





Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests

Marko Topalovic¹, Nilakash Das¹, Pierre-Régis Burgel ², Marc Daenen³, Eric Derom⁴, Christel Haenebalcke⁵, Rob Janssen⁶, Huib A.M. Kerstjens ⁷, Giuseppe Liistro⁸, Renaud Louis⁹, Vincent Ninane¹⁰, Christophe Pison¹¹, Marc Schlessler¹², Piet Vercauter¹³, Claus F. Vogelmeier¹⁴, Emiel Wouters¹⁵, Jocke Wynants^{16,17} and Wim Janssens¹ on behalf of the Pulmonary Function Study Investigators

 @ERSpublications

There is poor accuracy and substantial disagreement between pulmonologists when interpreting complex pulmonary function data. Automating interpretation with artificial intelligence provides a powerful decision support tool in clinical practice. <http://ow.ly/Tj9h30nxw4U>

Cite this article as: Topalovic M, Das N, Burgel P-R, *et al.* Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests. *Eur Respir J* 2019; 53: 1801660 [<https://doi.org/10.1183/13993003.01660-2018>].

ABSTRACT The interpretation of pulmonary function tests (PFTs) to diagnose respiratory diseases is built on expert opinion that relies on the recognition of patterns and the clinical context for detection of specific diseases. In this study, we aimed to explore the accuracy and interrater variability of pulmonologists when interpreting PFTs compared with artificial intelligence (AI)-based software that was developed and validated in more than 1500 historical patient cases.

120 pulmonologists from 16 European hospitals evaluated 50 cases with PFT and clinical information, resulting in 6000 independent interpretations. The AI software examined the same data. American Thoracic Society/European Respiratory Society guidelines were used as the gold standard for PFT pattern interpretation. The gold standard for diagnosis was derived from clinical history, PFT and all additional tests.

The pattern recognition of PFTs by pulmonologists (senior 73%, junior 27%) matched the guidelines in 74.4±5.9% of the cases (range 56–88%). The interrater variability of $\kappa=0.67$ pointed to a common agreement. Pulmonologists made correct diagnoses in 44.6±8.7% of the cases (range 24–62%) with a large interrater variability ($\kappa=0.35$). The AI-based software perfectly matched the PFT pattern interpretations (100%) and assigned a correct diagnosis in 82% of all cases ($p<0.0001$ for both measures).

The interpretation of PFTs by pulmonologists leads to marked variations and errors. AI-based software provides more accurate interpretations and may serve as a powerful decision support tool to improve clinical practice.

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

Received: Aug 30 2018 | Accepted after revision: Jan 25 2019

This study is registered at ClinicalTrials.gov with identifier number NCT03264417.

Copyright ©ERS 2019

Introduction

Pulmonary function tests (PFTs) are our primary tool to evaluate the function of the respiratory system [1]. In practice, the interpretation is based on expert opinion, and involves the recognition of a pattern (obstructive, restrictive, mixed and normal) and the grading of its severity according to international guidelines [2–4]. To arrive at the final diagnosis the results of PFTs are combined with patient information, symptoms and, possibly, the results of other tests, such as imaging, blood analysis, biopsies and exercise tests [5, 6].

In 2005, an American Thoracic Society/European Respiratory Society (ATS/ERS) task force designed a simplified algorithm to assess lung function in clinical practice [2]. However, when these recommended guidelines were translated into software for diagnostic decision support, it led to only 38% of correct disease predictions. Adding patient characteristics into such an algorithm improved the accuracy to 68%, highlighting a vast potential for automated diagnostic labelling when combining PFTs with clinical information [7]. In fact, the Belgian Pulmonary Function Study (BPFS) demonstrated that expert panels could reach 77% accuracy when predicting the diagnosis based on PFTs and clinical history alone [8]. Although one may doubt if a computer algorithm carries any added value to a group of experts, the question of whether it may help individual readers is yet to be answered.

The number of successful applications of artificial intelligence (AI) is quickly rising. Supported by various outstanding achievements in the field and because of its unlimited potential to deal with big data, high expectations are also emerging for healthcare. For instance, one study demonstrated the ability of an AI algorithm to identify and classify skin cancer with similar expertise as 21 board-certified dermatologists [9]. Another study reached the same level of performance when analysing retinal fundus images for the identification of diabetic retinopathy [10]. Moreover, there are multiple examples from radiology in detecting traces of breast and lung cancer [11, 12]. Notwithstanding these technical superiorities of AI-based systems, translation into clinical practice with broad acceptance has been very challenging [13–15]. As PFTs are entirely standardised and used worldwide [16], they are ideally suited for the development of AI algorithms for test interpretation and diagnostics. PFTs provide an extensive series of numeric outputs, easily controllable by computers, yet the patterns are not always easily perceptible or appropriately recognised by the human eye. Moreover, the example of automated interpretation for ECGs, which is widely adopted and standardised in most equipment, highlights its potential use.

In this study, we hypothesised that AI can improve the clinical reading of PFTs and overcome the variable test interpretation of individual pulmonologists. We explored the accuracy and interrater variability of pulmonologists when interpreting patterns of PFTs, and when suggesting a specific category of respiratory disease diagnosis based on limited clinical information and PFTs. In addition, we compared the pulmonologists' performance with that of AI-based software developed and validated in more than 1500 historical cases.

Methods

Study design

120 pulmonologists from 16 hospitals in five European countries participated in this multicentre non-interventional study. They independently evaluated complete PFTs (pre- and/or post-bronchodilator spirometry, whole-body plethysmography for lung volumes and airway resistance, and diffusing capacity) and limited clinical information (smoking history, cough, sputum and dyspnoea) of 50 randomly selected

Affiliations: ¹Respiratory Medicine, University Hospital Leuven, Chronic Diseases, Metabolism and Ageing, KU Leuven, Leuven, Belgium. ²Cochin Hospital, AP-HP, Université Paris Descartes, Sorbonne Paris Cité, Paris, France. ³Dept of Respiratory Medicine, Hospital Oost-Limburg, Genk, Belgium. ⁴Dept of Respiratory Medicine, Ghent University Hospital, Ghent, Belgium. ⁵Dept of Respiratory Medicine, AZ Sint-Jan Hospital, Bruges, Belgium. ⁶Dept of Pulmonary Medicine, Canisius Wilhelmina Hospital, Nijmegen, The Netherlands. ⁷Dept of Pulmonary Medicine and Tuberculosis, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. ⁸Dept of Pneumology, Cliniques Universitaires St-Luc, Université Catholique de Louvain, Brussels, Belgium. ⁹Dept of Respiratory Medicine, University Hospital, Liege, Belgium. ¹⁰Dept of Respiratory Medicine, Saint-Pierre Hospital, Université Libre de Bruxelles, Brussels, Belgium. ¹¹Service Hospitalier Universitaire de Pneumologie et Physiologie, CHU Grenoble Alpes, Université Grenoble Alpes, Grenoble, France. ¹²Dept of Pulmonary Medicine, Centre Hospitalier de Luxembourg, Luxembourg, Luxembourg. ¹³Dept of Respiratory Medicine, Onze-Lieve-Vrouw Hospital, Aalst, Belgium. ¹⁴Dept of Medicine, Pulmonary and Critical Care Medicine, University Medical Center Giessen and Marburg, Member of the German Center for Lung Research (DZL), Marburg, Germany. ¹⁵Dept of Respiratory Medicine, Maastricht University Medical Center, Maastricht, The Netherlands. ¹⁶Dept of Pneumology, Jessa Hospital, Hasselt, Belgium. ¹⁷For a full list of Pulmonary Function Study Investigators, please refer to the Acknowledgements section.

Correspondence: Wim Janssens, Respiratory Medicine, University Hospital Leuven, Chronic Diseases, Metabolism and Ageing, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. E-mail: wim.janssens@uzleuven.be

patients, admitted to the University Hospital Leuven (Leuven, Belgium) for a respiratory problem. Evaluation sessions were performed in each hospital in the period from August 15, 2017 to December 13, 2017. All pulmonologists independently examined different patient cases according to a pre-established protocol by providing: 1) PFT pattern interpretation: obstructive, restrictive, mixed or normal pattern; 2) choice of one of nine preferred diagnostic categories: asthma, chronic obstructive pulmonary disease (COPD), other obstructive disease (OBD) (including bronchiectasis, bronchiolitis and cystic fibrosis), interstitial lung disease (including idiopathic pulmonary fibrosis, non-specific interstitial pneumonitis and sarcoidosis), pulmonary vascular disease (including pulmonary hypertension, embolism and vasculitis), neuromuscular disease (including paralysis of the diaphragm, poliomyelitis and myopathy), thoracic deformity (including pneumectomy, lobectomy, chest wall problems and kyphoscoliosis), healthy and other diseases; and 3) confidence in their decision on a Likert scale: from 1 point (“absolutely not sure”) to 5 points (“absolutely sure”) (an example is shown in supplementary figures S1 and S2). Finally, 4) the same patient files were examined by in-house developed AI-based software for PFT interpretation and diagnostic suggestion.

Study population

The study included a random sample of 50 subjects prospectively collected at the outpatient clinic of University Hospital Leuven in August 2017. All enrolled subjects were Caucasians aged >18 years who had performed a complete PFT and provided clinical information. The gold standard diagnosis was derived from clinical history, PFT and all necessary additional tests, and finally confirmed by an expert panel in Leuven. This *ad hoc* expert panel consisted of three experienced clinicians that reviewed all baseline and clinical follow-up data to agree on a final gold standard diagnosis out of the nine categories. Consensus was reached for all these cases. Baseline characteristics are shown in table 1, covering a wide range of respiratory diseases that may present with an abnormal PFT. Other conditions (such as lung cancer, cardiovascular disease, and ear, nose and throat problems) were excluded from the test sample (n=3). The Ethics Committee of the University Hospital Leuven approved the study protocol (approval S60619; August 4, 2017). The study design can be found at ClinicalTrials.gov (identifier NCT03264417). All included patients provided informed consent for the use of their data (approval S60243; June 23, 2017).

AI software

The development of software for automated reading of PFTs was performed in R language and its machine learning framework. The software used the same lung function data as input as presented to the

TABLE 1 Population characteristics of the 50 subjects whose lung function was evaluated in the study

	Asthma	COPD	OBD	NMD	TD	ILD	PVD	Healthy
Subjects	8	11	4	3	5	10	4	5
Male/ female	5/3	8/3	3/1	2/1	4/1	6/4	3/1	3/2
Age years	57 (27–70)	64 (38–77)	53 (34–77)	65 (48–72)	60 (52–68)	70 (51–83)	80 (62–81)	64 (38–74)
FEV₁ z-score	−0.57 (−2.70–0.73)	−1.41 (−3.95–0.41)	−2.97 (−4.05–−1.39)	−2.47 (−2.87–−1.97)	−2.76 (−2.94–−1.77)	−0.65 (−2.74–1.01)	0.17 (−2.23–0.78)	0.24 (−0.01–1.63)
FVC z-score	−0.41 (−1.86–2.00)	0.70 (−2.48–2.07)	−2.51 (−4.37–−0.48)	−2.58 (−2.85–−1.93)	−2.68 (−2.79–−2.30)	−0.97 (−3.42–0.79)	0.66 (−1.83–1.59)	0.11 (−0.06–1.22)
FEV₁/FVC z-score	−1.01 (−2.79–0.29)	−2.54 (−4.86–−1.54)	−2.51 (−4.37–−0.48)	−0.41 (−0.60–0.07)	−0.83 (−1.41–1.47)	0.85 (−0.25–2.05)	−0.90 (−1.10–−0.53)	0.32 (−0.26–0.50)
TLC z-score	0.01 (−1.04–2.39)	1.55 (−1.49–2.80)	0.17 (−0.74–1.20)	−2.23 (−3.01–−2.17)	−2.98 (−5.05–−1.10)	−2.54 (−4.96–−1.00)	−0.29 (−1.53–0.11)	−0.13 (−0.43–1.50)
RV z-score	−0.02 (−2.81–4.49)	1.10 (−1.59–6.24)	2.24 (1.38–3.22)	−0.95 (−1.08–−0.19)	−1.50 (−2.98–1.67)	−2.45 (−4.20–−1.35)	−0.79 (−2.34–0.16)	−0.99 (−2.42–3.14)
D_{lco} z-score	−0.84 (−1.96–1.25)	−2.77 (−4.39–−0.54)	−1.89 (−3.98–−0.67)	−2.08 (−2.30–−1.74)	−2.44 (−4.77–−1.98)	−2.91 (−4.30–−0.06)	−2.80 (−4.17–−2.33)	−0.67 (−2.37–−0.29)
K_{co} z-score	0.09 (−0.93–1.48)	−2.05 (−2.93–−0.27)	−0.17 (−1.94–1.95)	0.53 (0.48–0.55)	0.18 (−1.86–1.73)	−1.09 (−2.04–1.27)	−2.02 (−3.53–−1.17)	−0.32 (−1.34–−0.07)

Data are presented as n or median [range]. COPD: chronic obstructive pulmonary disease; OBD: other obstructive disease; NMD: neuromuscular disease; TD: thoracic deformity; ILD: interstitial lung disease; PVD: pulmonary vascular disease; FEV₁: forced expiratory volume in 1 s; FVC: forced vital capacity; TLC: total lung capacity; RV: residual volume; D_{lco}: diffusing capacity of the lung for carbon monoxide; K_{co}: transfer coefficient of the lung for carbon monoxide.

pulmonologists (absolute values, percent predicted of normal reference values and z-scores; also shown in supplementary figure S1) combined with patient characteristics, age, pack-years, sex and body mass index. For pattern interpretations, the PFT algorithm was in line with ATS/ERS strategies [2]. However, the engine for complex diagnostic categorisation had to be developed and a machine learning approach was adopted.

The machine learning model was built using data from 1430 subjects used in our previous work to ensure a broad variety of data [7, 8, 17]. This data came from two cohorts: 1) BPFs, a prospective cohort study that enrolled a clinical population-based sample (n=851) of all successive undiagnosed patients admitted for the first time to one of the 33 participating Belgian hospitals due to respiratory symptoms [8], and 2) a retrospectively collected PFT data cohort of patients followed at the outpatient clinic of the University Hospital Leuven based on predefined established diagnoses (neuromuscular disease (n=112), chest/pleural wall problems, including pneumectomy and lobectomy (n=64), pulmonary vascular disease (n=76), OBD (n=100), COPD (n=47), asthma (n=40), healthy (n=50) and interstitial lung disease (n=90)). Briefly, all subjects were Caucasians aged between 18 and 85 years who had performed a complete PFT (including post-bronchodilator spirometry, whole-body plethysmography for lung volumes and airway resistance, and diffusing capacity). The final diagnosis was established with all additional tests deemed necessary by the responsible clinician, the patients' history and PFTs. Subsequently, it was validated by an *ad hoc* installed expert panel (BPFs) or by the clinical expert panel taking care of the patients in follow-up (Leuven data). The expert discussions of the BPFs were organised during the local meetings of physicians, at which all individual cases were presented to obtain a final diagnosis by consensus. In case there was disagreement, voting was used for a final gold standard diagnosis and, if needed, a secondary diagnosis [8]. For the retrospective PFT data collection of patients followed at the University Hospital Leuven, corresponding medical records were verified on the final diagnosis. For the few cases in which there was doubt about the diagnosis, the PFT data were not extracted and these cases were rejected. 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% [17]. Probabilistic output for each of the diagnostic categories obtained by the machine learning model was summarised in a report (supplementary figure S3).

Pulmonary function tests

All PFTs were performed with standardised equipment by respiratory technicians (MasterLab; Jäger, Würzburg, Germany), according to ATS/ERS criteria [18]. Spirometry data, as well as plethysmography and single-breath diffusing capacity data, were given as absolute values, but also expressed as percent predicted of normal reference values and as z-scores [19–21]. In the current prospective study, these data were presented to the AI software and pulmonologists, the latter also having access to the corresponding flow–volume loops, plethysmography and diffusing capacity manoeuvres.

Statistical analysis

Statistical analysis was performed using R version 3.3.3 (Foundation for Statistical Computing, Vienna, Austria). Figures were produced using Prism version 6 (GraphPad, La Jolla, CA, USA). The interobserver agreements were assessed using Fleiss' κ for multiple raters on categorical data. Interpretative strategies for lung function tests from the ATS/ERS task force were used as the gold standard to define a correct lung function pattern [2]. Preferred diagnostic category, by pulmonologists or software, was considered as correct if it corresponded to the gold standard diagnosis made historically by the expert panel based on all data. For both measures, *i.e.* PFT pattern interpretation and diagnostic category suggestion, accuracy was defined as the percentage of correctly labelled cases. The t-test and Mann–Whitney U-test were used to evaluate differences between groups with normal and non-parametric distribution, respectively. The Kruskal–Wallis test was used to determine the statistical difference between multiple groups. The one-sample t-test was used to assess the difference of AI performance and the average accuracy of pulmonologists. Results are presented as mean with standard deviation or as median with range.

Results

There were 120 pulmonologists who all together made 6000 evaluations of PFTs with clinical information. The pulmonologist group consisted of more senior members (n=88, established pulmonologists) than junior members (n=32, pulmonologists in training). A minimum number of five pulmonologists per centre was needed to participate.

PFT pattern interpretations

Applying the ATS/ERS interpretative strategies for PFTs revealed that the population consisted of 18 patients with an obstructive pattern, 10 patients with a restrictive pattern and 22 patients with a normal

lung function pattern, while there were no subjects with a mixed pattern. The interpretations of 118 pulmonologists (data were missing from two) matched with the reference PFT pattern in $74.4 \pm 5.9\%$ of the 50 cases, ranging from 56% to 88% per individual. The identification of a restrictive pattern was more difficult (positive predictive value 59% and sensitivity 75%) compared with normal and obstructive patterns (table 2). Even though a mixed pattern was not present, 376 (6%) cases were interpreted as mixed. A $\kappa=0.67$ signified a considerable interrater variability or disagreement between different pulmonologists. When the accuracy between different centres was compared, no significant differences in correct detections were found ($p=0.06$) (figure 1a). There were no significant differences between university and non-university centres ($p=0.06$) or between senior and junior readers ($p=0.49$). Interestingly, out of the 285 misclassified normal patterns falsely labelled into an obstructive pattern, 216 (76%) were on the four cases having a forced expiratory volume in 1 s (FEV₁)/forced vital capacity (FVC) ratio the above lower limits of normal but still below the 0.7 fixed cut-off.

Preferred diagnostic categories

For an individual pulmonologist, it was rather difficult to assign a correct preferred diagnostic category based on complete PFT data and clinical information. The mean accuracy of 6000 evaluations was only $44.6 \pm 8.7\%$, and it ranged from 39% to 51% per centre and from 24% to 62% per individual pulmonologist (figure 1b). A low κ score of 0.35 was indicative of a common disagreement between pulmonologists. Interestingly, age or clinical experience of the examiners did not influence the mean accuracy (seniors $45 \pm 4.2\%$ versus juniors $43.6 \pm 4.8\%$; $p=0.46$). Likewise, results were neither different between hospitals ($p=0.44$) nor affected by hospital type (university $44.1 \pm 9.4\%$ versus non-university $45.2 \pm 7.8\%$; $p=0.47$) or by country ($p=0.26$).

Due to a higher sensitivity, patterns of healthy subjects (true positive rate 71%) and subjects with COPD (true positive rate 65%) were more often identified on lung function than any of the other categories. Patient cases of less prevalent conditions, without a straightforward pattern (“fingerprint”) on lung function, were more difficult for the pulmonologists (thoracic deformity and neuromuscular disease, true positive rate 25%; asthma, true positive rate 20%). A detailed statistical group comparison is shown in table 3 and supplementary figure S4.

Confidence in decision making

Rarely, pulmonologists were “absolutely not sure” (in 2.7% of cases) or “not sure” (11.5%) when suggesting the preferred diagnostic category. Most commonly they were “sure” (36.5%) and “absolutely sure” (16%) in their decisions. Higher confidence in diagnostic suggestion was observed in decisions that were correct ($p<0.0001$) compared with the incorrect decisions. However, high confidence did not necessarily lead to correct diagnosis. From all “sure” and “absolutely sure” records, only 51.8% of the diagnoses were correct (supplementary figures S5 and S6).

TABLE 2 Confusion matrix with counts of all correctly and incorrectly labelled subjects per pulmonary function test (PFT) pattern

	Pulmonologist pattern				Total	Subjects
	Obstructive	Restrictive	Normal	Mixed		
Reference pattern						
Obstructive	1636	196	180	112	2124	18
Restrictive	34	883	14	249	1180	10
Normal	285	424	1872	15	2596	22
Mixed	0	0	0	0	0	0
Total	1955	1503	2066	376	5900	
Subjects (averaged)#	17	13	17	3		50
Specificity %	92	87	94			
Sensitivity %	77	75	72			
PPV %	84	59	91			
NPV %	88	93	81			

Data are presented as n, unless otherwise stated. PPV: positive predictive value; NPV: negative predictive value. Boxed rows show true reference PFT patterns, while boxed columns show patterns labelled by pulmonologists. There are 4391 (74.4%) correctly given interpretations [true positive in bold]. #: averaged number of subjects for each pattern given by each pulmonologist.

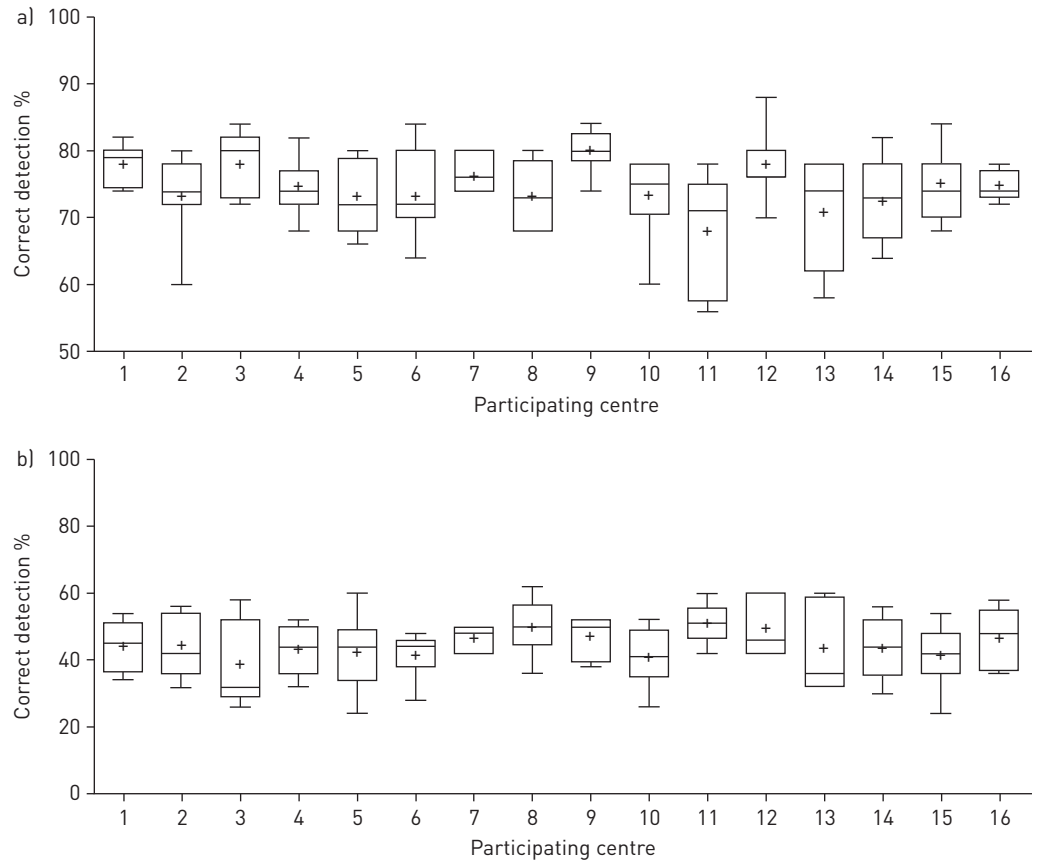


FIGURE 1 Comparison of correct detections per each participating centre. a) Pulmonary function test pattern interpretation. No significant difference between centres ($p=0.06$). b) Preferred diagnostic category. No significant difference between centres ($p=0.44$). Centres are anonymised. Box-and-whisker plots show median with interquartile range (box) and range (whiskers); the mean is indicated by “+”.

Comparison with the AI software

The in-house developed AI-based software perfectly matched the pattern interpretations of the ATS/ERS guidelines (100%). Software response was 0.2 s, giving immediate and consistent interpretations. Moreover, it assigned a correct diagnostic category in 82% of the cases, which was greatly superior to the average 44.6% accuracy of the pulmonologists ($p<0.0001$) (figure 2). It also proved to be highly sensitive in recognising COPD, neuromuscular disease, interstitial lung disease and healthy subjects. Concerning positive predictive value, the software showed powerful results for the majority of the respiratory disease diagnoses (figure 3 and table 4). Both the sensitivity and positive predictive value of the AI-based algorithm were superior to expert-based diagnostic category allocation in each of the eight disease groups (figure 3). AI lacked sensitivity for the OBD group, which was recouped by the very high positive predictive value.

Discussion

In this study, we explored the accuracy and consistency between pulmonologists when interpreting PFT patterns and providing a preferred diagnostic category. PFT pattern interpretations matched the ATS/ERS guidelines in 74.4% of cases with an interrater variability of $\kappa=0.67$, demonstrating that such a fundamental task is prone to mistakes and disagreements. PFTs combined with limited clinical information were difficult for pulmonologists as the only tool for reaching an accurate diagnostic category (accuracy of 44.6% and significant variability of $\kappa=0.35$). However, our advanced AI-based software for the automated clinical reading of PFTs perfectly interpreted (100%) PFT patterns and pointed to the correct diagnostic category in 82% of all cases. Consequently, it outperformed the pulmonologists in both tasks by 34% and 84%, respectively, which demonstrates that individual pulmonologists do not sufficiently capture the information available in PFTs.

Facilitating clinical practice with decision support systems is not a new idea and it has been shown that the majority (64%) of such systems do improve the performance of individual clinicians [22]. Nowadays, we

TABLE 3 Confusion matrix with counts of all correctly and incorrectly labelled subjects by the pulmonologists per each diagnostic category

	Pulmonologist diagnosis								Total	Subjects	
	Asthma	COPD	OBD	NMD	TD	ILD	PVD	Healthy			Other
Reference diagnosis											
Asthma	189	82	141	23	49	4	5	395	72	960	8
COPD	157	859	154	4	6	28	49	22	41	1320	11
OBD	77	139	162	13	15	5	6	45	18	480	4
NMD	1	2	7	90	156	70	3	4	27	360	3
TD	10	103	56	68	152	133	7	15	56	600	5
ILD	2	9	5	58	168	533	167	205	53	1200	10
PVD	2	55	27	8	18	75	266	11	18	480	4
Healthy	21	24	10	6	9	7	49	426	48	600	5
Other	0	0	0	0	0	0	0	0	0	0	0
Total	459	1273	562	270	573	855	552	1123	333	6000	
Subjects [averaged]#	3.8	10.6	4.7	2.3	4.8	7.1	4.6	9.4	2.8		50
Specificity %	90	81	86	93	86	87	89	76			
Sensitivity %	20	65	34	25	25	44	55	71			
PPV %	41	67	29	33	27	62	48	38			
NPV %	76	80	89	91	85	76	92	93			

Data are presented as n, unless otherwise stated. COPD: chronic obstructive pulmonary disease; OBD: other obstructive disease; NMD: neuromuscular disease; TD: thoracic deformity; ILD: interstitial lung disease; PVD: pulmonary vascular disease; PPV: positive predictive value; NPV: negative predictive value. Boxed rows show true reference diagnostic category, while boxed columns show diagnosis labelled by the pulmonologists. There are 2667 (44.6%) correctly suggested diagnoses (true positive in bold). #: averaged number of subjects for each diagnostic category given by each pulmonologist.

regularly use them to interpret ECGs, to analyse mammogram irregularities or as reminders for drug prescription [23, 24]. Although automated analyses of PFTs have been evaluated previously [25, 26], none has become a clinical reality. First, there is an obvious difficulty in reaching a preferred diagnosis without knowing the clinical context [27, 28]. Second, there is a lack of clear international diagnostic guidelines to label respiratory diseases based on PFTs, with controversial and often arbitrary choices of cut-offs to label abnormality. This implies that not all pulmonologists are using the same interpretative strategies in their daily routine [29, 30]. For example, a typical conflict is often seen in the first interpretative step: should we take the lower limits of normal or fixed 0.7 cut-off for the FEV₁/FVC ratio [31]. Undoubtedly, this will explain some of the differences between the interpretations of pulmonologists, but it also highlights a more general concern. Different recommendations on which cut-offs to use will reclassify individual patients from healthy to diseased and *vice versa*, while in real life the disease processes will present as a continuum around pre-fixed values. The strength of complete PFTs lies in the variety and multitude of tests in order to recognise disease-specific patterns, regardless of these fixed cut-off points.

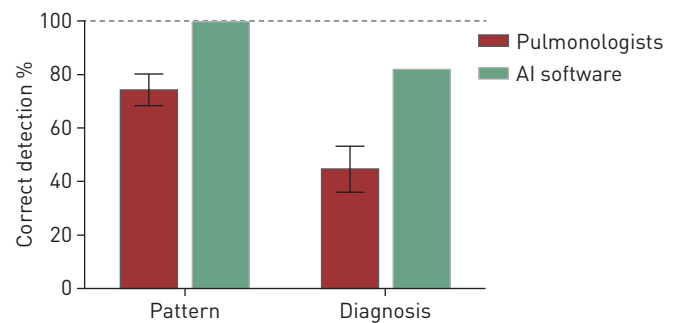


FIGURE 2 Comparison of results obtained by pulmonologists *versus* results achieved by the artificial intelligence (AI) software. Correct detections are significantly ($p < 0.0001$) higher for the AI software (improvement of 34% for pulmonary function test pattern interpretation and 84% for diagnostic category detection). Error bars indicate standard deviation.

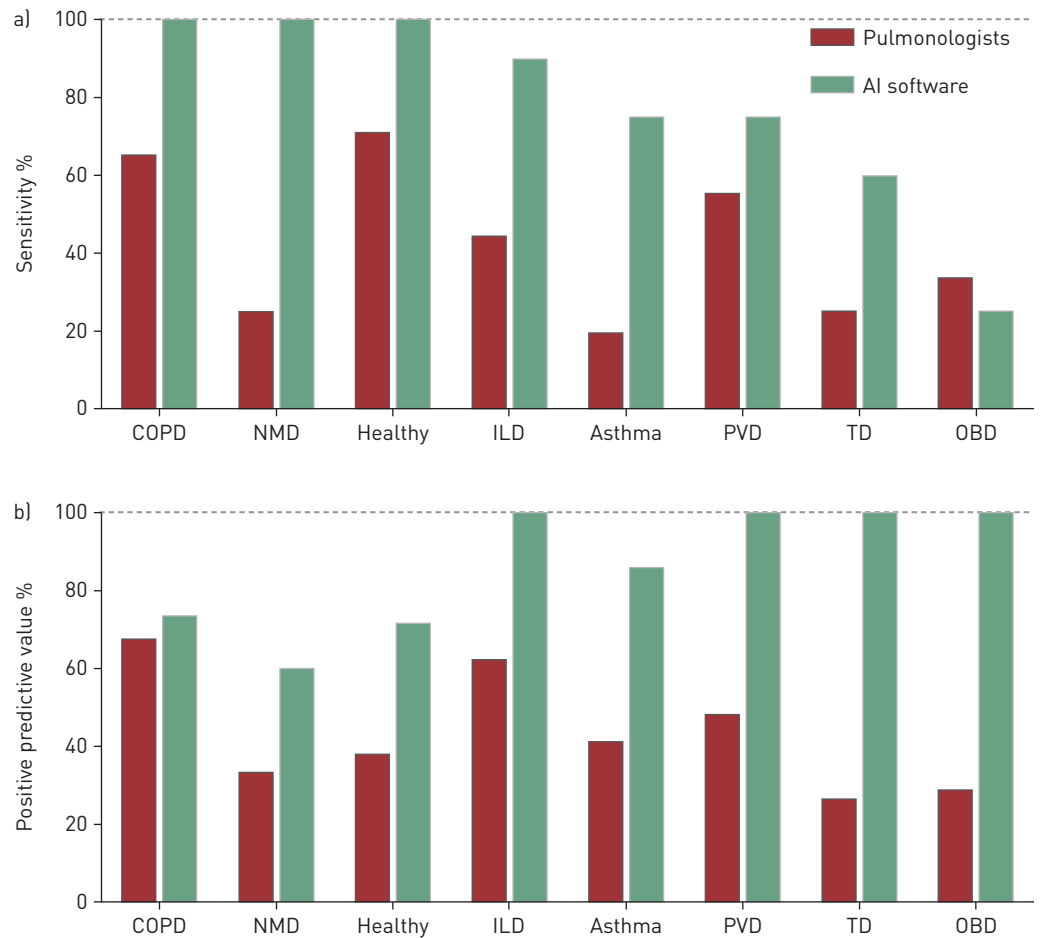


FIGURE 3 Performance of pulmonologists in comparison with the artificial intelligence software for each disease category. COPD: chronic obstructive pulmonary disease; NMD: neuromuscular disease; ILD: interstitial lung disease; PVD: pulmonary vascular disease; TD: thoracic deformity; OBD: other obstructive disease. a) Sensitivity (*i.e.* true positive/(true positive+false negative)) shows how many relevant subjects (from a specific group) were correctly identified. b) Positive predictive value (*i.e.* true positive/(true positive+false positive)) shows how many labelled subjects rightly belonged to the specific group.

Using AI, we approached each disease as having a unique fingerprint on the PFT. As such, AI identifies subtle and defining characteristics that are challenging for humans to detect, and incorporates them into a powerful discriminating diagnostic algorithm. In our case, the AI system takes complete input data and maps them into a high-dimensional space. As a result of a large number of known disease cases, with known magnitudes and patterns between all input data, AI will construct the most optimal hyperplanes that categorise new examples. Once presented with the data of a new patient, AI maps them into the same high-dimensional space and predicts to which category they belong. Such a multidimensional approach exceeds human capabilities to observe the same data in terms of accuracy. Fundamentally, the AI algorithm is no longer dependent on the arbitrary cut-offs, but is a purely patient data-driven knowledge system. In fact, with the increase in computing resources, modern AI algorithms have entirely moved away from rule-based systems and currently adopt a probabilistic approach. Our study confirms that a unique data-driven fingerprint of each disease often exists in the PFTs.

A fascinating characteristic of AI-based software is its ability to improve over time by being exposed to new and more difficult cases. In other words, the developed software may improve (as do physicians) by learning from mistakes and gaining experience. It is too ambitious to expect the software to be correct in 100% of cases, as some respiratory diseases do not show characteristic lung function abnormalities. Particularly for early disease stages or combined complex disease processes, disease-specific characteristics may be hidden. As the current accuracy of the AI software is situated within the range that clinical expert panels reached during the BPFS [8], there is probably little room for improvement. However, it also indicates that a computer can process all necessary information as effectively as a group of experts (not the individual), yet at a much higher speed and with 100% consistency for the same data input. The further usefulness of the AI software will be demonstrated if it decreases the time to final diagnosis, reduces the

TABLE 4 Confusion matrix with counts of all correctly and incorrectly labelled subjects by the artificial intelligence (AI) software per each diagnostic category

	AI software diagnosis									Subjects
	Asthma	COPD	OBD	NMD	TD	ILD	PVD	Healthy	Other	
Reference diagnosis										
Asthma	6	1	0	0	0	0	0	1	0	8
COPD	0	11	0	0	0	0	0	0	0	11
OBD	1	2	1	0	0	0	0	0	0	4
NMD	0	0	0	3	0	0	0	0	0	3
TD	0	0	0	2	3	0	0	0	0	5
ILD	0	0	0	0	0	9	0	1	0	10
PVD	0	1	0	0	0	0	3	0	0	4
Healthy	0	0	0	0	0	0	0	5	0	5
Other	0	0	0	0	0	0	0	0	0	0
Total	7	15	1	5	3	9	3	7	0	50
Specificity %	97	88	100	95	100	100	100	95		
Sensitivity %	75	100	25	100	60	90	75	100		
PPV %	86	73	100	60	100	100	100	71		
NPV %	95	100	93	100	95	97	97	100		

Data are presented as n, unless otherwise stated. COPD: chronic obstructive pulmonary disease; OBD: other obstructive disease; NMD: neuromuscular disease; TD: thoracic deformity; ILD: interstitial lung disease; PVD: pulmonary vascular disease; PPV: positive predictive value; NPV: negative predictive value. Boxed rows show true reference diagnostic category, while boxed columns show diagnosis labelled by the AI software. There are 41 (82%) correctly suggested diagnoses (true positive in bold).

number of tests needed for a final diagnosis and, if by standardising PFT interpretation, a number of misdiagnoses can be avoided.

Comparable with the human examiner marking their confidence on a Likert scale, AI expresses its certainty as a probability of a patient belonging to one of the disease categories. In the situations where AI made a wrong diagnostic suggestion, it should be noted that it never attributed a high probability to this diagnosis. More specifically, probability barely exceeded 50% in two out of the nine mislabels and it was <50% in the seven other cases. Surprisingly, the use of the COPD Assessment Test for the quantification of symptoms in the BPFs study did not contribute to further improving the accuracy of our AI software. This suggests that most respiratory diseases present with similar non-specific symptoms such as cough and dyspnoea. It is tempting to speculate that more input, e.g. more extensive history taking, and tests like exhaled nitric oxide fraction, forced oscillometry and/or blood/radiological markers, could enhance its future potential. In particular, for diseases such as asthma that can present with a normal PFT, the added value of such tests when integrated into our AI-based software is obvious.

A limitation of the current study is that we underestimated the accuracy of the pulmonologists by limiting the amount of clinical data to suggest a preferred diagnosis. In reality, a diagnosis is reached by a synergy of multiple factors, including expanded history, clinical examination, imaging and blood sampling. The real-life situation may therefore yield better outcomes. Additionally, the test sample we used may not entirely reflect the prevalence of diseases that pulmonologists confront in daily clinical practice. It is clear that we only explored the maximum output that could be reached from PFTs and clinical information, representative of the first diagnostic encounter. Furthermore, we did not formally test the level of agreement within the *ad hoc* expert panel to define the final diagnosis. Although the experts relied on all available test information, one may speculate that providing the AI interpretation would have favoured their initial agreement. A final limitation is that the risk of misinterpretation and misdiagnosis increases if tests are poorly performed [32]. However, sufficient quality of the tests is needed for both human and computer interpretations.

To conclude, our data indicate that interpretation of PFTs and the suggestion of primary respiratory disease diagnosis by pulmonologists is highly variable. The AI-based software has superior performance and may provide a powerful decision support tool for clinicians. The significance of such technology in improving clinical practice will drive real-life acceptance by the medical community.

Acknowledgements: We thank all the pulmonologists, pulmonary function technicians, patients and hospitals who participated in the study for providing and analysing data.

Author contributions: All authors critically revised the manuscript and approved the final version. All authors organised evaluation sessions in hospitals, examined patient files and interpreted results. M. Topalovic performed the data acquisition, analysis, interpretation as well as contributed to the study design and wrote the manuscript. N. Das contributed to data acquisition. W. Janssens takes responsibility for the content of the manuscript, contributed to the study design, and assisted in the data analysis, interpretation and writing of the manuscript.

The Pulmonary Function Study Investigators: R. De Pauw, C. Depuydt, C. Haenebalcke, S. Muyldermans, V. Ringoet, D. Stevens (AZ Sint-Jan Hospital, Bruges, Belgium); S. Bayat, J. Benet, E. Catho, J. Claustre, A. Fedi, M.A. Ferjani, R. Guzun, M. Isnard, S. Nicolas, T. Pierret, C. Pison, S. Rouches, B. Wuyam (CHU Grenoble Alpes, Grenoble, France); J.L. Corhay, J. Guiot, K. Ghysen, L. Renaud, A. Sibille (University Hospital, Liege, Belgium); H. De La Barriere, C. Charpentier, S. Corhut, K.A. Hamdan, M. Schlessler, G. Wirtz (Centre Hospitalier de Luxembourg, Luxembourg, Luxembourg); E. Alabadian, G. Birsén, P.R. Burgel, A. Chohra, C. Hamard, B. Lemarié, M.N. Lothe, C. Martin, A.C. Sainte-Marie, L. Sebane (Cochin Hospital, Paris, France); Y. Berk, B. de Brouwer, R. Janssen, J. Kerckhoff, A. Spaanderman, M. Stegers, A. Termeer, I. van Grimbergen, A. van Veen, L. van Ruitenbeek, L. Vermeer, R. Zaai, M. Zijlker (Canisius Wilhelmina Hospital, Nijmegen, The Netherlands); J. Aumann, K. Cuppens, D. Degraeve, K. Demuyne, B. Dieriks, K. Pat, L. Spaas, R. Van Puijenbroek, K. Weytjens, J. Wynants (Jessa Hospital, Hasselt, Belgium); V. Adam, B.J. Berendes, E. Hardeman, P. Jordens, E. Munghen, K. Tournoy, P. Vercauter (Onze-Lieve-Vrouw Hospital, Aalst, Belgium); T. Alame, M. Bruyneel, M. Gabrovska, I. Muylle, V. Ninane, D. Rozen, P. Rummens, S. Van Den Broecke (Saint-Pierre Hospital, Brussels, Belgium); A. Froidure, S. Gohy, G. Liistro, T. Pieters, C. Pilette, F. Pirson (Université Catholique de Louvain, Brussels, Belgium); H. Kerstjens, M. Van den Berge, N. Ten Hacken, M. Duiverman, D. Koster (University Medical Center Groningen, Groningen, The Netherlands); B. Vosse, L. Conemans, M. Maus, M. Bischoff, M. Rutten, D. Agterhuis, R. Sprooten (Maastricht University Medical Center, Maastricht, The Netherlands); B. Beutel, A. Jerrentrup, A. Klemmer, C. Viniol, C. Vogelmeier (University Medical Center, Marburg, Germany); H. Bode, C. Dooms, D. Gullentops, W. Janssens, K. Nackaerts, D. Rutens, E. Wauters, W. Wuys (University Hospital Leuven, Leuven, Belgium); E. Derom, S. Dobbelaere, S. Loof, G. Serry, B. Putman, L. Van Acker, Y. Vandeweygaerde (Ghent University Hospital, Ghent, Belgium); M. Criel, M. Daenen, R. Gubbelmans, S. Klerkx, E. Michiels, M. Thomeer, A. Vanhauwaert (Hospital Oost-Limburg, Genk, Belgium).

Conflict of interest: M. Topalovic has nothing to disclose. N. Das has nothing to disclose. P-R. Burgel reports personal fees from AstraZeneca, Boehringer Ingelheim, Chiesi, Novartis, Teva and Vertex, outside the submitted work. M. Daenen has nothing to disclose. E. Derom has nothing to disclose. C. Haenebalcke reports personal fees from Novartis, Chiesi, GSK and AstraZeneca, outside the submitted work. R. Janssen has nothing to disclose. H.A.M. Kerstjens has nothing to disclose. G. Liistro has nothing to disclose. R. Louis reports grants and personal fees from GSK and Novartis, personal fees from AstraZeneca, and grants from Chiesi, outside the submitted work. V. Ninane has nothing to disclose. C. Pison has nothing to disclose. M. Schlessler has nothing to disclose. P. Vercauter has nothing to disclose. C.F. Vogelmeier reports personal fees from Almirall, Cipla, Berlin-Chemie/Menarini, CSL Behring and Teva, grants and personal fees from AstraZeneca, Boehringer Ingelheim, Chiesi, GSK, Grifols, Mundipharma, Novartis and Takeda, grants from German Federal Ministry of Education and Research (BMBF) Competence Network Asthma and COPD (ASCONET), Bayer Schering Pharma AG, MSD and Pfizer, outside the submitted work. E. Wouters reports personal fees for board membership from Nycomed and Boehringer, grants from AstraZeneca and GSK, and personal fees for lectures from AstraZeneca, GSK, Novartis and Chiesi, outside the submitted work. J. Wynants has nothing to disclose. W. Janssens has nothing to disclose.

Support statement: This work was supported by the Vlaams Agentschap Innoveren & Ondernemen (VLAIO, government body, 2016–2018). The funder had no role in study design and conduct of the study; collection, management, analysis and interpretation of the data; preparation, review or approval of the manuscript; and decision to submit the manuscript for publication. Funding information for this article has been deposited with the Crossref Funder Registry.

References

- 1 Crapo RO. Pulmonary-function testing. *N Engl J Med* 1994; 331: 25–30.
- 2 Pellegrino R, Viegi G, Brusasco V, *et al.* Interpretative strategies for lung function tests. *Eur Respir J* 2005; 26: 948–968.
- 3 Reddel HK, Bateman ED, Becker A, *et al.* A summary of the new GINA strategy: a roadmap to asthma control. *Eur Respir J* 2015; 46: 622–639.
- 4 Vogelmeier CF, Criner GJ, Martinez FJ, *et al.* Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report: GOLD Executive Summary. *Eur Respir J* 2017; 49: 1700214.
- 5 Galie N, Humbert M, Vachiery JL, *et al.* 2015 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension: The Joint Task Force for the Diagnosis and Treatment of Pulmonary Hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS); Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC), International Society for Heart and Lung Transplantation (ISHLT). *Eur Respir J* 2015; 46: 903–975.
- 6 Martinez FJ, Chisholm A, Collard HR, *et al.* The diagnosis of idiopathic pulmonary fibrosis: current and future approaches. *Lancet Respir Med* 2017; 5: 61–71.
- 7 Topalovic M, Laval S, Aerts JM, *et al.* Automated interpretation of pulmonary function tests in adults with respiratory complaints. *Respiration* 2017; 93: 170–178.
- 8 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.
- 9 Esteva A, Kuprel B, Novoa RA, *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; 542: 115–118.
- 10 Ting DSW, Cheung CY, Lim G, *et al.* Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* 2017; 318: 2211–2223.
- 11 Ehteshami Bejnordi B, Veta M, Johannes van Diest P, *et al.* Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017; 318: 2199–2210.

- 12 Thrall JH, Li X, Li Q, *et al.* Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. *J Am Coll Radiol* 2018; 15: 504–508.
- 13 Armstrong S. The computer will assess you now. *BMJ* 2016; 355: i5680.
- 14 Fridsma DB. Health informatics: a required skill for 21st century clinicians. *BMJ* 2018; 362: k3043.
- 15 The Lancet. Artificial intelligence in health care: within touching distance. *Lancet* 2018; 390: 2739.
- 16 Culver BH, Graham BL, Coates AL, *et al.* Recommendations for a standardized pulmonary function report. An official American Thoracic Society technical statement. *Am J Respir Crit Care Med* 2017; 196: 1463–1472.
- 17 Topalovic M, Das N, Troosters T, *et al.* Applying artificial intelligence on pulmonary function tests improves the diagnostic accuracy. *Eur Respir J* 2017; 50: Suppl. 61, OA3434.
- 18 Miller MR, Crapo R, Hankinson J, *et al.* General considerations for lung function testing. *Eur Respir J* 2005; 26: 153–161.
- 19 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.
- 20 Quanjer PH, Tammeling G, Cotes J, *et al.* Lung volumes and forced ventilatory flows. *Eur Respir J* 1993; 6: Suppl. 16, 5–40.
- 21 Stanojevic S, Graham BL, Cooper BG, *et al.* Official ERS technical standards: Global Lung Function Initiative reference values for the carbon monoxide transfer factor for Caucasians. *Eur Respir J* 2017; 50: 1700010.
- 22 Garg AX, Adhikari NK, McDonald H, *et al.* Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA* 2005; 293: 1223–1238.
- 23 Filippi A, Sabatini A, Badioli L, *et al.* Effects of an automated electronic reminder in changing the antiplatelet drug-prescribing behavior among Italian general practitioners in diabetic patients: an intervention trial. *Diabetes Care* 2003; 26: 1497–1500.
- 24 Willems JL, Abreu-Lima C, Arnaud P, *et al.* The diagnostic performance of computer programs for the interpretation of electrocardiograms. *N Engl J Med* 1991; 325: 1767–1773.
- 25 Hankinson JL. Automated pulmonary function testing: interpretation and standardization. *Ann Biomed Eng* 1981; 9: 633–643.
- 26 Krumpal P, Weigt G, Martinez N, *et al.* Computerized rapid analysis of pulmonary function test: use of a least mean squares correlation for interpretation of data. *Comput Biol Med* 1982; 12: 295–307.
- 27 Berry CE, Wise RA. Interpretation of pulmonary function test: issues and controversies. *Clin Rev Allergy Immunol* 2009; 37: 173–180.
- 28 Enright P. Flawed interpretative strategies for lung function tests harm patients. *Eur Respir J* 2006; 27: 1322–1323.
- 29 Miller MR, Quanjer PH, Swanney MP, *et al.* Interpreting lung function data using 80% predicted and fixed thresholds misclassifies more than 20% of patients. *Chest* 2011; 139: 52–59.
- 30 Visentin E, Nieri D, Vagaggini B, *et al.* An observation of prescription behaviors and adherence to guidelines in patients with COPD: real world data from October 2012 to September 2014. *Curr Med Res Opin* 2016; 32: 1493–1502.
- 31 Quanjer PH, Enright PL, Miller MR, *et al.* The need to change the method for defining mild airway obstruction. *Eur Respir J* 2011; 37: 720–722.
- 32 Leuppi JD, Miedinger D, Chhajed PN, *et al.* Quality of spirometry in primary care for case finding of airway obstruction in smokers. *Respiration* 2010; 79: 469–474.