 826–830.
24 Kaminsky DA, Marcy TW, Bachand M, *et al.* Knowledge and use of office spirometry for the detection of chronic obstructive pulmonary disease by primary care physicians. *Respir Care* 2005; 50: 1639–1648.
25 Tollånes MC, Sjaastad GE, Aarli BB, *et al.* Spirometry in chronic obstructive pulmonary disease in Norwegian general practice. *BMC Fam Pract* 2020; 21: 235.
26 Das N, Happaerts S, Gyselinck I, *et al.* Collaboration between explainable artificial intelligence and pulmonologists improves the accuracy of pulmonary function test interpretation. *Eur Respir J* 2023; 61: 2201720.
27 Martinez FJ, Han MK, Lopez C, *et al.* Discriminative accuracy of the CAPTURE tool for identifying chronic obstructive pulmonary disease in US primary care settings. *JAMA* 2023; 329: 490–501.
28 Siddharthan T, Pollard SL, Quaderi SA, *et al.* Discriminative accuracy of chronic obstructive pulmonary disease screening instruments in 3 low- and middle-income country settings. *JAMA* 2022; 327: 151–160.
29 Nardini S, Annesi-Maesano I, Simoni M, *et al.* Accuracy of diagnosis of COPD and factors associated with misdiagnosis in primary care setting. E-DIAL (Early DIAgnosis of obstructive lung disease) study group. *Respir Med* 2018; 143: 61–66.
30 Louis R, Satia I, Ojanguren I, *et al.* European Respiratory Society guidelines for the diagnosis of asthma in adults. *Eur Respir J* 2022; 60: 2101585.

31 Alyami SM, Moran-Mendoza O. Increased expiratory flows identify early interstitial lung disease. *Ann Thorac Med* 2023; 18: 152–155.

32 Pramono RXA, Bowyer S, Rodriguez-Villegas E. Automatic adventitious respiratory sound analysis: a systematic review. *PLoS One* 2017; 12: e0177926.

33 Aggarwal R, Sounderajah V, Martin G, *et al.* Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digit Med* 2021; 4: 65.

34 Aaron SD, Vandemheen KL, Whitmore GA, *et al.* Early diagnosis and treatment of COPD and asthma – a randomized, controlled trial. *N Engl J Med* 2024; 390: 2061–2073.