EUROPEAN RESPIRATORY *journal*

FLAGSHIP SCIENTIFIC JOURNAL OF ERS

Early View

Original research article

# Collaboration between explainable artificial intelligence and pulmonologists improves the accuracy of pulmonary function test interpretation

Nilakash Das, Sofie Happaerts, Iwein Gyselinck, Michael Staes, Eric Derom, Guy Brussselle, Felip Burgos, Marco Contoli, Anh Tuan Dinh-Xuan, Frits M. E. Franssen, Sherif Gonem, Neil Greening, Christel Haenebalcke, William D-C. Man, Jorge Moisés, Rudi Peché, Vitalii Poberezhets, Jennifer K. Quint, Michael C. Steiner, Eef Vanderhelst, Mustafa Abdo, Marko Topalovic, Wim Janssens

Please cite this article as: 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; in press (https://doi.org/10.1183/13993003.01720-2022).

**Collaboration between explainable artificial intelligence and pulmonologists improves the accuracy of pulmonary function test interpretation.**


Authors

Nilakash Das[1], Sofie Happaerts[2], Iwein Gyselinck[1,2], Michael Staes[1,2], Eric Derom[3], Guy Brussselle[3], Felip Burgos[4], Marco Contoli[5], Anh Tuan Dinh-Xuan[6], Frits M.E. Franssen[7], Sherif Gonem[8], Neil Greening[9], Christel Haenebalcke[10], William D-C.Man[11,12], Jorge Moisés[13], Rudi Peché[14], Vitalii Poberezhets[15], Jennifer K Quint[11,12], Michael C. Steiner[9], Eef Vanderhelst[16], Mustafa Abdo[17], Marko Topalovic[20], and Wim Janssens[1,2]

Author' affiliations

[1] Laboratory of Respiratory Diseases and Thoracic Surgery, Department of Chronic Diseases Metabolism and Ageing, KU LEUVEN, Leuven, Belgium

[2] Clinical department of Respiratory Diseases, University Hospitals Leuven, Leuven, Belgium

[3] UZ Gent, University of Ghent, Belgium

[4] Department of Pulmonary Medicine, Hospital Clinic-Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), University of Barcelona, Barcelona, Spain

[5] Department of Translational Medicine, University of Ferrara, Ferrara, Italy

[6] Service de Physiologie-Explorations Fonctionnelles, Assistance Publique-Hôpitaux de Paris, Hôpital Cochin, Université Paris Cité, F-75014 Paris, France

[7] Department of Respiratory Medicine and School of Nutrition and Translational Research in Metabolism (NUTRIM), Maastricht University Medical Center, Maastricht, Netherlands

[8] Nottingham University Hospitals NHS trust, UK

[9] Leicester NIHR Biomedical Research Centre – respiratory, Department of Respiratory Sciences, University of Leicester, UK

[10] AZ Sint-Jan Brugge-Oostende, Belgium

[11] National Heart and Lung Institute, Imperial College London, UK

[12] Royal Brompton and Harefield Clinical Group, Guy's and St.Thomas' NHS Foundation Trust, UK

[13] Biomedical Research Networking Center on Respiratory Diseases (CIBERES), Madrid, Spain

[14] CHU Charleroi, Belgium

[15] Department of Propedeutics of Internal Medicine, National Pirogov Memorial Medical University, Vinnytsya, Ukraine

[16] University Hospital of Brussels, Vrije Universiteit Brussel, Belgium

[17] LungenClinic Grosshansdorf, Germany

[18] ArtiQ NV, Leuven, Belgium

Corresponding author

Wim Janssens

O&N1 Herestraat 49 bus 706

3000 Leuven, Belgium

Email: wim.janssens@uzleuven.be

Manuscript word count

3171

**Conflicts of interest statement**

ND, SH, IG, MS, ATD, FF, SG, NG, CH, JM, RP, VP, MS, MA report no conflicts of interest. ED reports consultancy fees from Chiesi, GSK, AstraZeneca, Boehringer-Ingelheim. GB reports payment honoraria for lectures from Astra Zeneca, Boehringer-Ingelheim, Chiesi, GlaxoSmithKline, Novartis, Sanofi. FB reports consultancy fees from Medical Graphics Coorporation Diagnostics. MC reports grants from Chiesi and GlaxoSmithKline, consultancy fees and honoraria from Astra Zeneca, Boehringer-Ingelheim, Chiesi, GlaxoSmithKline, Novartis as well as support for attending meetings from Chiesi, Astra Zeneca, GSK and ALK-ABELLO. WM reports grants from NIHR Research for Patient Benefit and British Lung Foundation, as well as honoraria from Mundipharma, Novartis, European Conference and Incentive Services DMC and to be Honorary President of Association for Respiratory Technology and Physiology (ARTP, UK). JQ reports grants from MRC, HDR UK,

GlaxoSmithKline, Astra Zeneca, Chiesi and consultancy fees from Insmed, Evidera. EV reports grants from Chiesi and consultancy fees and honoraria from Boehringer Ingelheim, Vertex, GlaxoSmithKline. MT reports to have stock options from ARTIQ. WJ reports grants from Chiesi and AstraZeneca, consultancy and lecture fees from AstraZeneca, Chiesi and GSK, stock options from ARTIQ.

**Funding:** The study was supported by a VLAIO research grant of ARTIQ and KU Leuven (HB.2020.2406). ND, IG and WJ are supported by the Flemish Research Funds (FWO Vlaanderen)

**Abstract**

**Rationale**

Few studies have investigated the collaborative potential between artificial intelligence (AI) and pulmonologists for diagnosing pulmonary disease. We hypothesized that the collaboration between pulmonologist and AI with explanations (explainable AI, XAI) is superior in diagnostic interpretation of pulmonary function tests (PFTs) than a pulmonologist without support.

**Materials and methods**

The study was conducted in two phases, a mono-centre (P1) and a multi-centre intervention study (P2). Each phase utilized two different sets of 24 PFT reports of patients with a clinically validated gold-standard diagnosis. Each PFT was interpreted without (control) and with XAI's suggestions (intervention). Pulmonologists provided a differential diagnosis consisting of a preferential diagnosis and optionally up to three additional diagnoses. Primary endpoint compared accuracy of preferential and additional diagnoses between control and intervention. Secondary endpoints were number of diagnoses in differential diagnosis, diagnostic confidence and inter-rater agreement. We also analysed how XAI influenced pulmonologists' decisions.

**Results**

In P1 (N=16 pulmonologists), mean preferential and differential diagnostic accuracy significantly increased by 10.4% and 9.4%, respectively between control and intervention ($p<0.001$). Improvements were somewhat lower but highly significant ($p<0.0001$) in P2 (5.4% and 8.7% respectively, N=62 pulmonologists). In both phases, the number of diagnoses in differential diagnosis did not reduce, but diagnostic confidence and inter-rater agreement significantly increased during intervention. Pulmonologists updated their decisions with XAI's feedback and consistently improved their baseline performance if AI provided correct predictions.

**Conclusion**

A collaboration between pulmonologist and XAI is better at interpreting PFTs than individual pulmonologists reading without XAI support or XAI alone.

## Introduction

When correctly interpreted, pulmonary function tests (PFTs) are a useful tool to address the differential diagnosis of respiratory diseases (1). However, interpretation of PFTs requires expertise in combining the understanding of normal values, lung function patterns (obstructive, restrictive, mixed and normal) and appearance of flow-volume curves within the patient's medical history, clinical presentation and results of other diagnostic assessments.(2,3) Although various algorithms exist to aid the interpretation of PFTs (4,5), it has been shown that neither pulmonologists nor ATS/ERS's guidelines derived algorithms are sufficiently accurate for a correct reading (6,7).

It could be argued that artificial intelligence (AI) may help in automating the complex reasoning that entails the process of interpreting PFTs. Indeed, when all the PFT indices are taken together, the data-based AI approach captures subtle characteristics of respiratory disorders that are not always identified by the clinician, resulting in a powerful algorithm for differential diagnosis [8]. In the past, such AI driven algorithms have been shown to perform as well, if not better than pulmonologists alone and might help support pulmonologists to interpret lung function [6]. However, most clinical studies often report AI outperforming clinicians' diagnostic performance in head-to-head comparisons [9, 10], giving way to an irrational claim that clinicians will soon be replaced by AI-equipped devices. Unlike the narrow task-based scope of AI, clinicians carry out a multitude of duties involving diagnostics, treatment and management of patients, while also bringing a vital element of empathy to healthcare [11]. While clinicians are irreplaceable, there remains a vast potential for AI and clinicians to work together in improving routine clinical outcomes [11]. Presently, there exists no data on the benefits of a collaboration between AI and a pulmonologist at interpreting PFTs. Further, AI algorithms are often regarded as black boxes, i.e. they cannot provide explanations on their output [12]. Understanding the rationale behind a prediction is critical to gaining trust, especially if a clinician plans an action based on the algorithm's output. On the other hand, it has also been suggested that explanations may help in mitigating automation bias, errors that arise from over-reliance on AI systems [13]. Today, several methods exist that allow us to produce explanations, rendering AI more transparent, hence easier to decipher. This new paradigm of AI is called explainable AI (XAI) [14].

In this study, we hypothesized that a pulmonologist with the help of XAI's suggestions would be superior at interpreting PFTs to the pulmonologist working alone. Our primary goal was to

compare the preferential and differential diagnostic accuracy between the pulmonologist's view (control) and the pulmonologist's view assisted with suggestions provided by a machine-learning model (intervention) [6]. We also compared whether the intervention was better than the AI's standalone diagnostic performance. Additionally, we investigated how pulmonologists updated their diagnostic choices following the assistance of XAI.

## Methods

### Study design

In this study with a repeated measures design, pulmonologists were requested to interpret 24 anonymized PFT reports including pre-and/or post-bronchodilator spirometry, lung volumes, airway resistance and diffusing capacity (with access to Z-scores and data colour coding indicating deviation from normal). Limited clinical information (smoking history and symptom presentation) was also provided. Each PFT report was interpreted in two steps: first, a control step **(a)** in which pulmonologists provided their responses after reading the PFT report only, then an intervention step **(b)** in which pulmonologists provided their responses for the same report with suggestions of XAI available to them. Thus, each pulmonologist performed 48 interpretations in one exercise.

We carried out the study in two phases. The first phase (P1) was a monocentric study in which 16 out of 25 invited pulmonologists from University Hospital Leuven completed the study. In the second phase (P2), 62 out of 88 invited pulmonologists from across European institutions completed the study (supplement S1). P2 was initiated only after we observed that primary endpoints in P1 were met. The set of 24 PFT reports differed completely between the two phases.

We used an online platform called Gorilla to carry out the study [15]. Participants could complete the study at their own pace with no time limits. They began by indicating their informed consent, years of clinical experience (< or >5 years), any prior experience with AI-based clinical decision support system (Yes/No), and their enthusiasm on AI applications in general on a 5-point Likert scale (supplement S2).

Afterwards, participants were guided to complete a tutorial to familiarise themselves with the online platform and XAI's suggestions (supplement S3). During the main tasks, pulmonologists provided a differential diagnosis including a mandatory preferential diagnosis and up to three

additional diagnoses ranked in the order of preference. The diagnostic choices were: 1) healthy or normal, 2) asthma (including obstructive or non-obstructive), 3) chronic obstructive pulmonary disease (COPD) (including emphysema or chronic bronchitis), 4) interstitial lung disease (ILD) (including idiopathic pulmonary fibrosis and non-idiopathic pulmonary fibrosis), 5) neuromuscular disease (NMD) (including diaphragm paralysis), 6) other obstructive disease (OBD) (including cystic fibrosis, bronchiectasis, bronchiolitis), 7) thoracic deformity (TD) (including pleural disease, pneumonectomy), and 8) pulmonary vascular disease (PVD) (including pulmonary hypertension, vasculitis, chronic thromboembolic pulmonary hypertension).

The pulmonologists also provided an overall confidence of their diagnosis on a 5-point Likert scale (1=least confidence, 5=highest confidence). In addition, they indicated their level of agreement with XAI's suggestion on a 5-point Likert scale (1=strongly disagree, 2=disagree, 3=neutral, 4=agree, 5=strongly agree) in the intervention phases. Supplement S3 shows an example of a control and interventional phase for one particular PFT report.

An ethics committee approval was obtained for P1 (study no S60243) while a separate ethics committee approval was obtained for the international multicentre P2 phase (S65162).

### Pulmonary function test cases

Between November 2017 and August 2018 at UZ Leuven, 1003 subjects performed complete lung function testing. All PFTs were performed with standardised equipment by respiratory operators (Masterlab, Würzburg, Germany), according to the ATS/ ERS criteria [16]. GLI equations were used to calculated reference values for spirometric $FEV_1$, FVC and $FEV1_1$/FVC [17], while ECCS 93 was used at that time period for diffusion capacity, lung-volumes and airway resistance measurements[18]. A single clinician assigned a preliminary diagnosis across each of the eight disease categories in 794 subjects by referring electronic health records of clinical history, symptoms, PFT reports and additional tests. A high prevalence of COPD (23%), ILD (25%), asthma (9%) and normal (30%) subjects characterized the sample. All subjects were Caucasians older than 18 years. From this group, we shortlisted 92 subjects, by randomly selecting 15 subjects from each of the most prevalent groups (COPD, asthma, ILD and normal lung function), and 8 subjects from each of the least prevalent diseases (NMD, TD, PVD and OBD). Two pulmonologists jointly adjudicated the gold standard diagnosis in each of these cases using all available clinical data including PFT. If there was disagreement or doubt on the

diagnosis another case was selected to end-up with a set of 24 PFT cases with a gold standard diagnosis, for P1 and P2 separately. In each set, we randomly included four subjects from the most prevalent disease and two subjects from the least prevalent diseases. We then slightly inflated the sample of incorrectly predicted cases by the AI to study how clinicians would respond to incorrect AI's suggestions. Following an additional review by the pulmonologists, three cases in each set that were correctly predicted by the AI were deliberately replaced by cases in which the AI did not correctly predict the adjudicated gold standard diagnosis. Thus in both sets, the preferential diagnostic accuracy of the AI was set at 62.5% or 15/24 cases, which was lower than its reported validation accuracy of 74%) [6].

### Explainable Artificial Intelligence (XAI)

We used our previously reported machine-learning model that predicts eight respiratory disorders (COPD, asthma, ILD, healthy, NMD, TD, PVD and OBD) [6]. Its preferential diagnostic accuracy (disease with the highest calculated probability) was reported at 74% during inter-validation, while similar accuracies (76%-82%) were also observed during testing on external cohorts [6]. In this study we also reported explanations on AI's second diagnostic suggestion when its probability was higher than 15%, in addition to explanations for AI's preferential diagnosis. To render the AI model explainable, we used a game-theoretic concept called Shapley values (SVs) to estimate the evidence of different PFT indices towards AI's diagnostic suggestions [19]. A positive SV is interpreted as evidence supporting the model's prediction while a negative SV is counter-evidence. The magnitude of SV denotes the strength of the contribution. For each diagnostic suggestion, we included a Shapley value plot of the top five PFT indices in descending order of magnitude of evidence. We also normalised the SVs by dividing them by the highest magnitude. We show an example of a PFT case with XAI's suggestions in figure 1.

### Study endpoints

Our primary endpoint was to compare pulmonologists' mean preferential and differential diagnostic accuracy between the control and the interventional setting. The mean preferential accuracy is calculated as the number of cases in which a pulmonologists' preferential diagnosis matched the gold standard, averaged over the entire cohort. Mean differential accuracy is calculated as the number of times in which a pulmonologists' differential diagnosis (preferential diagnosis + additional diagnoses) included the gold standard, averaged over the

entire cohort. As secondary endpoints, we explored the number of additional diagnoses, clinicians' diagnostic confidence on the overall diagnostic performance as well as their inter-rater agreement on the preferential diagnosis. We also analysed how pulmonologists updated their diagnostic decisions between control and intervention, and further studied if pulmonologists followed XAI's incorrect suggestions indicating automation bias.

### Sample size calculation

The minimum sample size for pulmonologists was calculated at 11, using two-sided paired t-test with the assumption that the accuracy of both preferential and differential diagnosis improves between control and intervention with a mean of 3 cases out of 24 (12.5%), a standard deviation of 3 cases, a significance of 0.05 and power of 0.8. The premise of our assumption is that the interventional setting will show a mean improvement in preferential and differential diagnostic accuracy of at-least 10% [6].

### Statistical analysis

We evaluated our quantitative endpoints using paired t-test. Inter-observer agreement in preferential diagnostic choice was assessed using Fleiss' Kappa. Continuous variables were assumed to be normally distributed with homogenous variance, and Shapiro-Wilk test was used to test assumptions of normality. We performed all our analysis on R statistical software using a significance level of 0.05.

## Results

### Participant demographics

P1 and P2 saw the participation of 16 and 62 pulmonologists respectively (table, online supplement S4). More than 3/4$^{th}$ of the participants in both phases had at least 5 years of clinical experience. Over half of P1 participants had prior experience with AI-based decision support systems, but that percentage was much lower in P2 (11%). Mean baseline enthusiasm in AI on a 5-point Likert scale was high in both groups (3.56 and 3.92 respectively) suggesting an overall bias towards accepting AI decisions.

### PFT sample characteristics and baseline XAI's performance

PFT sample characteristics were similar for P1 and P2 (N=24 each), and shown in Table 1. Both samples included four groups each of high prevalence (COPD, asthma, ILD and normal lung function) and two diseases each of low prevalence (NMD, TD, PVD and OBD).

AI's preferential diagnosis was set to match the gold standard (GS) in 15 out of 24 cases (62.5%) in both P1 and P2 samples, while its differential diagnosis (preferential diagnosis + second diagnostic suggestion) included the gold standard in 22 (91.7%) of the P1 and 21 (87.5%) of the P2 cases. A breakdown of AI's diagnostic performance across different disease groups is given (table, online supplement S5).

### Primary endpoints

In P1, the use of explainable AI improved the mean preferential and differential diagnostic accuracy by 10.4% and 9.4% respectively ($p<0.001$), which was somewhat higher than in P2 (5.4% and 8.7% respectively, $p<0.0001$). Thus, primary endpoints were met as mean diagnostic accuracies significantly increased between control (pulmonologist) to intervention (pulmonologist + XAI) (Table 2, Figure 2A/2B). However, the improvements were smaller than anticipated (12.5%) from our sample size estimation.

When we compared the diagnostic performance between XAI and the intervention setting (pulmonologists + XAI) as an exploratory analysis, we also observed a mean improvement of 13% ($p<0.0001$) and 3.1% ($p=0.01$) for preferential and differential diagnostic accuracy in P1 (N=16), which was similar to P2 (N=62) with a mean improvement of 12.25% and 2.9% respectively. Thus, we noted that pulmonologists with the help of XAI's suggestion not only

improved their individual performance, but they also significantly outperformed AI's predictive performance in both P1 and P2 (Figure, online supplement S6).

### Secondary endpoints

We included a number of secondary endpoints in our study (Table 3). In both studies, mean Likert scale confidence in diagnosis significantly increased ($p<0.01$), while number of differential diagnostic choices remained unchanged between control and intervention. Fleiss's kappa quantifying inter-clinician agreement in preferential diagnosis also increased. Pulmonologists indicated a moderately high level of agreement with suggestions of XAI.

### Demographics based performance

In P2 (N=62), we further analysed the diagnostic performance of the enhanced setting (pulmonologist + XAI) by stratifying on experience. We observed no significant differences in interventional diagnostic accuracies between participants with <5 (N=12) and > 5 years (N=50) of experience. Similarly, no significant differences were observed when the subjects were stratified on their baseline enthusiasm in AI applications (online supplement S7).

### Change in responses

In both phases, pulmonologists' diagnostic responses changed between control and intervention in almost half of the 24 cases (Table 4). Diagnostic confidence at baseline was significantly lower in cases where responses changed as compared to cases in which responses remained unchanged. Whenever responses changed, we observed a significant improvement ($p<0.001$) in differential diagnostic accuracy: In the 55% changed cases of P1, the differential diagnosis contained the GS in 78% within the control arm compared to 95% after the intervention; in the 48% changed cases in P2, the differential diagnosis included the GS in 73% of control arm versus 91% after the intervention. The changed responses always contained at-least one diagnostic suggestion of XAI.

### Automation bias

We studied if pulmonologists' performance reduced between control and intervention whenever AI suggested a correct or incorrect preferential diagnosis (9 cases in P1 and P2 respectively) (online supplement S8). While it was found that preferential diagnostic accuracy reduced slightly but significantly in case an incorrect XAI diagnosis was given, we observed much larger increases in accuracy when the XAI diagnosis was correct. We also observed that pulmonologists placed a significantly higher ($p<0.001$) level of agreement with XAI's

suggestions in cases with correct preferential predictions as opposed to with incorrect preferential predictions, indicating little risk for automation bias.

## Discussion

In this study conducted in two separate phases, we observed that pulmonologists when aided by XAI significantly improved on their individual preferential and differential diagnostic accuracy in interpreting PFTs. Among secondary endpoints, we noted a significant increase in diagnostic confidence but no reduction in the number of differential diagnostic choices. Our results support the hypothesis that a pulmonologist aided by XAI improves on the interpretation of PFTs for differential diagnosis of respiratory diseases when compared to individual pulmonologists with no support. Interestingly, we also observed that pulmonologists when aided by XAI significantly outperformed XAI itself in preferential and differential diagnostic accuracy.

Most clinical studies involving AI have emphasized the diagnostic superiority of AI using head-to-head comparisons [10], while few have studied the benefits of a collaborative approach. In-fact, our post-hoc head-to-head comparison revealed no clear differences in diagnostic accuracy between AI and individual pulmonologists in both P1 and P2. This was expected because unlike most studies that typically compare AI with non-experts diluting average human performance, our participants were respiratory medicine specialists. It is likely that the use of XAI will be even more beneficial when used by medical practitioners less experienced in interpreting PFTs. Although this was not the aim of our study, the use of XAI could be expanded to these populations if proven advantageous. Secondly, a lower than expected improvement can also be explained by the fact that we purposefully included PFT cases in which AI made mistakes to study the effect of incorrect predictions on clinicians' decision making. A random selection of cases based on actual disease prevalence in the real word would have seen a higher AI accuracy and pushed up pulmonologist's performance by a larger margin.

The superiority of the collaborative approach is in-line with several clinical decision support systems (CDSSs) that have been reported to improve practitioners' performance in the past [20]. Our study adopted a repeated measures design instead of a placebo-controlled trial, not only due to the limited availability of participants. We also wanted to recreate a setting in which the pulmonologist arrives at a diagnostic work-up and updates, if needed, based on an automated protocol. Although there might be an element of learning effect present through the repeated measure design, our results showed that XAI's suggestions effected a change in pulmonologists' responses in almost of half of the cases. Whenever responses changed,

pulmonologists were more likely to improve over their baseline performance. An analysis of changed responses revealed that the updated diagnosis always contained at-least one diagnostic suggestion of XAI.

Our study also allowed a preliminary investigation into automation bias, a known error that arises due to clinicians over relying on CDSS's output even when it is incorrect [13]. Present results showed that pulmonologists preferential diagnostic performance decreased slightly whenever AI made incorrect predictions, but increased largely when a correct diagnostic suggestion was made. Moreover, agreement with XAI's suggestions was significantly higher ($p<0.0001$) with correct suggestions as compared to those in which AI made incorrect predictions, indicating only limited risk for automation bias. Researchers have suggested that explanations, as we provided with the Shapley values, allow the clinician to develop an internal picture on how the system operates. It has the potential to mitigate misplaced trust and over-reliance on CDSS [13, 21]. Nonetheless, a controlled study with and without explanations must be conducted to conclusively establish the impact of explanations on automation bias.

The current study is in line with the novel ATS/ERS standards for lung function interpretation stating that PFT are to detect and quantify disturbances of the respiratory system[22]. Based on certain patterns, clinicians will use PFT in their diagnostic work-up towards a preferential diagnosis and a reduced list of differential diagnoses. As the AI and XAI algorithms provide probability estimates for diagnostic disease clusters but no final disease diagnoses, they completely support this clinical diagnostic process. A major limitation of our study is that our definition of diagnostic superiority as a positive outcome may be construed as narrow in scope. In a real life, a diagnostic work-up is achieved through an extensive anamnesis, clinical exam and a multitude of tests like FeNO and histamine challenge, blood samples and even CT scan, which were not available to the pulmonologists in the current study. Vice versa, future AI models may also benefit from this multimodal layers of information to improve on their granularity and accuracy. Our study could also have benefitted from a larger sample of PFT reports as the current sample over represents disease groups like NMD, TD, OBD and PVD. It distorts the actual prevalence of diseases that pulmonologists routinely encounter in clinical practice. Due to the limited sample size and the good individual baseline performance of clinicians, the improvements in diagnostic accuracy from introducing AI were small and maybe clinically not very relevant. The lack of ethnic diversity was also a major limitation that hinders extrapolation of current results to the general population. In the future, prospective studies

using randomised clinical trial settings including less experienced practitioners and using PFTs of a more diverse population, with specific endpoints like time to final diagnosis, number of diagnostic or redundant tests, total costs for the healthcare system etc. are required to establish the real effectiveness of XAI.

To conclude, our study demonstrates that pulmonologists can improve their individual diagnostic interpretation of PFTs with the help of AI. Such teamwork between AI and clinicians may become commonplace in the future, with the potential to drive healthcare improvements particularly in areas where clinical expertise is less available.

**Table 1**

Overview of pulmonary function test (PFT) characteristics in monocentric phase 1 (P1) and multicentric phase 2 (P2) studies, with 24 PFT reports each. Values are expressed as minimum-maximum.

| | P1 study | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Healthy** | **COPD** | **Asthma** | **ILD** | **NMD** | **OBD** | **TD** | **PVD** |
| **N** | 4 | 4 | 4 | 4 | 2 | 2 | 2 | 2 |
| **Sex, F/M** | 3/1 | 3/1 | 2/2 | 3/1 | 0/2 | 2/0 | 0/2 | 1/1 |
| **Age, years** | 36 to 62 | 58 to 72 | 26 to 48 | 51 to 84 | 59 to 59 | 20 to 49 | 65 to 67 | 70 to 82 |
| **PY** | 0 to 0 | 30 to 56 | 0 to 5 | 0 to 12 | 10 to 35 | 0 to 0 | 0 to 25 | 0 to 30 |
| **FEV1, Z-score** | -0.78 to 1.05 | -3.08 to -1.14 | -4.13 to -0.41 | -3.84 to 0.64 | -4.41 to -3.25 | -4.87 to -3.88 | -4.09 to -1.34 | -0.33 to 1.19 |
| **FVC, Z-score** | -0.93 to 0.93 | -1.16 to 0.22 | -1.61 to -0.41 | -4.31 to -1.65 | -5.02 to -3.7 | -3.59 to -0.95 | -4.8 to -1.63 | -1.01 to 1.44 |
| **FEV1/FVC, %** | 77 to 86 | 54 to 64 | 49 to 82 | 83 to 90 | 77 to 81 | 43 to 60 | 79 to 80 | 72 to 89 |
| **RV, Z-score** | -1.44 to 0.19 | 1.21 to 2.28 | -0.63 to 2.89 | -3.43 to -2.08 | -1.64 to -0.61 | 3.68 to 3.73 | -3.39 to -2.27 | 0.44 to 0.88 |
| **TLC, Z-score** | -1.49 to 1.3 | -0.34 to 1.13 | -0.37 to 1.06 | -4.12 to -2.01 | -3.37 to -3.37 | -0.07 to 1.84 | -4.78 to -2.92 | -0.09 to 0.54 |
| **DLCO, Z-score** | -0.81 to 0.45 | -2.66 to 0.4 | -1.12 to -0.38 | -3.86 to -1.81 | -1.96 to -0.91 | -2.45 to -2.25 | -3.24 to -2.46 | -3.29 to -2.39 |
| **KCO, Z-score** | -0.97 to 2.24 | -1.88 to 0.39 | -0.26 to 0.71 | -2.24 to 0.95 | 1.52 to 4.72 | -0.22 to 1.31 | 0.02 to 3.36 | -2.3 to -1.79 |

| | P2 study | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Healthy** | **COPD** | **Asthma** | **ILD** | **NMD** | **OBD** | **TD** | **PVD** |
| **N** | 4 | 4 | 4 | 4 | 2 | 2 | 2 | 2 |
| **Sex, F/M** | 3/1 | 1/3 | 0/4 | 2/2 | 2/0 | 0/2 | 2/0 | 1/1 |
| **Age, years** | 27 to 67 | 48 to 84 | 21 to 59 | 35 to 85 | 31 to 56 | 30 to 68 | 54 to 90 | 50 to 64 |
| **PY** | 0 to 25 | 18 to 50 | 0 to 20 | 0 to 25 | 0 to 3 | 0 to 0 | 0 to 0 | 10 to 20 |
| **FEV1, Z-score** | -0.69 to 0.79 | -4.96 to -1.63 | -1.02 to 1.38 | -4.35 to -0.06 | -4.6 to -4.28 | -5.35 to -2.17 | -3.05 to -2.7 | -1.16 to -1.1 |
| **FVC, Z-score** | -1.21 to 0.7 | -4.1 to 0.19 | -0.16 to 1.85 | -4.39 to -0.18 | -4.83 to -4.82 | -3.56 to -1.39 | -3.03 to -2.92 | -1.35 to -0.16 |
| **FEV1/FVC, %** | 77 to 92 | 49 to 60 | 69 to 73 | 77 to 87 | 81 to 81 | 42 to 61 | 74 to 80 | 67 to 82 |
| **RV, Z-score** | -1.08 to 0.72 | -1.14 to 4.56 | -0.53 to 3.43 | -1.84 to 2.53 | -1.11 to -1.04 | 1.77 to 4.64 | -2.07 to -0.97 | -0.55 to 0.48 |
| **TLC, Z-score** | -0.06 to 0.01 | -2.7 to 2.05 | -0.32 to 2.81 | -3.63 to -1.39 | -2.74 to -2.40 | -0.77 to -0.41 | -3.42 to -3.19 | -1.04 to 0.16 |
| **DLCO, Z-score** | -1.28 to -0.32 | -3.73 to -0.63 | -0.62 to 0.76 | -5.2 to -2.32 | -4.95 to -4.27 | -1.69 to 1.15 | -2.28 to -2.07 | -2.19 to -1.99 |
| **KCO, Z-score** | -0.78 to 0.23 | -0.44 to 0.39 | -0.44 to 0.58 | -1.79 to -0.63 | -1.17 to 2.27 | 0.41 to 1.4 | 0.89 to 1.15 | -1.47 to -0.61 |

**Abbreviations**: PY = pack-years; FEV1 = forced expiratory volume in one second; F = Female; FVC = forced vital capacity; DLCO = diffusing capacity for carbon monoxide; KCO = transfer coefficient for carbon monoxide; M= Male; TLC = total lung capacity; NMD = neuromuscular disease; ILD = interstitial lung diseases; PVD = pulmonary vascular diseases; OBD = other obstructive diseases; TD = Thoracic deformity/ Pleural diseases; COPD = chronic obstructive pulmonary disease.

**Table 2**

Primary endpoints in monocentric phase 1 (P1) and multicentric phase 2 (P2) studies. Values are mean (standard deviation) unless stated otherwise. Differential diagnosis includes preferential diagnosis and up to three additional diagnoses.

| **P1 study (16 pulmonologists)** | | | | | |
|---|---|---|---|---|---|
| | **XAI alone** | **Control (pulmonologist)** | **Intervention (pulmonologist + XAI)** | **Mean improvement: intervention on control** | **Mean improvement: intervention on XAI alone** |
| **Preferential diagnosis= GS** | 62.5% | 65.1% (8.2%) | 75.5% (9.3%) | 10.4% (p<0.001) | 13% (p<0.0001) |
| **Differential diagnosis includes GS** | 91.7% | 85.4% (10.5%) | 94.8% (5.8%) | 9.4% (p<0.0001) | 3.1% (p<0.05) |
| **P2 study (62 pulmonologists)** | | | | | |
| | **XAI alone** | **Control (pulmonologist)** | **Intervention (pulmonologist + XAI)** | **Mean improvement: intervention on control** | **Mean improvement: intervention on XAI alone** |
| **Preferential diagnosis= GS** | 62.5% | 69.3% (9.1%) | 74.6% (7.6%) | 5.4% (p<0.0001) | 12.1% (p<0.0001) |
| **Differential diagnosis includes GS** | 87.6% | 81.7% (11.2%) | 90.4% (8.8%) | 8.7% (p <0.0001) | 2.9% (p<0.05) |

**Abbreviations**: GS= Gold standard

**Table 3**

Secondary endpoints in monocentric phase 1 (P1) and multicentric phase 2 (P2) studies. Values are mean (standard deviation) unless stated otherwise.

| | P1 study (16 pulmonologists) | | |
|---|---|---|---|
| | **Control (pulmonologist)** | **Intervention (pulmonologist + XAI)** | **p** |
| **No of additional diagnoses in the differential diagnosis** | 1.86 (0.32) | 1.8 (0.33) | 0.197 |
| **Diagnostic confidence on Likert Scale (1=least confidence, 5=Most confidence)** | 3.71 (0.5) | 3.98 (0.42) | <0.01 |
| **Agreement with XAI on Likert Scale (1=strongly disagree, 2=disagree, 3=neutral, 4=agree, 5=strongly agree)** | | 3.76 (0.3) | NC |
| **Inter-rate agreement on preferential diagnosis (Fleiss's Kappa)** | 0.52 | 0.64 | NC |
| | **P2 study (62 pulmonologists)** | | |
| | **Control (pulmonologist)** | **Intervention (pulmonologist + XAI)** | **p** |
| **No of additional diagnoses in differential diagnosis** | 1.67 (0.35) | 1.64 (0.32) | 0.22 |
| **Diagnostic confidence on Likert Scale (1=least confidence, 5=Most confidence)** | 3.93 (0.34) | 4.03 (0.34) | <0.0001 |
| **Agreement with XAI on Likert Scale (1=strongly disagree, 2=disagree, 3=neutral, 4=agree, 5=strongly agree)** | | 3.49 (0.36) | NC |
| **Inter-rate agreement on preferential diagnosis (Fleiss's Kappa)** | 0.53 | 0.63 | NC |

Abbreviations: NC= Not calculated

**Table 4**

Change (percentage of cases) in diagnostic responses between control and intervention in phase P1 study with 16 pulmonologists and in phase P2 with 62 pulmonologists. Baseline confidence is the overall diagnostic confidence on a 5-point Likert Scale indicated by pulmonologists during control. Values are mean (+- standard deviation) unless stated otherwise. T-test comparison between baseline Likert scales is given with p value.

| | **P1 study (16 pulmonologists)** | |
|---|---|---|
| | **Percentage of cases** | **Baseline confidence** |
| **Differential diagnosis unchanged** | 45% (16.3%) | 3.87 (0.54) |
| **Differential diagnosis changed** | 55% (16.3%) | 3.56 (0.48) |
| *Preferential diagnosis changed* | 27.1% (10%) | *p<0.01* |
| *Additional diagnoses changed* | 27.9% (11.5%) | |

| | **P2 study (62 pulmonologists)** | |
|---|---|---|
| | **Percentage of cases** | **Baseline confidence** |
| **Differential diagnosis unchanged** | 51.7% (15.8%) | 4.09 (0.36) |
| **Differential diagnosis changed** | 48.5% (15.8%) | 3.76 (0.39) |
| *Preferential diagnosis changed* | 18% (14.2%) | *p< 0.01* |
| *Additional diagnoses changed* | 30.4% (13.9%) | |

**Figure legends**

## 1.1 Figure 1

Figure showing (a) a sample pulmonary function test (PFT) report, and (b) AI's diagnostic suggestions with Shapley value (SV) evidence. GS diagnosis was COPD based on emphysema on CT scan and passive smoke exposure during childhood (normal alpha-1 levels). In this case, AI makes two diagnostic suggestions (COPD and OBD), since the probability of the second disease (OBD) is greater than 15%. Additionally, we show a normalised SV plot of the top 5 PFT indices that contributed towards the prediction of COPD and OBD respectively. A positive SV (in green) is supporting evidence while a negative SV (in red) is counter evidence.

**Abbreviations**: NMD = neuromuscular disease; ILD = interstitial lung diseases; PVD = pulmonary vascular diseases; OBD = other obstructive diseases; TD = Thoracic deformity/ Pleural diseases; COPD = chronic obstructive pulmonary disease.

## 1.2 Figure 2

Percentual change of preferential (figure on the left) and differential diagnostic performance between control (individual pulmonologists) and intervention (pulmonologists and explainable AI (XAI)) in (a) phase P1 study with 16 pulmonologists, and (b) phase P2 with 62 pulmonologists.

**Abbreviations:** GS: Gold standard ; ***= $p<0.001$ ; ****=$p<0.0001$

References


1. Decramer M, Janssens W, Derom E, Joos G, Ninane V, Deman R, Van Renterghem D, Liistro G, Bogaerts K. Contribution of four common pulmonary function tests to diagnosis of patients with respiratory symptoms: A prospective cohort study. *Lancet Respir. Med.* 2013; 1(9):705-13

2. Ranu H, Wilde M, Madden B. Pulmonary function tests. *Ulster Med. J.* 2011; 80: 84–90.

3. Robert OC. Pulmonary-function testing. *N. Engl. J. Med.* 1994; 331: 25–30.

4. Pellegrino R, Viegi G, Brusasco V, Crapo RO, Burgos F, Casaburi R, Coates A, van der Grinten CPM, Gustafsson P, Hankinson J, Jensen R, Johnson DC, MacIntyre N, McKay R, Miller MR, Navajas D, Pedersen OF, Wanger J. Interpretative strategies for lung function tests. *Eur. Respir. J.* 2005; 26: 948–968.

5. Johnson JD, Theurer WM. A stepwise approach to the interpretation of pulmonary function tests. *Am. Fam. Physician.* 2014; 89(5): 359-366.

6. Topalovic M, Das N, Burgel PR, Daenen M, Derom E, Haenebalcke C, Janssen R, Kerstjens HAM, Liistro G, Louis R, Ninane V, Pison C, Schlesser M, Vercauter P, Vogelmeier CF, Wouters E, Wynants J, Janssens W. Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests. *Eur. Respir. J.* 2019; 53(4):1801660.

7. Topalovic M, Laval S, Aerts J-M, Troosters T, Decramer M, Janssens W. Automated Interpretation of Pulmonary Function Tests in Adults with Respiratory Complaints. *Respiration.* 2017; 93(3):170-178.

8. Das N, Topalovic M, Janssens W. Artificial intelligence in diagnosis of obstructive lung disease: current status and future potential. *Curr. Opin. Pulm. Med.* 2018; 24(2):117-123

9. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, Topol EJ, Ioannidis JPA, Collins GS, Maruthappu M. Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies in medical imaging. *BMJ.* 2020 25; 368:m68910.

10. Shen J, Zhang CJP, Jiang B, Chen J, Song J, Liu Z, He Z, Wong SY, Fang PH, Ming WK. Artificial intelligence versus clinicians in disease diagnosis: Systematic review. J*MIR Med. Informatics.* 2019 16;7(3):e10010.

11. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ.* 2019 4;7:e7702.

12. London AJ. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Cent. Rep.* 2019; 49(1):15-21 .

13. Gretton C. Trust and Transparency in Machine Learning-Based Clinical Decision Support. *Human an Machine learning*, 2018. ISBN : 978-3-319-90402-3

14. Adadi A, Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 2018; 2870052.

15. Anwyl-Irvine AL, Massonnié J, Flitton A, Kirkham N, Evershed JK. Gorilla in our midst: An online behavioral experiment builder. *Behav. Res. Methods* 2020; 52: 388–407

16. Miller MR, Crapo R, Hankinson J, Brusasco V, Burgos F, Casaburi R, Coates A, Enright P, van der Grinten CPM, Gustafsson P, Jensen R, Johnson DC, MacIntyre N, McKay R, Navaja D, Pedersen OF, Pellegrino R, Viegi G, Wagner J. General considerations for lung function testing. *Eur. Respir. J.* 2005; 26(1):153-61.

17. Quanjer PH, Stanojevic S, Cole TJ, Baur X, Hall GL, Culver BH, Enright PL, Hankinson JL, Ip MSM, Zheng J, Stocks J, Schindler C. Multi-ethnic reference values for spirometry for the 3-95-yr age range: The global lung function 2012 equations. *Eur. Respir. J.* 2012; 40(6):1324-43.

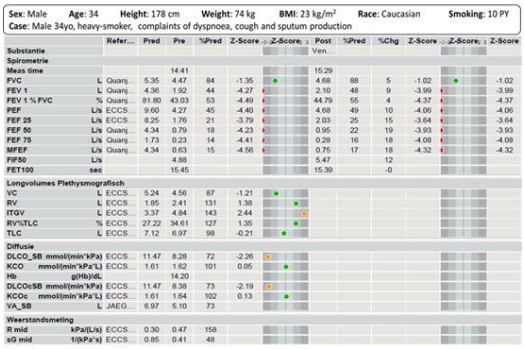
18. Standardized lung function testing. Official statement of the European Respiratory Society. Eur. Respir. J. Suppl. England; 1993. p. 1–100.

19. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. NIPS'17 *Adv. Neural Inf. Process. Syst.* 2017. 4768–4777

20. Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, Sam J, Haynes RB. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review. JAMA. 2005; 293(10):1223-1238.

21. Bussone A, Stumpf S, O'Sullivan D. The role of explanations on trust and reliance in clinical decision support systems. *Proc. - 2015 IEEE Int. Conf. Healthc. Informatics, ICHI 2015*.

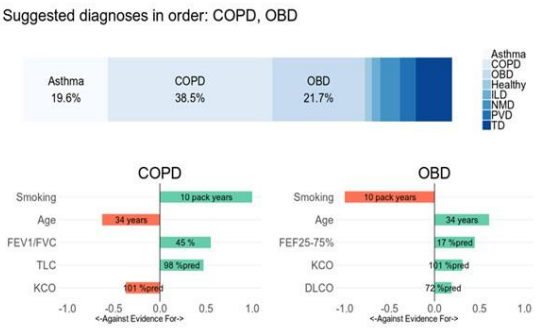
22. Stanojevic S, Kaminsky D, Miller MR, Thompson B, Aliverti A, Barjaktarevic I, Cooper BG, Culver B, Derom E, Hall GL, Hallstrand ST, Leuppi JD, MacIntyre N, McCormack M, Rosenfeld M and Swenson E. ERS/ATS technical standard on interpretive strategies for routine lung function tests *Eur. Respir. J.* 2022; 60(1):2101499.



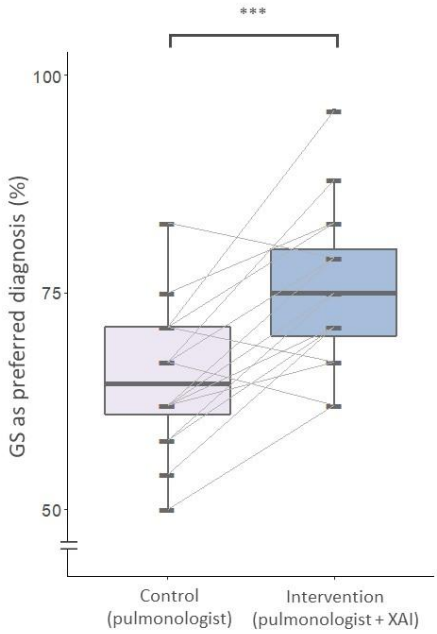
**Sex**: Male **Age**: 34 **Height**: 178 cm **Weight**: 74 kg **BMI**: 23 kg/m² **Race**: Caucasian **Smoking**: 10 PY
**Case**: Male 34yo, heavy-smoker, complaints of dyspnoea, cough and sputum production

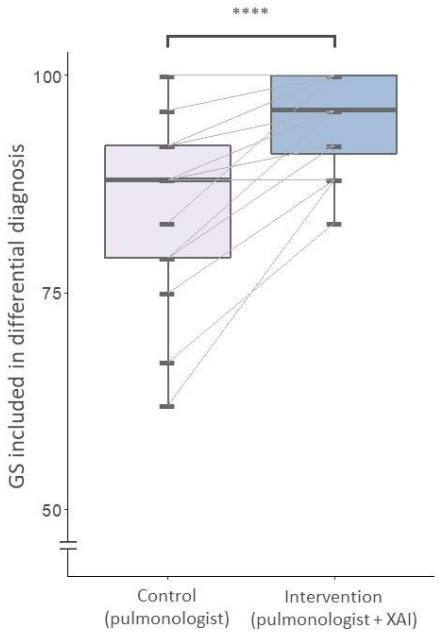
| | | Refer... | Pred | Pre | %Pred | Z-Score | Z-Score₀ | Post Ven... | %Pred | %Chg | Z-Score | Z-Score₀ | Z-Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Substantie** | | | | | | | | | | | | | |
| **Spirometrie** | | | | | | | | | | | | | |
| **Meas time** | | | | 14:41 | | | | 15:29 | | | | | |
| **FVC** | L | Quanj... | 5.35 | 4.47 | 84 | -1.35 | | 4.68 | 88 | 5 | -1.02 | | -1.02 |
| **FEV 1** | L | Quanj... | 4.36 | 1.92 | 44 | -4.27 | | 2.10 | 48 | 9 | -3.99 | | -3.99 |
| **FEV 1 % FVC** | % | Quanj... | 81.80 | 43.03 | 53 | -4.49 | | 44.79 | 55 | 4 | -4.37 | | -4.37 |
| **PEF** | L/s | ECCS... | 9.60 | 4.27 | 45 | -4.40 | | 4.68 | 49 | 10 | -4.06 | | -4.06 |
| **FEF 25** | L/s | ECCS... | 8.25 | 1.76 | 21 | -3.79 | | 2.03 | 25 | 15 | -3.64 | | -3.64 |
| **FEF 50** | L/s | Quanj... | 4.34 | 0.79 | 18 | -4.23 | | 0.95 | 22 | 19 | -3.93 | | -3.93 |
| **FEF 75** | L/s | Quanj... | 1.73 | 0.23 | 14 | -4.41 | | 0.28 | 16 | 18 | -4.08 | | -4.08 |
| **MFEF** | L/s | Quanj... | 4.34 | 0.63 | 15 | -4.56 | | 0.75 | 17 | 18 | -4.32 | | -4.32 |
| **FIF50** | L/s | | | 4.88 | | | | 5.47 | | 12 | | | |
| **FET100** | sec | | | 15.45 | | | | 15.39 | | -0 | | | |
| **Longvolumes Plethysmografisch** | | | | | | | | | | | | | |
| **VC** | L | ECCS... | 5.24 | 4.56 | 87 | -1.21 | | | | | | | |
| **RV** | L | ECCS... | 1.85 | 2.41 | 131 | 1.38 | | | | | | | |
| **ITGV** | L | ECCS... | 3.37 | 4.84 | 143 | 2.44 | | | | | | | |
| **RV%TLC** | % | ECCS... | 27.22 | 34.61 | 127 | 1.35 | | | | | | | |
| **TLC** | L | ECCS... | 7.12 | 6.97 | 98 | -0.21 | | | | | | | |
| **Diffusie** | | | | | | | | | | | | | |
| **DLCO_SB** | mmol/(min*kPa) | ECCS... | 11.47 | 8.28 | 72 | -2.26 | | | | | | | |
| **KCO** | mmol/(min*kPa*L) | ECCS... | 1.61 | 1.62 | 101 | 0.05 | | | | | | | |
| **Hb** | g(Hb)/dL | | | 14.20 | | | | | | | | | |
| **DLCOcSB** | mmol/(min*kPa) | ECCS... | 11.47 | 8.38 | 73 | -2.19 | | | | | | | |
| **KCOc** | mmol/(min*kPa*L) | ECCS... | 1.61 | 1.64 | 102 | 0.13 | | | | | | | |
| **VA_SB** | L | JAEG... | 6.97 | 5.10 | 73 | | | | | | | | |
| **Weerstandsmeting** | | | | | | | | | | | | | |
| **R mid** | kPa/(L/s) | ECCS... | 0.30 | 0.47 | 158 | | | | | | | | |
| **sG mid** | 1/(kPa*s) | ECCS... | 0.85 | 0.41 | 48 | | | | | | | | |

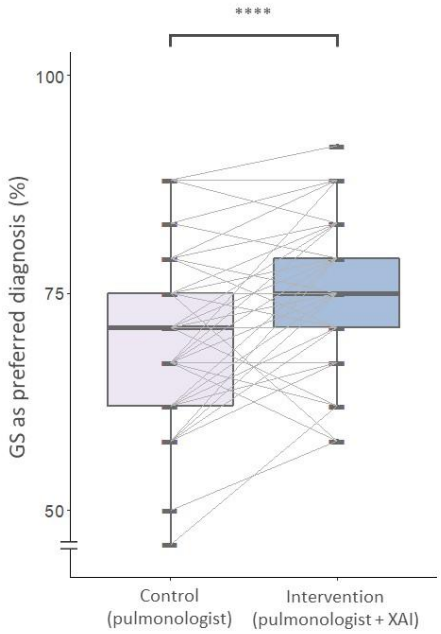
(a)

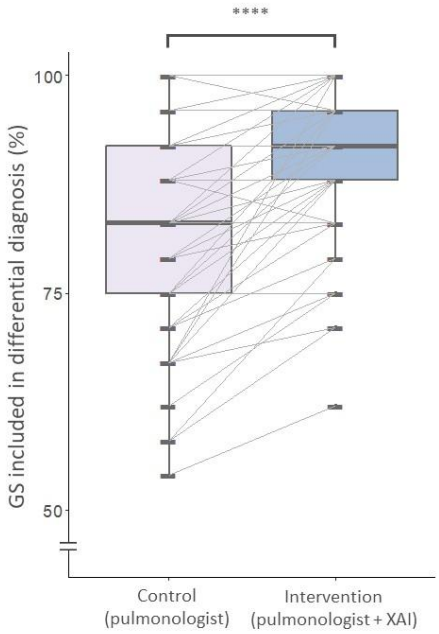



(b)


***
100
75
50
GS as preferred diagnosis (%)
Control
(pulmonologist)
Intervention
(pulmonologist + XAI)


****
100
75
50
GS included in differential diagnosis
Control
(pulmonologist)
Intervention
(pulmonologist + XAI)


****
100
75
50
GS as preferred diagnosis (%)
Control
(pulmonologist)
Intervention
(pulmonologist + XAI)


****
100
75
50
GS included in differential diagnosis (%)
Control
(pulmonologist)
Intervention
(pulmonologist + XAI)

**Explainable artificial intelligence supports pulmonologists in the accurate interpretation of pulmonary function tests.**

Supplemental document

Supplement table S1

Table showing centre wise participation

| **P1 (N=16)** | |
|---|---|
| **Centre** | **N** |
| UZ Leuven | 16 |
| **P2 (N=62)** | |
| **Centre** | **N** |
| AZ Sint-Jan Brugge-Oostende | 4 |
| CHU Charleroi | 2 |
| Cochin Hospital Paris | 4 |
| Hospital Clinic de Barcelona | 2 |
| Imperial College London | 4 |
| LungenClinic Grosshansdorf | 4 |
| Maastricht UMC | 2 |
| National Pirogov Memorial Medical University | 6 |
| Nottingham University Hospitals NHS | 4 |
| Royal Brompton Hospital | 5 |
| University of Ferrara Italy | 2 |
| University of Leicester | 5 |
| UZ Brussels | 2 |
| UZ Gent | 16 |

# Informed Consent

**Title of the study**:
Diagnosing respiratory diseases with artificial intelligence

**Research organisation - Sponsor**:
KU Leuven, Belgium

**Medical Ethics Committee:** University Hospital Leuven, Belgium

**Local investigators**:
Dr. Nilakash Das
Post-doc scientist
Laboratory of respiratory diseases and thoracic surgery
KU Leuven, Belgium
neel.das@kuleuven.be

Sofie Happaerts
Clinical resident
Faculty of medicine
KU Leuven, Belgium
University Hospital Leuven
sofie.happaerts@uzleuven.be

Prof. Dr. Wim Janssens
Principal investigator
Laboratory of respiratory diseases and thoracic surgery
KU Leuven, Belgium
University Hospital Leuven
wim.janssens@uzleuven.be

# I Information vital to your decision to take part

## Introduction

You are invited to participate in a study on how artificial intelligence (AI) and clinicians can collaborate on diagnosing respiratory diseases. In order to help you decide whether or not to take part in this study, please take the time to review the following information for participants so that you can make an informed decision. This is called "informed consent".

We ask you to read the following information carefully. If you have any questions, please contact the researcher.

This page consists of essential information that you need to make your decision, and an option to provide your consent digitally.

This page consists of essential information that you need to make your decision, and an option to provide your consent digitally.

## If you are participating in this study, you should know that:

1. This study was drawn up after evaluation by the Ethical Committee (EC) Research UZ/KU Leuven.
2. Your participation is voluntary; there can be no question of coercion. Your signed consent is required for participation. Even after you have signed, you can let the researcher know that you want to stop your participation without giving any reason.
3. The information collected within the framework of your participation is confidential. Your anonymity is guaranteed when the results are published.
4. If you would like additional information, you can always contact the researchers.

## Objectives and conduct of the study

In this retrospective study, we want to assess **how AI and clinicians can collaborate together in diagnosing respiratory diseases using pulmonary function tests (PFT) reports**. We invite you to participate in this study because you are an expert in the field of the respiratory medicine and experienced in interpreting pulmonary function test reports.

Specifically, we will request you to **perform a series of 24 PFT report interpretations in two parts, first without the aid of AI and then with the aid of AI**. We will record your responses like your preferred diagnoses, and your confidence in diagnosis on a likert scale. Filling out these responses will take around 2-3 minutes of your time for each interpretation. However, there is no time limitation and you can provide all your responses within a week or two.

## Description of the risks and benefits

Your participation in this study **does not present any risk to subjects**. The subjects will be anonymized and your responses will not be used to interfere with their current clinical strategy.

Your participation will be beneficial in understanding the interaction between AI and clinicians.

## Privacy and security

Your **responses will be anonymized** and downloaded on a KU Leuven hard-drive prior to data analysis. Afterwards, they will be deleted from the servers of the online platform. Aggregated data reports will be provided per participating center.

## Publication policy

You will be invited to contribute as co-authors to the manuscript that is written as outcome of this research project in accordance with **ICMJE guidelines**.

## Withdrawal of your consent

It is up to you to decide whether you want to take part in the study. **Participation is voluntary**. If you decide not to participate, you do not need to do anything else. You do not have to sign anything. You also do not have to say why you do not want to participate.

If you do participate, you can always change your mind and still stop, even during the study. You do not have to give a reason for this.

No new data will be collected and that if consent to participate in the study is withdrawn, the coded data already collected before withdrawal will be retained.

## If you wish to participate in this study

We would like to **request you to cooperate fully** to ensure the proper conduct of the study.

## Contact

If you require additional information or have any concerns, or if you encounter any problems, you can contact Dr. Nilakash Das (neel.das@kuleuven.be, +32 484576481) or Prof.Dr.Wim Janssens, (wim.janssens@kuleuven.be, +32 16377265).

# II Informed consent

☐ I declare that I have been informed of the nature of the study, its purpose, its duration, the possible side effects and what is expected of me. I have taken note of the information document and the appendices to this document.

I have had the opportunity to ask any questions that came to mind and have obtained a favorable response to my questions.

I understand that data about me will be collected throughout my participation in this study and that the investigator and the sponsor of the study will guarantee the confidentiality of these data in accordance with applicable European and Belgian legislation. I understand that the performance of this study by UZ Leuven serves the general interest and that the processing of my personal data is necessary for the performance of this study.

I have received a copy of the information to the participant and the informed consent form.

## Investigator

I, **Wim Janssens**, the principal investigator of this study, confirm that no pressure was applied to persuade the participants to agree to take part in the study and that I am willing to answer any additional questions if required.

I confirm that I operate in accordance with the ethical principles set out in the latest version of the "Helsinki Declaration", the "Good Clinical Practices" and the Belgian Law of 7 May 2004 related to experiments on humans.

# III Supplementary information

## Supplementary information on the organization of the study

The study involves retrospective evaluation of PFT cases using an online platform

## Supplementary information on the risks associated with participation in the study

Not applicable

## Supplementary information on the protection and rights of the participant in a clinical study

**Ethics Committee**
This study has been reviewed by an independent Ethics Committee, namely the Ethics Committee of UZ Leuven. It is the task of the Ethics Committees to protect people who take part in a clinical trial. They make sure that your rights as a patient and as a participant in a clinical study are respected, that based on current knowledge, the study is scientifically relevant and ethical.

You should not under any circumstances take the favorable opinion of the Ethics Committee as an incentive to take part in this study.

**Voluntary participation**
Before signing, do not hesitate to ask any questions you feel are appropriate. Take the time to discuss matters with a trusted person if you so wish.

Your participation in the study is voluntary and must remain free of any coercion: this means that you have the right not to take part in the study or to withdraw without giving a reason, even if you previously agreed to take part. Your decision will not affect your relationship with the investigator or the quality of your future therapeutic care.

If you agree to take part in this study, you will sign the informed consent form. The investigator will also sign this form to confirm that he/she has provided you with the necessary information about the study. You will receive a copy of the form.

**Costs associated with your participation**
You will not receive any compensation for your participation in this study. Furthermore, the study will not involve any additional costs for you.

**Guarantee of confidentiality**
Your participation in the study means that you agree to the investigator collecting data about you and to the study sponsor using these data for research purposes and in connection with scientific and medical publications.

The processing of your personal data is necessary to achieve the scientific research purposes as set out herein. The conduct of scientific research is one of the core missions of UZ Leuven as defined by law. As a university hospital, part of KU Leuven, UZ Leuven is indeed required to support research and education in the public interest. We would therefore like to inform you that the necessity of the processing for the conduct of scientific research as a task of public interest constitutes the lawful basis on which we process your information in the context of the study in which you are participating. UZ Leuven is also subject to specific legal requirements which require the processing of your personal in the context of safety reporting (such as for example the notification of adverse events to the regulatory authorities).

Your data will be processed in accordance with the European General Data Protection Regulation (GDPR) and Belgian framework law. The sponsor UZ Leuven is responsible for the data collection with Data protection officer (DPO) contact: DPO - UZ Leuven, Herestraat 49, 3000 Leuven, Belgium e-mail: dpo@uzleuven.be. Data will be kept secured for a minimal period of 20 years.

You are entitled to ask the investigator what data are being collected about you and what is their use in connection with the study. This data concerns your current clinical situation but also some of your background, the results of examinations carried out within the context of care of your health in accordance with current standards. You have the right to inspect these data and correct them if they are incorrect.

The investigator has a duty of confidentiality vis-à-vis the data collected. This means that he/she undertakes not only never to reveal your name in the context of a publication or conference but also that he/she will encode your data before sending them to the manager of the database of collected data (Laboratory of respiratory diseases and thoracic surgey, CHROMETA department, KU Leuven).

The investigator and his team will therefore be the only ones to be able to establish a link between the data transmitted throughout the study and your medical records.

The personal data transmitted will not contain any combination of elements that might despite everything allow you to be identified.
For the study data manager designated by the sponsor, the data transmitted will not allow you to be identified. The latter is responsible for collecting the data gathered by all investigators taking part in the study, processing them and protecting them in accordance with the requirements of the Belgian law on the protection of privacy.

These (encoded) data will be able to be sent to Belgian or other regulatory authorities, to the relevant ethics committees, to other doctors and/or to organisations working in collaboration with the sponsor.

The sponsor will use the data collected within the context of the study in which you are taking part, but would also like to be able to use them in connection with other research concerning the same disease as yours and its treatment. Any use of your data outside the context described in this document is only possible with the approval of the ethics committee.

If you withdraw your consent to take part in the study, to guarantee the validity of the research, the data encoded up to the point at which you withdraw will be retained. No new data may be sent to the sponsor.

If you have any questions relating to how your data are being processed, you may contact the investigator. The data protection officer in your hospital can be contacted as well: DPO - UZ Leuven, Herestraat 49, 3000 Leuven, e-mail dpo@uzleuven.be.

Finally, if you have a complaint concerning the processing of your data, you can contact the Belgian supervisory authority who ensures that privacy is respected when personal data are processed.

The Belgian supervisory authority is called:
Data Protection Authority (DPA)
Drukpersstraat 35,

1000 Brussels, Belgium
Tel. +32 2 274 48 00
Email: contact@apd-gba.be
Website: https://www.dataprotectionauthority.be

Next

# About you

Please answer the following questions about you

Your years of experience as a physician

Please Select...

Have you worked with AI-based decision support systems before?

- Yes
- No

On a scale from 1 (least) to 5 (most), are you enthusiastic about AI for clinical outcomes in general?

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

Next

## Supplement S3: Tutorial task for the participants

### Tutorial task 1a

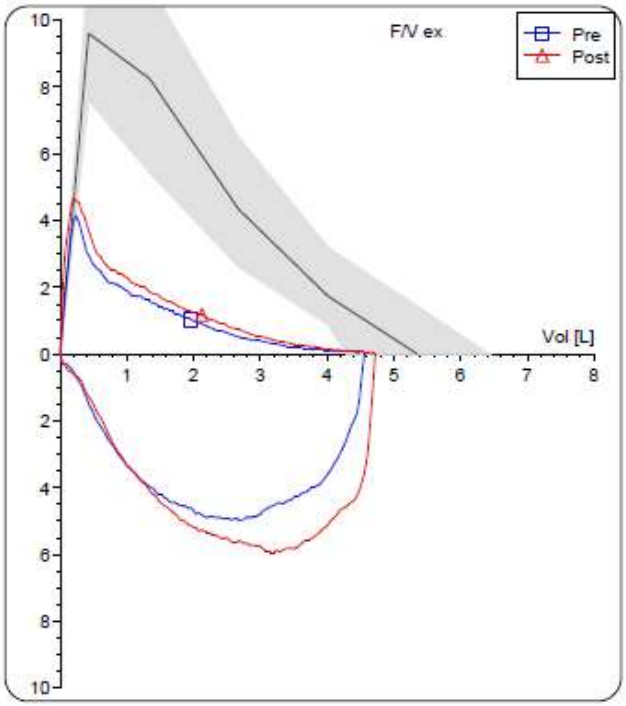
**Instruction** The following is a PFT report with clinical characteristics, flow-volume and volume-time curves from spirometry, resistance curves from body-plethysmography, volume-time and CO concentration-time curves from diffusion capacity test.

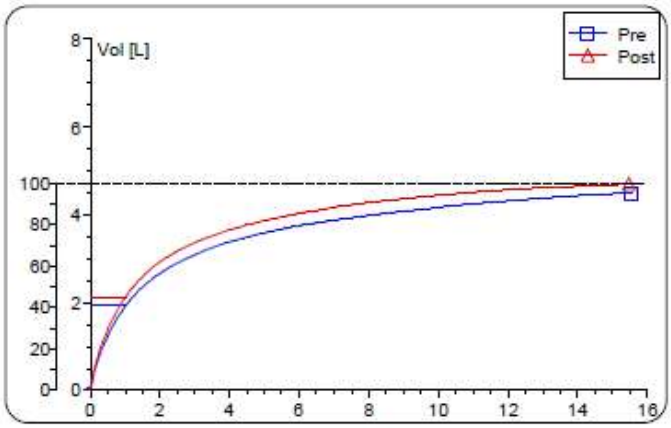
Please scroll down to view the entire PFT report. At the end, we will ask your responses.

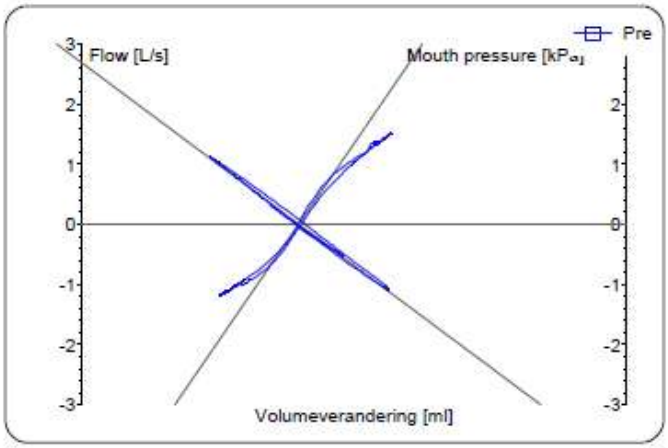
To improve visibility, press *Ctrl and scroll* to zoom in or out.

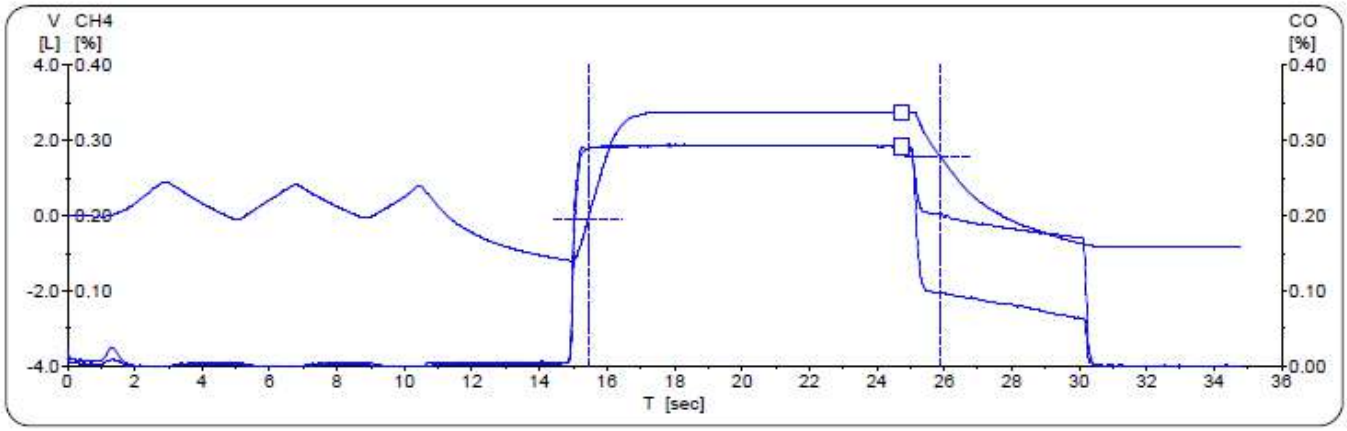
**Sex**: Male **Age**: 34 **Height**: 178 cm **Weight**: 74 kg **BMI**: 23 kg/m² **Race**: Caucasian **Smoking**: 10 PY
**Case**: Male 34yo, heavy-smoker, complaints of dyspnoea, cough and sputum production

| | | Refer... | Pred | Pre | %Pred | Z-Score | Z-Score | Post | %Pred | %Chg | Z-Score | Z-Score | Z-Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Substantie** | | | | | | | | Ven... | | | | | |
| **Spirometrie** | | | | | | | | | | | | | |
| **Meas time** | | | | 14:41 | | | | 15:29 | | | | | |
| **FVC** | L | Quanj... | 5.35 | 4.47 | 84 | -1.35 | | 4.68 | 88 | 5 | -1.02 | | -1.02 |
| **FEV 1** | L | Quanj... | 4.36 | 1.92 | 44 | -4.27 | | 2.10 | 48 | 9 | -3.99 | | -3.99 |
| **FEV 1 % FVC** | % | Quanj... | 81.80 | 43.03 | 53 | -4.49 | | 44.79 | 55 | 4 | -4.37 | | -4.37 |
| **PEF** | L/s | ECCS... | 9.60 | 4.27 | 45 | -4.40 | | 4.68 | 49 | 10 | -4.06 | | -4.06 |
| **FEF 25** | L/s | ECCS... | 8.25 | 1.76 | 21 | -3.79 | | 2.03 | 25 | 15 | -3.64 | | -3.64 |
| **FEF 50** | L/s | Quanj... | 4.34 | 0.79 | 18 | -4.23 | | 0.95 | 22 | 19 | -3.93 | | -3.93 |
| **FEF 75** | L/s | Quanj... | 1.73 | 0.23 | 14 | -4.41 | | 0.28 | 16 | 18 | -4.08 | | -4.08 |
| **MFEF** | L/s | Quanj... | 4.34 | 0.63 | 15 | -4.56 | | 0.75 | 17 | 18 | -4.32 | | -4.32 |
| **FIF50** | L/s | | | 4.88 | | | | 5.47 | | 12 | | | |
| **FET100** | sec | | | 15.45 | | | | 15.39 | | -0 | | | |
| **Longvolumes Plethysmografisch** | | | | | | | | | | | | | |
| **VC** | L | ECCS... | 5.24 | 4.56 | 87 | -1.21 | | | | | | | |
| **RV** | L | ECCS... | 1.85 | 2.41 | 131 | 1.38 | | | | | | | |
| **ITGV** | L | ECCS... | 3.37 | 4.84 | 143 | 2.44 | | | | | | | |
| **RV%TLC** | % | ECCS... | 27.22 | 34.61 | 127 | 1.35 | | | | | | | |
| **TLC** | L | ECCS... | 7.12 | 6.97 | 98 | -0.21 | | | | | | | |
| **Diffusie** | | | | | | | | | | | | | |
| **DLCO_SB** | mmol/(min*kPa) | ECCS... | 11.47 | 8.28 | 72 | -2.26 | | | | | | | |
| **KCO** | mmol/(min*kPa*L) | ECCS... | 1.61 | 1.62 | 101 | 0.05 | | | | | | | |
| **Hb** | g(Hb)/dL | | | 14.20 | | | | | | | | | |
| **DLCOcSB** | mmol/(min*kPa) | ECCS... | 11.47 | 8.38 | 73 | -2.19 | | | | | | | |
| **KCOc** | mmol/(min*kPa*L) | ECCS... | 1.61 | 1.64 | 102 | 0.13 | | | | | | | |
| **VA_SB** | L | JAEG... | 6.97 | 5.10 | 73 | | | | | | | | |
| **Weerstandsmeting** | | | | | | | | | | | | | |
| **R mid** | kPa/(L/s) | ECCS... | 0.30 | 0.47 | 158 | | | | | | | | |
| **sG mid** | 1/(kPa*s) | ECCS... | 0.85 | 0.41 | 48 | | | | | | | | |




Your preferred diagnosis?

Please Select...

Second diagnosis? (Optional)

None

Third diagnosis? (Optional)

None

Fourth diagnosis? (Optional)

None

Your overall confidence in diagnosis from 1(least) to 5 (most)?

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

Any comments? (Optional)

**Instruction** Please click Next to proceed to tutorial task 1b

Back Next

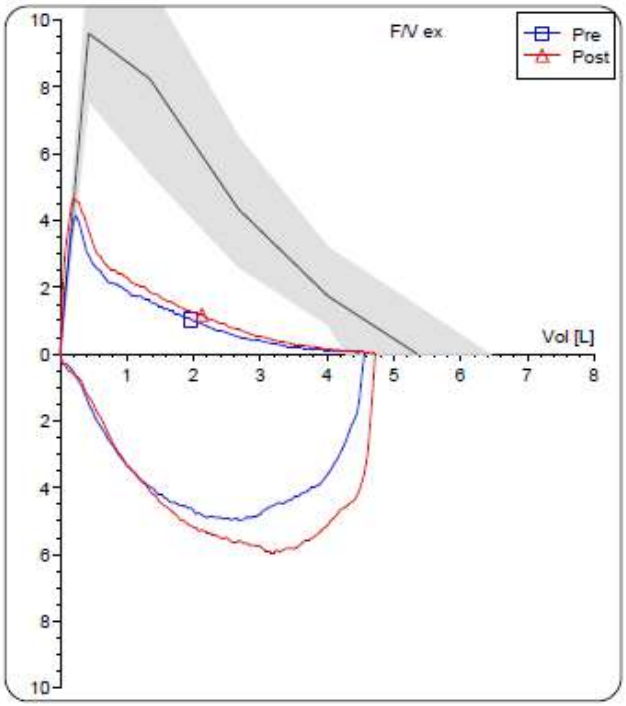
**Tutorial task 1b**

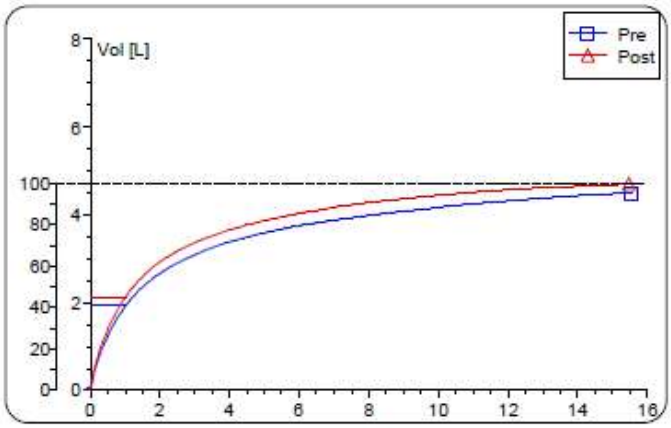
**Instruction** The same PFT report from the task a is displayed here, but now the suggestions of AI are included at the end of the report.

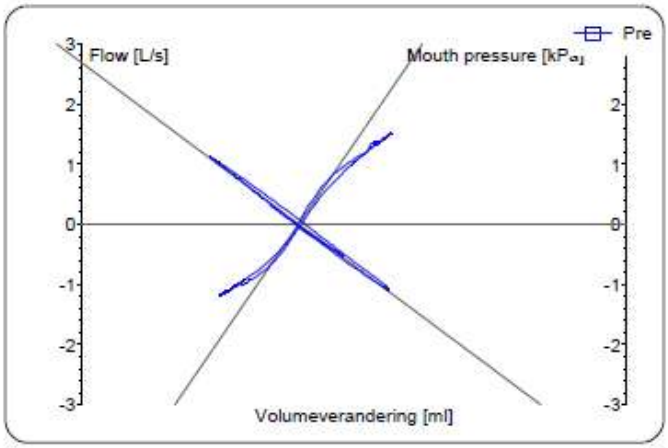
Please scroll down to view the PFT report and AI's suggestions

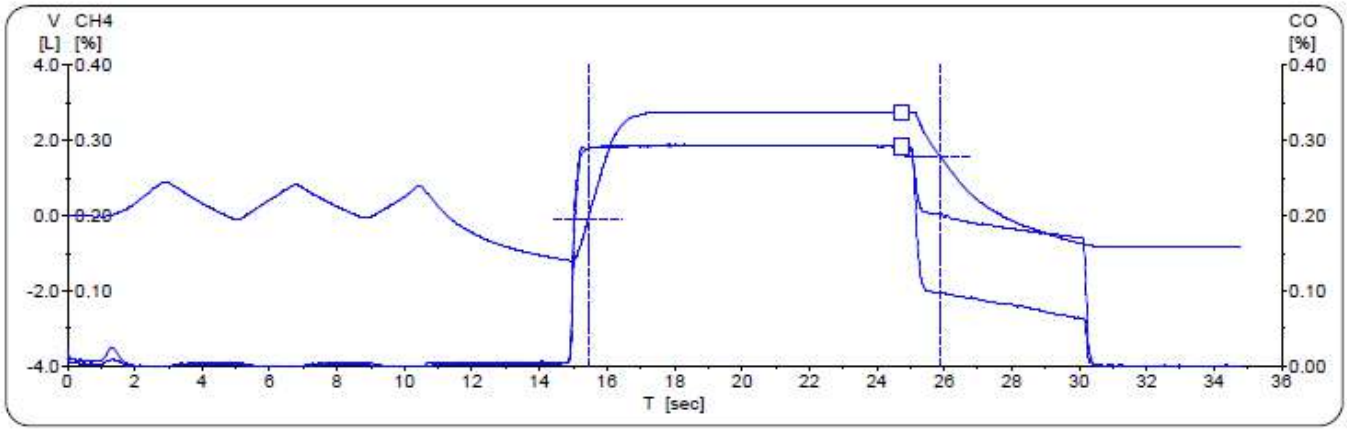
**Sex**: Male **Age**: 34 **Height**: 178 cm **Weight**: 74 kg **BMI**: 23 kg/m² **Race**: Caucasian **Smoking**: 10 PY
**Case**: Male 34yo, heavy-smoker, complaints of dyspnoea, cough and sputum production

| | | Refer... | Pred | Pre | %Pred | Z-Score | Z-Score | Post | %Pred | %Chg | Z-Score | Z-Score | Z-Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Substantie** | | | | | | | | Ven... | | | | | |
| **Spirometrie** | | | | | | | | | | | | | |
| **Meas time** | | | | 14:41 | | | | 15:29 | | | | | |
| **FVC** | L | Quanj... | 5.35 | 4.47 | 84 | -1.35 | | 4.68 | 88 | 5 | -1.02 | | -1.02 |
| **FEV 1** | L | Quanj... | 4.36 | 1.92 | 44 | -4.27 | | 2.10 | 48 | 9 | -3.99 | | -3.99 |
| **FEV 1 % FVC** | % | Quanj... | 81.80 | 43.03 | 53 | -4.49 | | 44.79 | 55 | 4 | -4.37 | | -4.37 |
| **PEF** | L/s | ECCS... | 9.60 | 4.27 | 45 | -4.40 | | 4.68 | 49 | 10 | -4.06 | | -4.06 |
| **FEF 25** | L/s | ECCS... | 8.25 | 1.76 | 21 | -3.79 | | 2.03 | 25 | 15 | -3.64 | | -3.64 |
| **FEF 50** | L/s | Quanj... | 4.34 | 0.79 | 18 | -4.23 | | 0.95 | 22 | 19 | -3.93 | | -3.93 |
| **FEF 75** | L/s | Quanj... | 1.73 | 0.23 | 14 | -4.41 | | 0.28 | 16 | 18 | -4.08 | | -4.08 |
| **MFEF** | L/s | Quanj... | 4.34 | 0.63 | 15 | -4.56 | | 0.75 | 17 | 18 | -4.32 | | -4.32 |
| **FIF50** | L/s | | | 4.88 | | | | 5.47 | | 12 | | | |
| **FET100** | sec | | | 15.45 | | | | 15.39 | | -0 | | | |
| **Longvolumes Plethysmografisch** | | | | | | | | | | | | | |
| **VC** | L | ECCS... | 5.24 | 4.56 | 87 | -1.21 | | | | | | | |
| **RV** | L | ECCS... | 1.85 | 2.41 | 131 | 1.38 | | | | | | | |
| **ITGV** | L | ECCS... | 3.37 | 4.84 | 143 | 2.44 | | | | | | | |
| **RV%TLC** | % | ECCS... | 27.22 | 34.61 | 127 | 1.35 | | | | | | | |
| **TLC** | L | ECCS... | 7.12 | 6.97 | 98 | -0.21 | | | | | | | |
| **Diffusie** | | | | | | | | | | | | | |
| **DLCO_SB** | mmol/(min*kPa) | ECCS... | 11.47 | 8.28 | 72 | -2.26 | | | | | | | |
| **KCO** | mmol/(min*kPa*L) | ECCS... | 1.61 | 1.62 | 101 | 0.05 | | | | | | | |
| **Hb** | g(Hb)/dL | | | 14.20 | | | | | | | | | |
| **DLCOcSB** | mmol/(min*kPa) | ECCS... | 11.47 | 8.38 | 73 | -2.19 | | | | | | | |
| **KCOc** | mmol/(min*kPa*L) | ECCS... | 1.61 | 1.64 | 102 | 0.13 | | | | | | | |
| **VA_SB** | L | JAEG... | 6.97 | 5.10 | 73 | | | | | | | | |
| **Weerstandsmeting** | | | | | | | | | | | | | |
| **R mid** | kPa/(L/s) | ECCS... | 0.30 | 0.47 | 158 | | | | | | | | |
| **sG mid** | 1/(kPa*s) | ECCS... | 0.85 | 0.41 | 48 | | | | | | | | |



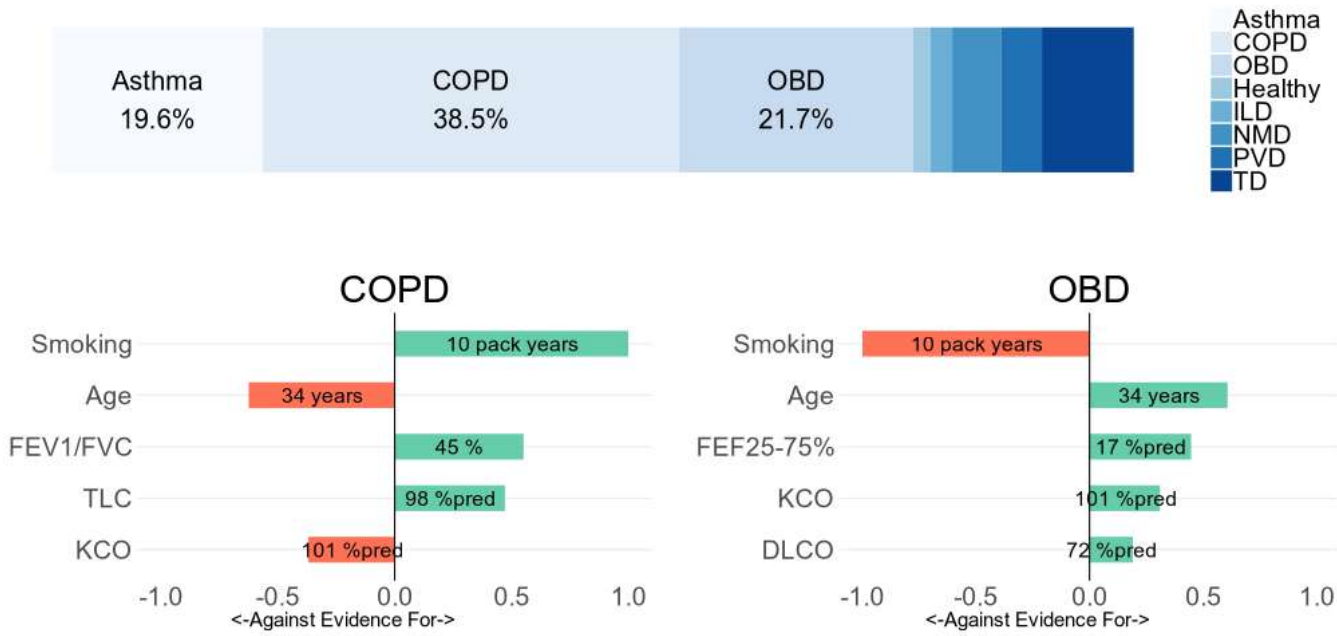




**Interpretation by AI**

## Suggested diagnoses in order: COPD, OBD




**Instruction** Key points to note

1. AI's disease suggestions are based on a descending order of predicted probabilities (bar plot) , and the suggestions are limited to two diseases.
2. The evidence plots show the relative evidence of the top 5 PFT report parameters towards AI's suggestions. The parameter with the highest evidence has a magnitude of 1.
3. Evidence can be **positive** implying supporting evidence, or **negative** implying counter-evidence

**Abbreviations**
COPD: Chronic obstructive pulmonary disease
OBD: Other obstructive disease
ILD: Interstitial lung disease
NMD: Neuromuscular disease
PVD: Pulmonary vascular disease
TD: Thoracic deformity

Your preferred diagnosis?

Please Select...

Second diagnosis? (Optional)

None

Third diagnosis? (Optional)

None

Fourth diagnosis? (Optional)

None

Your overall confidence in your diagnosis from 1(least) to 5 (most)?

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

Do you agree with the suggestions and the evidence provided by AI?

| Strongly disagree | Disagree | Neutral | Agree | Strongly agree |
|---|---|---|---|---|

Any comments? (Optional)

**Instruction** Please click Next to conclude the tutorial

Back Next

Supplement Table S4

Participant demographics in phase P1 and P2 studies

| | **P1** | **P2** |
|---|---|---|
| **N** | 16 | 62 |
| **Enrolment** | Monocentric | Multicentric |
| **Years of experience > 5 years** | 75% | 81% |
| **Past experience with AI** | 56% | 11% |
| **Mean baseline enthusiasm in AI on Likert Scale (1-least, 5-most)*** | 3.56 (0.96) | 3.92 (0.93) |

Supplement Table S5

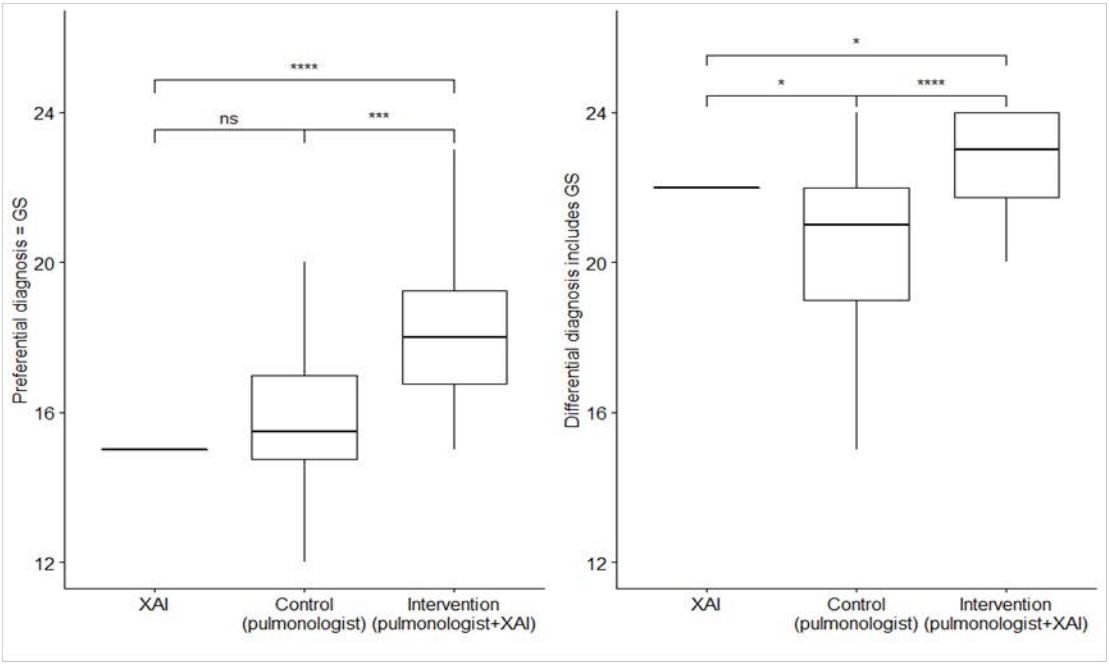
Preferential and differential diagnostic performance of the explainable artificial intelligence (XAI) model across different disease cohorts.

| | P1 study | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Overall** | **Healthy** | **COPD** | **Asthma** | **ILD** | **NMD** | **OBD** | **TD** | **PVD** |
| **N** | 24 | 4 | 4 | 4 | 4 | 2 | 2 | 2 | 2 |
| **Preferential diagnosis (disease with maximum probability) = GS** | 15 | 2 | 4 | 2 | 4 | 2 | 1 | 0 | 0 |
| **Differential diagnosis (preferential + second suggestion if probability > 15%) includes GS** | 22 | 4 | 4 | 4 | 4 | 2 | 2 | 1 | 1 |

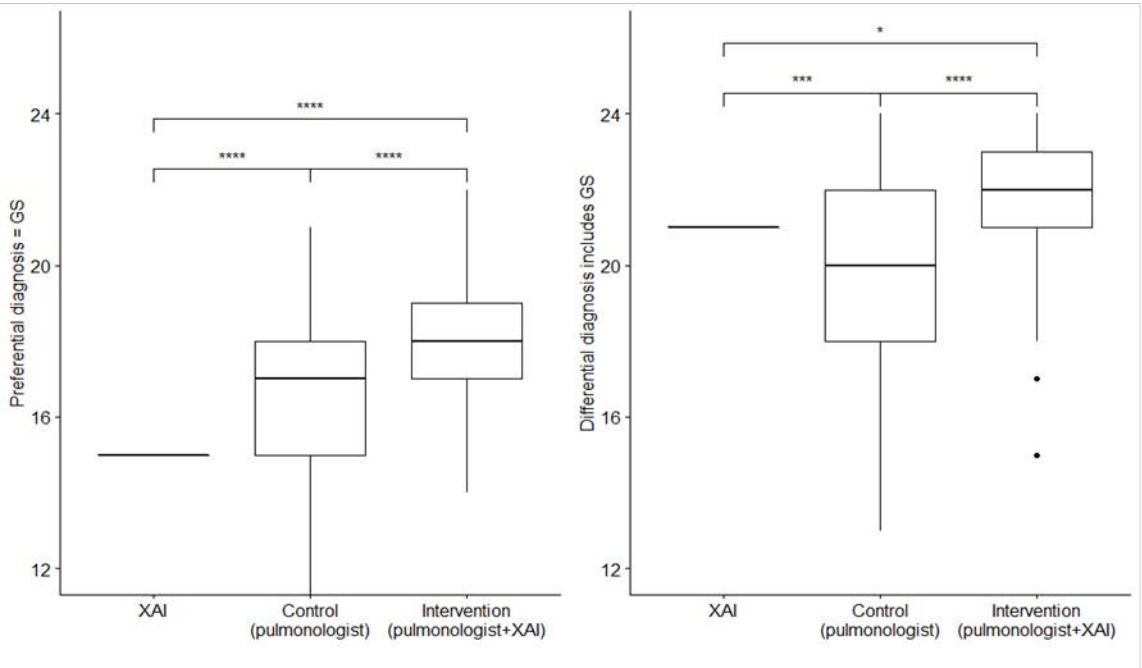
| | P2 study | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Overall** | **Healthy** | **COPD** | **Asthma** | **ILD** | **NMD** | **OBD** | **TD** | **PVD** |
| **N** | 24 | 4 | 4 | 4 | 4 | 2 | 2 | 2 | 2 |
| **Preferential diagnosis (disease with maximum probability) = GS** | 15 | 2 | 3 | 3 | 4 | 2 | 1 | 0 | 0 |
| **Differential diagnosis (preferential + second suggestion if probability > 15%) includes GS** | 21 | 4 | 4 | 4 | 4 | 2 | 1 | 1 | 1 |

**Abbreviations**: GS= Gold standard; NMD = neuromuscular disease; ILD = interstitial lung diseases; PVD = pulmonary vascular diseases; OBD = other obstructive diseases; TD = Thoracic deformity/ Pleural diseases; COPD = chronic obstructive pulmonary disease

(a) P1 with 16 pulmonologists




(b) P2 with 62 pulmonologists



Supplement S7

Table : Interventional diagnostic performance stratified on years of experience and baseline enthusiasm in AI applications in P2 study (N=62 pulmonologists)

| **Diagnostic performance based on years of experience (pulmonologist + XAI, N = 62)** | | | |
|---|---|---|---|
| | **< 5 years (N=12)** | **> 5 years (N=50)** | **p** |
| **Preferential diagnosis= GS** | 18.42 (1.93) | 17.82 (1.79) | 0.34 |
| **Differential diagnosis includes GS** | 22 (1.81) | 21.62 (2.18) | 0.54 |
| **Diagnostic performance based on baseline enthusiasm in AI applications measured on Likert Scale (LS) (pulmonologist + XAI, N=62)** | | | |
| | **LS > 3 (N=42)** | **LS <= 3 (N=20)** | **p** |
| **Preferential diagnosis= GS** | 17.69 (1.83) | 18.45 (1.73) | 0.12 |
| **Differential diagnosis includes GS** | 21.23 (2.31) | 22.65 (1.14) | 0.06 |

Supplement S8

Table showing how pulmonologists' diagnostic performance whenever XAI's preferential diagnosis was incorrect.

| | P1 study (16 pulmonologists) | | | | | |
|---|---|---|---|---|---|---|
| | **XAI's preferential diagnosis was incorrect (N=9)** | | | **XAI's preferential diagnosis was correct (N=15)** | | |
| | **Control (pulmonologist)** | **Intervention (pulmonologist + XAI)** | **p** | **Control (pulmonologist)** | **Intervention (pulmonologist + XAI)** | **p** |
| **Preferential diagnosis= GS** | 5.94 (1.44) | 5.5 (1.55) | 0.032 | 9.69 (1.08) | 12.62 (1.89) | <0.0001 |
| **Mean level of agreement with XAI on Likert scale** | 3.47 (0.43) | | | 4.07 (0.13) (p<0.0001) | | |

| | P2 study (62 pulmonologists) | | | | | |
|---|---|---|---|---|---|---|
| | **XAI's preferential diagnosis was incorrect (N=9)** | | | **XAI's preferential diagnosis was correct (N=15)** | | |
| | **Control (pulmonologist)** | **Intervention (pulmonologist + XAI)** | **p** | **Control (pulmonologist)** | **Intervention (pulmonologist + XAI)** | **p** |
| **Preferential diagnosis= GS** | 6.45 (1.29) | 5.47 (1.64) | <0.0001 | 10.19 (1.49) | 12.47 (1.26) | <0.0001 |
| **Mean level of agreement with XAI on Likert Scale** | 2.95 (0.52) | | | 3.75 (0.06) (p<0.00001) | | |

**Abbreviations**: GS= Gold standard; XAI: Explainable AI