Deep-learning algorithm helps to standardise ATS/ERS spirometric acceptability and usability criteria

Nilakash Das[1], Kenneth Verstraete [1], Sanja Stanojevic[2], Marko Topalovic[1,3], Jean-Marie Aerts[4] and Wim Janssens [1]

Affiliations: [1]Laboratory of Respiratory Diseases and Thoracic Surgery, Dept of Chronic Diseases, Metabolism and Ageing, Katholieke Universiteit Leuven, Leuven, Belgium. [2]Translational Medicine, Division of Respiratory Medicine, Hospital for Sick Children, Toronto, ON, Canada. [3]ArtiQ NV, Leuven, Belgium. [4]Division Animal and Human Health Engineering, Dept of Biosystems, Katholieke Universiteit Leuven, Leuven, Belgium.

Correspondence: Wim Janssens, O&N1 Herestraat 49 bus 706, 3000 Leuven, Belgium.
E-mail: wim.janssens@uzleuven.be

@ERSpublications
Deep-learning models were developed to standardise ATS/ERS spirometric acceptability and usability criteria. This approach reduces the intertechnician variability and provides instant feedback to the user https://bit.ly/3dNFe1i

Cite this article as: Das N, Verstraete K, Stanojevic S, *et al.* Deep-learning algorithm helps to standardise ATS/ERS spirometric acceptability and usability criteria. *Eur Respir J* 2020; 56: 2000603 [https://doi.org/10.1183/13993003.00603-2020].

ABSTRACT

Rationale: While American Thoracic Society (ATS)/European Respiratory Society (ERS) quality control criteria for spirometry include several quantitative limits, it also requires manual visual inspection. The current approach is time consuming and leads to high intertechnician variability. We propose a deep-learning approach called convolutional neural network (CNN), to standardise spirometric manoeuvre acceptability and usability.

Methods and methods: In 36 873 curves from the National Health and Nutritional Examination Survey USA 2011–2012, technicians labelled 54% of curves as meeting ATS/ERS 2005 acceptability criteria with satisfactory start and end of test, but identified 93% of curves with a usable forced expiratory volume in 1 s. We processed raw data into images of maximal expiratory flow–volume curve (MEFVC), calculated ATS/ERS quantifiable criteria and developed CNNs to determine manoeuvre acceptability and usability on 90% of the curves. The models were tested on the remaining 10% of curves. We calculated Shapley values to interpret the models.

Results: In the test set (n=3738), CNN showed an accuracy of 87% for acceptability and 92% for usability, with the latter demonstrating a high sensitivity (92%) and specificity (96%). They were significantly superior ($p<0.0001$) to ATS/ERS quantifiable rule-based models. Shapley interpretation revealed MEFVC<1 s (MEFVC pattern within first second of exhalation) and plateau in volume–time were most important in determining acceptability, while MEFVC<1 s entirely determined usability.

Conclusion: The CNNs identified relevant attributes in spirometric curves to standardise ATS/ERS manoeuvre acceptability and usability recommendations, and further provides individual manoeuvre feedback. Our algorithm combines the visual experience of skilled technicians and ATS/ERS quantitative rules in automating the critical phase of spirometry quality control.

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

Received: 9 March 2020 | Accepted after revision: 3 June 2020

Introduction

The validity of spirometric indices such as forced expiratory volume in 1 s (FEV_1) and forced vital capacity (FVC) depends on the quality of forced expiratory manoeuvre. The American Thoracic Society (ATS) and European Respiratory Society (ERS) have published joint guidelines that describe the within-manoeuvre acceptability criteria [1]. Furthermore, ATS/ERS provide recommendations on the usability of a manoeuvre when it does not meet a satisfactory end of test (EOT), but still contains a valid FEV_1.

The evaluation of forced expiratory manoeuvre acceptability involves quantitative criteria, namely back-extrapolated volume (BEV), time of forced expiration (t_{FE}) and existence of a plateau (EOP) in a volume–time curve. Additionally, the current guidelines include descriptive protocols that require a technician to visually evaluate if a manoeuvre is free from artefacts such as cough in first second, glottis closure, early termination, nonmaximal effort, obstructed mouthpiece, *etc.* [1, 2]. For usability (*i.e.* whether an FEV_1 can be reported), the criteria are based on an acceptable BEV and the technician's judgement to interpret whether the measurement of FEV_1 was accurate [1, 3]. Based on these current guidelines, whether or not a spirometry manoeuvre is acceptable or usable can vary according to the training and experience of the technician. In some cases, agreement between technicians can be as low as 52% for an individual manoeuvre [4]. Differences in overall test quality in large population studies have been attributed to the complexity of acceptability criteria and professional background of the reviewers [5]. Moreover, spirometry manoeuvre quality can differ widely between the settings of epidemiological studies with standardised protocols on spirometry training [6, 7], specialised pulmonary function laboratories and primary care practice [8].

Advances in deep-learning techniques may provide a tool to automate spirometry manoeuvre acceptability and usability. Ideally, an algorithm should provide holistic feedback on manoeuvre acceptability in addition to standard BEV, t_{FE} and EOP criteria. In the past, researchers have proposed machine-learning models that utilise manually selected features from maximal expiratory flow–volume curve (MEFVC) to determine acceptability [4, 9]. However, the features employed were not only difficult to interpret, but also fail to capture all possible artefacts, making them less robust for application in practice. Deep-learning models such as convolutional neural networks (CNN) have become the dominant method in computer vision tasks, and their application in medical imaging has produced astonishing results [10]. A CNN learns representations directly from data for the required task [11], and thus is superior to prior feature-driven approaches. Since spirometry review is accomplished by visual inspection of the MEFVC and volume–time curves [2], we hypothesise that application of a CNN is ideal for automation of acceptability and usability criteria.

In this study, we aim to develop a CNN to determine spirometry manoeuvre ATS/ERS acceptability and usability using data from the National Health and Nutritional Examination Survey (NHANES) USA 2011–2012 [12]. In addition, we describe methodology that allows interpretation of the model to provide feedback to the user.

Materials and methods

Study subjects

NHANES 2011–2012 spirometry component

Spirometry data were collected in NHANES 2011–2012. Eligible participants aged 6–79 years performed baseline spirometry; a subset who met additional screening criteria performed post-bronchodilator spirometry [12].

Quality control

Technicians who completed the National Institute for Occupational Safety and Health (NIOSH)-approved spirometry course performed regular quality checks of equipment and spirometry measurements. A chief health technician on site supervised the performance of the technicians, while experts at NIOSH quality control reviewed all spirometry data on an ongoing basis. The goal of the technicians was to attain three acceptable and two reproducible curves using ATS/ERS 2005 recommendations [1]. More details on quality control measures can be referred in the NHANES spirometry procedures manual [13].

Data

The NHANES 2011–2012 dataset contained 6696 pre- and 458 post-bronchodilator tests, resulting in 37 661 trials. Only 2% of trials did not have curve data. For the remaining tests the raw data were available as a sequence of numeric values representing change in volume in millilitres over a 0.01-s interval during forced expiration [14]. Technicians had assigned an effort-rating variable “SPAQEFF” to individual curves as follows. A: the curve quality attributes were acceptable (n=19 828); B: the curve had a large time to peak flow or a nonrepeatable peak flow (n=5890); C: the curve had either <6 s of exhalation or no plateau (n=8605); D: the curve had either a cough or large extrapolated volume (n=2550). Furthermore, a variable "SPAACC" indicated that curves with ratings A, B and C (n=34 323) were used for obtaining FEV_1 or FVC. Only curves with rating D were rejected in NHANES 2011–2012.

Study design

In this retrospective analysis of the NHANES 2011–2012 dataset, we aimed to develop, validate and interpret a deep-learning model that determines acceptability and usability of spirometry manoeuvres. Curves with an effort-rating of A were deemed as the gold standard for ATS/ERS acceptability criteria (*i.e.* satisfactory start and end of test). Curves with effort-ratings A, B or C were considered as the gold standard for usable manoeuvers (*i.e.* reportable FEV_1).

Model development

Processing of input

First, we processed the raw data into time, flow and volume measurements, which began at the point of maximum inspiration until residual volume. Then, we generated 32×32 pixel matrices of MEFVC with an aspect ratio of two units of flow for each unit of volume in accordance with the display guidelines of ATS/ERS for MEFVC (supplement S1) [1]. Figure 1 shows an example of an MEFVC pixel matrix where the white spaces and black lines denote a pixel value of one and zero, respectively. In addition, we defined time zero (t_0) by the ATS/ERS back-extrapolation technique to calculate BEV (extrapolated volume $<$ maximum of 5% of FVC or 0.15 L) and t_{FE} (difference between time at residual volume and t_0) >6 s criteria. We defined EOP as no change in volume ($\leqslant$0.025 L) in volume–time curve for $\geqslant$1 s of expiration [1]. Finally, we calculated time to peak expiratory flow (t_{PEF}) from t_0. All input processing was done using Python scripts [15].

Development and test data

We split NHANES spirometry sessions (n=7154) randomly into 80% for training (5726 sessions or 29 452 curves), 10% for validation (711 sessions or 3683 curves) and 10% for testing (710 sessions or 3738 curves).

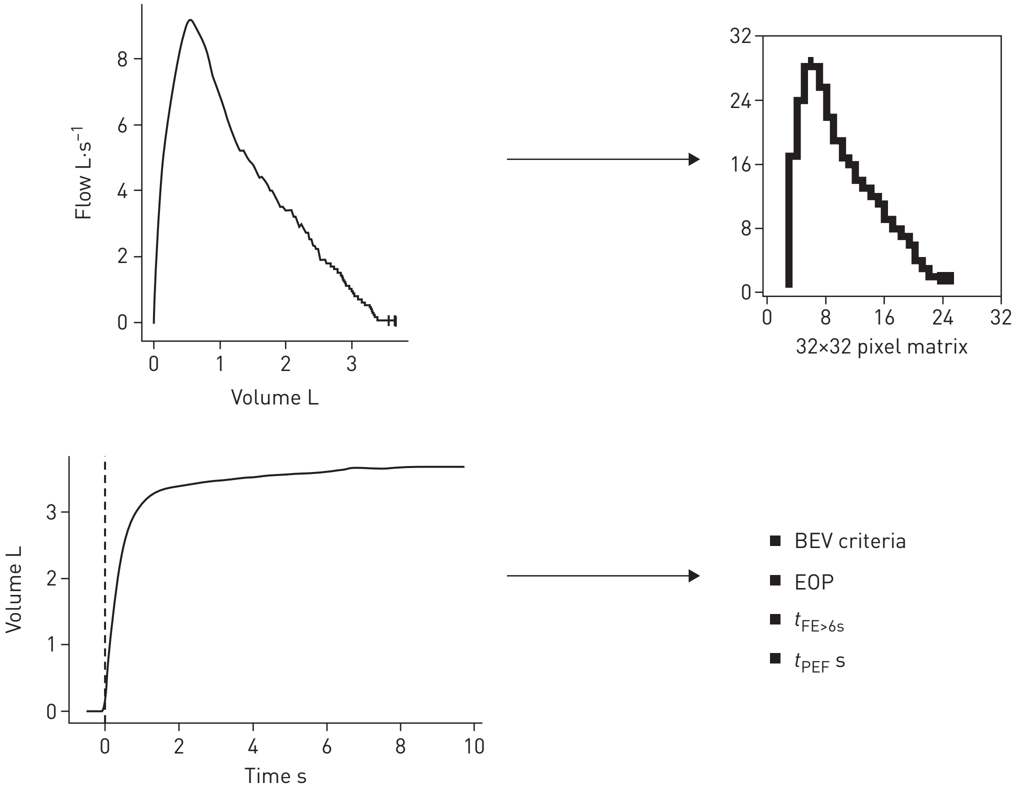

FIGURE 1 Processing of input. Flow–volume data are converted into 32×32-pixel matrices of maximal expiratory flow–volume loop (supplementary material). The white and the black regions represent a pixel value of 1 and 0, respectively. Furthermore, American Thoracic Society/European Respiratory Society recommendations are used to compute criteria (true/false) associated with back-extrapolated volume (BEV), time of forced expiration >6 s ($t_{\mathrm{FE>6\ s}}$) and existence of plateau (EOP) from the volume–time curve, along with time to peak expiratory flow (t_{PEF}).

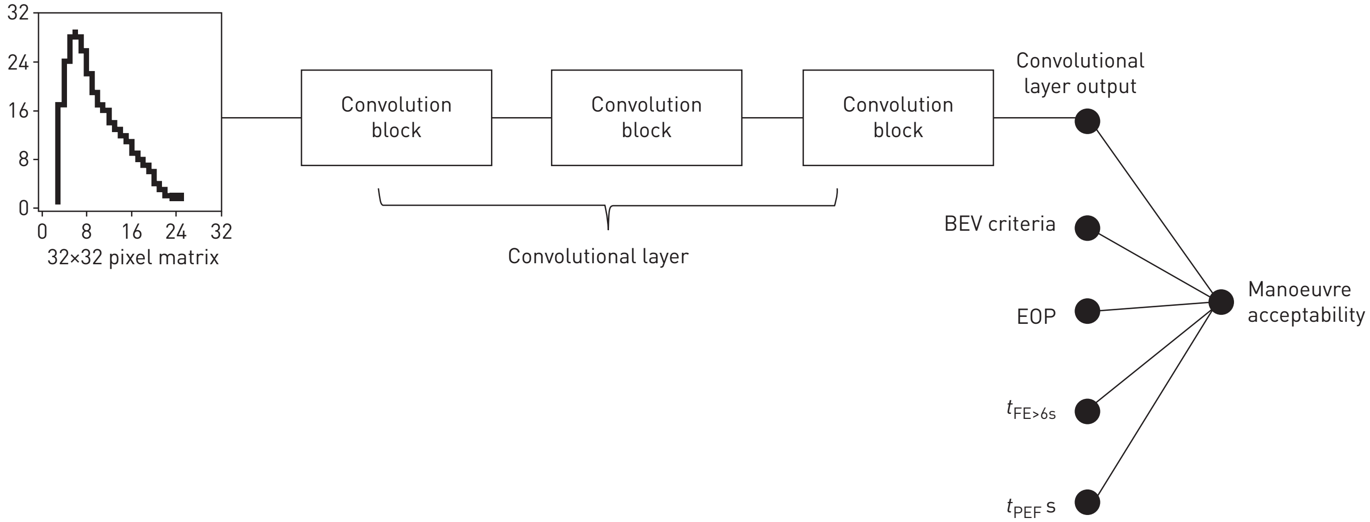

FIGURE 2 Convolutional neural network architecture. The pixel matrix of maximal expiratory flow–volume loop (MEFVC) is passed through a convolutional layer composed of three convolutional blocks (supplementary material). The convolved output, which represents features from the MEFVC, along with back-extrapolated volume (BEV) criteria, time of forced expiration >6 s ($t_{FE>6s}$), existence of plateau (EOP) and time to peak expiratory flow (t_{PEF}), are fed into a simple neural network. The final output is the probability of manoeuvre acceptability, which is considered acceptable if >0.5.

Additionally, we ensured that the training set included all post-bronchodilator sessions (n=458). Thus, each set had unique sessions with non-overlapping individuals. We developed our models using the development set (training and validation set) and evaluated them on the test set.

Manoeuvre acceptability model

We show the CNN architecture in figure 2. The MEFVC pixel matrix is passed into a convolutional layer composed of three convolutional blocks. The output of the convolutional layer, which represents features learned from the MEFVC, along with other input features, are passed into a fully connected network. The final output of the model is the probability of manoeuvre acceptability, with a value >0.5 considered as acceptable. We trained the model against the gold standard labels for acceptability (A-rated curves, 54%) in the training set and tuned its hyperparameters in the validation set. More details on CNN architecture and model development can be found in supplement S1. We used Keras deep-learning Python library with Tensorflow backend for model development [16].

Manoeuvre usability model

Since the proportion of usable curves (A-, B- or C- rated curves, 93%) were disproportionately larger than discarded curves (D-rated), we first subsampled the training set to include all the discarded curves (n=2022) along with 700 of each of A-, B- and C-rated curves to create a balanced dataset of usable and unusable curves. Then, using a transfer-learning approach [17], we recalibrated our acceptability model to predict manoeuvre usability. We considered a manoeuvre usable if the output probability was >0.5.

Model interpretability

We used a game theory based concept called Shapley value to estimate the evidence from different portions of the MEFVC and other ATS/ERS-based features towards the model's output of acceptability and usability [18]. To quantify the visual evidence from start of test, which affects FEV_1 measurement, we considered the pattern of MEFVC from maximum inspiration until 1 s after exhalation begins (MEFVC<1 s). This was achieved by dividing the MEFVC pixel matrix horizontally into two-pixel regions, MEFVC<1 s and MEFVC>1 s, at the point of FEV_1 (supplement S1). Then, we calculated the Shapley values of pixels associated with MEFVC<1 s and MEFVC>1 s along with BEV criteria, $t_{FE>6\ s}$, EOP and t_{PEF} using a technique called Shapley additive explanation [19]. A positive Shapley value is interpreted as evidence supporting a model's prediction, while a negative Shapley value is counter-evidence. The magnitude of Shapley value denotes the strength of the contribution.

Statistical analysis

We used one-way ANOVA to analyse differences in means between effort-ratings. Chi-squared tests were used to compare proportions across the effort ratings. We reported the model's performance using accuracy, sensitivity, specificity, positive predicted value, negative predicted value (NPV) and area under the receiver operating characteristic curve (AUROC) in the test set. We used McNemar's test to compare the model's accuracy against a simple ATS/ERS algorithm that predicts acceptability and usability when BEV criteria and EOT ($t_{\mathrm{FE}>6\ \mathrm{s}}$ or EOP) criteria are met. This comparison reflects the integration of the current ATS/ERS algorithm into commercial spirometry software. We interpreted the model using bar plots of Shapley values for individual curves and summarised the Shapley value evidence in the entire population using forest plots. Using multiple paired-proportion tests, we compared spirometry sessions that satisfied ATS/ERS repeatability criteria in gold-standard (technician evaluation), CNN and the ATS/ERS rules-based acceptable curves [1]. Finally, we used multiple paired t-tests to compare the distribution of best FEV_1 and FEV_1/FVC ratio after applying repeatability criteria or by selecting highest FEV_1 and FVC whenever repeatability failed. All results were reported on the test set (n=3738 curves or 710 sessions). The significance level was 0.05 and values were expressed as mean±SD or median (interquartile range). All statistical analysis was performed using R software version 3.3.3 [20].

Results

Baseline characteristics

There were a total of 36 873 curves available with raw data and a label from skilled technician in NHANES 2011–2012 (*e.g.* A–D). Just over half (54%) satisfied ATS/ERS acceptability criteria, whereas 93% were considered useful (*i.e.* reportable FEV_1). Table 1 shows a summary of the ATS/ERS quantitative features across the effort quality ratings. The BEV criteria and EOP were satisfied in 97% and 100% of the A-rated curves, respectively, supporting that A-rated curves contained a satisfactory start and EOT. Furthermore, t_{PEF} was highest in B-rated curves (0.13±0.07 s, n=5890) while EOP was achieved only in 8% of the C-rated curves (n=8605), agreeing with their respective label description. Finally, BEV criteria and EOP were satisfied in less than half of the rejected curves (D-rated, n=2550).

Model evaluation

In the test set (n=3738 curves), the CNN to determine ATS/ERS manoeuvre acceptability criteria (A-rating, prevalence 54%) resulted in an accuracy of 87% with a good sensitivity (87%) and specificity (86%). The model to determine ATS/ERS manoeuvre usability (A-, B- or C-rating) demonstrated an excellent accuracy (92%), sensitivity (92%) and specificity (96%) but a low NPV (50%), as prevalence of usable curves (93%) dominated over rejected curves (table 2). In addition, we observed high AUROC in both the cases (0.93 and 0.98, respectively; figure S1).

Comparison against ATS/ERS quantitative criteria

A simple ATS/ERS rule-based algorithm that predicted manoeuvre acceptability and usability when BEV criteria, $t_{\mathrm{FE}>6\ \mathrm{s}}$ and EOP were satisfied resulted in a significantly lower (p<0.0001) accuracy (78% and 66%, respectively) and AUROC (0.78 and 0.73, respectively; figure S1) when compared to the CNN models (table 2). This demonstrated the advantage of analysing the qualitative aspects of the MEFVC in addition to the simple ATS/ERS quantitative criteria.

TABLE 1 Baseline characteristics of spirometry curves and their effort ratings from the National Health and Nutritional Examination Survey USA 2011–2012

	A All curve quality attributes were acceptable	B Large time to PEF or a nonrepeatable PEF	C <6 s of exhalation or no plateau	D Cough or large extrapolated volume	p-value
Curves n	19 828	5890	8605	2550	
EOP	97	96	8	49	<0.001
$t_{\mathrm{FE}>6\ \mathrm{s}}$	91	87	10	45	<0.001
BEV criteria	100	99	99	44	<0.001
t_{PEF} **s**	0.07±0.02	0.13±0.07	0.10±0.07	0.21±0.18	<0.001

Data are presented as % or mean±SD, unless otherwise stated. PEF: peak expiratory flow; EOP: existence of plateau in volume–time curve defined as no change in volume (⩽0.025 L) for ⩾1 s of expiration; $t_{\mathrm{FE}>6\ \mathrm{s}}$: time of forced expiration, defined as difference between time at residual volume and time 0 obtained by back-extrapolation, is >6 s; BEV: back-extrapolated volume is less than the maximum of 5% of forced vital capacity or 0.15 L; t_{PEF}: time to PEF from time 0.

TABLE 2 The performance of a convolutional neural network (CNN) and a rule-based American Thoracic Society/European Respiratory Society algorithm at predicting manoeuvre acceptability and usability in the test set (n=3738 curves)

	Accuracy %	Sensitivity %	Specificity %	PPV %	NPV %	AUROC
CNN acceptability	87	90	85	85	89	0.93
Rule-based acceptability	78	88	67	73	84	0.78
CNN usability	92	92	96	99	50	0.98
Rule-based usability	66	65	82	98	16	0.73

PPV: positive predictive value; NPV: negative predictive value; AUROC: area under the receiver operating characteristic curve.

Model interpretability

For a manoeuvre predicted as acceptable (probability of acceptability (P(acceptability))=0.90), we show the MEFVC and volume–time curve along with a Shapley value evidence plot in figure 3a. A satisfactory start of test is supported by a positive Shapley value (0.13) from MEFVC<1s (MEFVC pattern in first second of exhalation) indicating no artefacts in this region, and dominant over a compliant BEV which was considered redundant (Shapley value 0). A satisfactory EOT is also evidenced by positive evidence from EOP (Shapley value 0.095) and $t_{\mathrm{FE}>6\ \mathrm{s}}$ (Shapley value 0.075).

In figure 3b, we show a manoeuvre predicted as unacceptable (P(acceptability)=0.21) but usable (P(usability)=0.91). The main reason for unacceptability stems from an unsatisfactory EOT due to a lack of EOP (Shapley value −0.31). However, we can observe that this curve might have a valid FEV_1 due to positive evidence from MEFVC<1s (Shapley value 0.11).

Finally, we show a manoeuvre predicted as unusable (P(usability)=0.16) with an evidence plot for usability in figure 3c. The highest counterevidence comes from MEFVC<1s (Shapley value −0.38), and we can clearly see cough artefacts on the corresponding highlighted section of MEFVC. Interestingly, this curve had a compliant BEV and a satisfactory EOP and $t_{\mathrm{FE}>6\ \mathrm{s}}$, but their evidence was comparatively very low.

Global interpretability

In figure 4a, we summarise the Shapley value evidence towards prediction of manoeuvre acceptability in the test set (n=3738 curves). MEFVC<1s (supporting evidence 0.137 (0.053), counter-evidence −0.249 (0.175)), and EOP (supporting evidence 0.095 (0.016), counter-evidence −0.276 (0.113)) were the most important factors in determining acceptability. Furthermore, MEFVC<1s was the single most important contributor (supporting evidence 0.365 (0.07), counterevidence −0.335 (0.086)) in determining curve usability (figure 3b). We further confirmed that a positive Shapley value for MEFVC<1s (0.142 (0.051)) and a EOP (0.096 (0.001)) were the strongest contributors in a subgroup that was predicted as acceptable (n=2021; figure S2a), while a negative Shapley value of MEFVC<1s (−0.336 (0.079)) was the strongest contributor in curves predicted as discarded (n=548; figure S2b).

On examining a subgroup of curves that were predicted as unacceptable but usable (n=1211, figure S2c), we observed that a lack of EOP provided the most evidence against acceptability (−0.22 (0.363)). This implied that curves with a failed EOT criteria (61% of n=1211) still contained a usable FEV_1. In addition, we observed strong counterevidence from MEFVC<1s (−0.172 (0.378)), which was due to presence of B-rated curves (25% of n=1211) that contained a nonrepeatable or a large time to peak flow.

Effect on distribution of best FEV_1 and FEV_1/FVC

In the test set spirometry sessions (n=710), we observed no significant differences (p>0.10) in the proportion of satisfactory repeatability criteria between the CNN (81%), gold standard (82%) and ATS/ERS algorithm (86%).

Furthermore, the best FEV_1 after checking repeatability criteria was not statistically different (p>0.05) in CNN and ATS/ERS approaches from the gold standard (2.76±0.83 L). However, significant differences (p<0.001) in FEV_1/FVC ratio were observed between the gold standard (82±7.7%) and ATS/ERS approach, but not between gold-standard and CNN approaches (p>0.05). Finally, we noted small differences in lower limit of normal (=mean−1.645 SD) between the three approaches (table S1).

Discussion

We developed a novel deep-learning approach using CNN to classify spirometry manoeuvre according to ATS/ERS acceptability criteria with a valid start and EOT. The model was both sensitive (90%) and

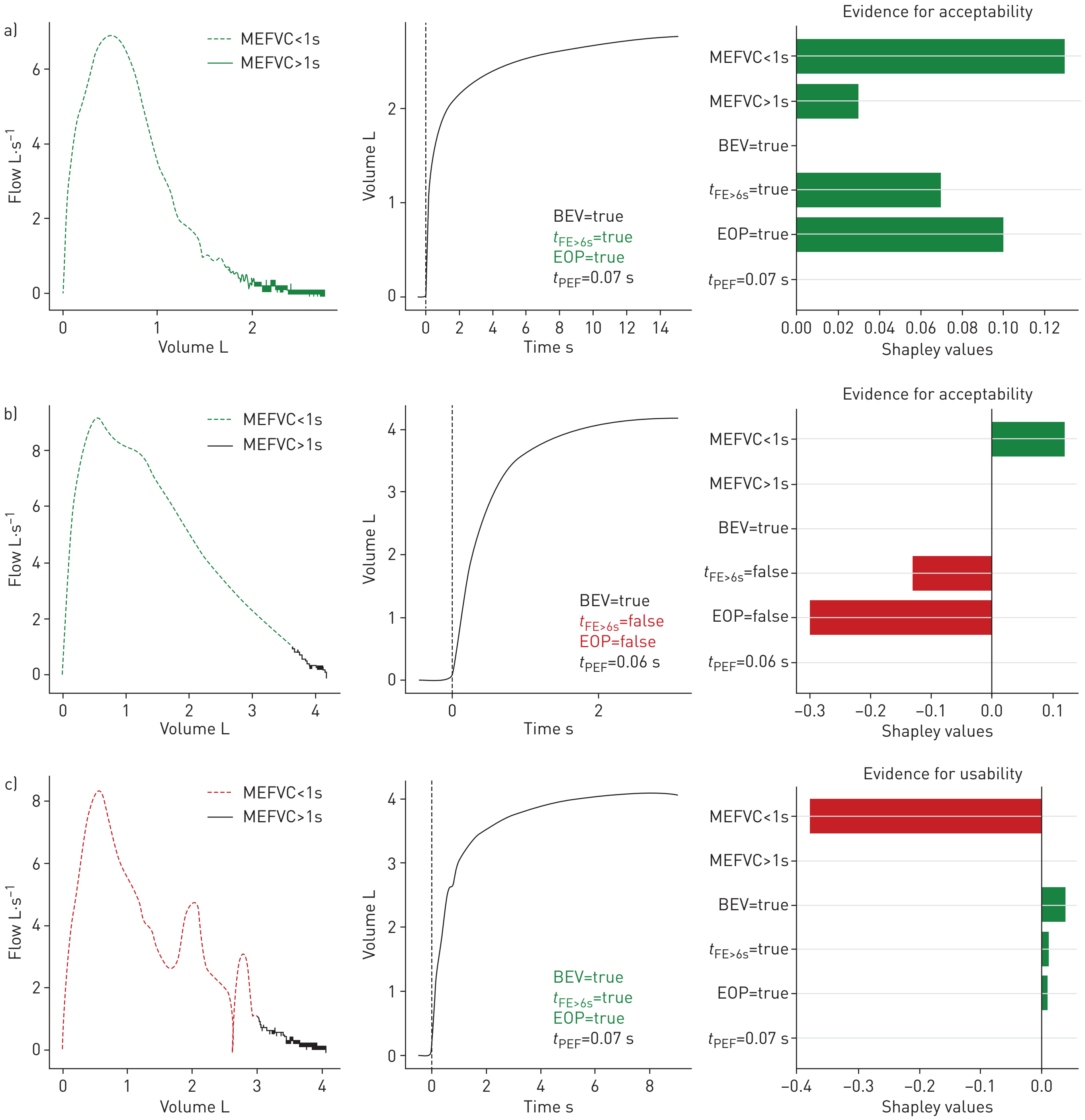

FIGURE 3 Interpreting the convolutional neural network (CNN). The evidence, as quantified by Shapley values, from maximal expiratory flow volume from maximum inspiration until 1 s after exhalation begins (MEFVC<1s), MEFVC after 1 s of exhalation (MEFVC>1s), back-extrapolated volume (BEV) criteria, time of forced expiration >6 s ($t_{FE>6\ s}$), existence of plateau (EOP) and time to peak expiratory flow (t_{PEF}) in examples of curves predicted as a) acceptable with satisfactory start and end of test (P(acceptability)=0.90); b) unacceptable but usable (P(acceptability)=0.25, P(usability)=0.91) and c) unacceptable and unusable (P(acceptability)=0.02, P(usability)=0.16). P(acceptability): probability of acceptability; P(usability): probability of usability.

specific (85%) when evaluated on a test set (n=3738 curves). The acceptability model was then recalibrated to identify usable curves (reportable FEV_1) with a high sensitivity (92%) and specificity (96%). Model interpretation results show that MEFVC<1s and EOP were the most important factors in determining acceptability, while MEFVC<1s entirely determined usability. Thus, our models captured the expected attributes from raw data to concur with ATS/ERS recommendations on manoeuvre acceptability that

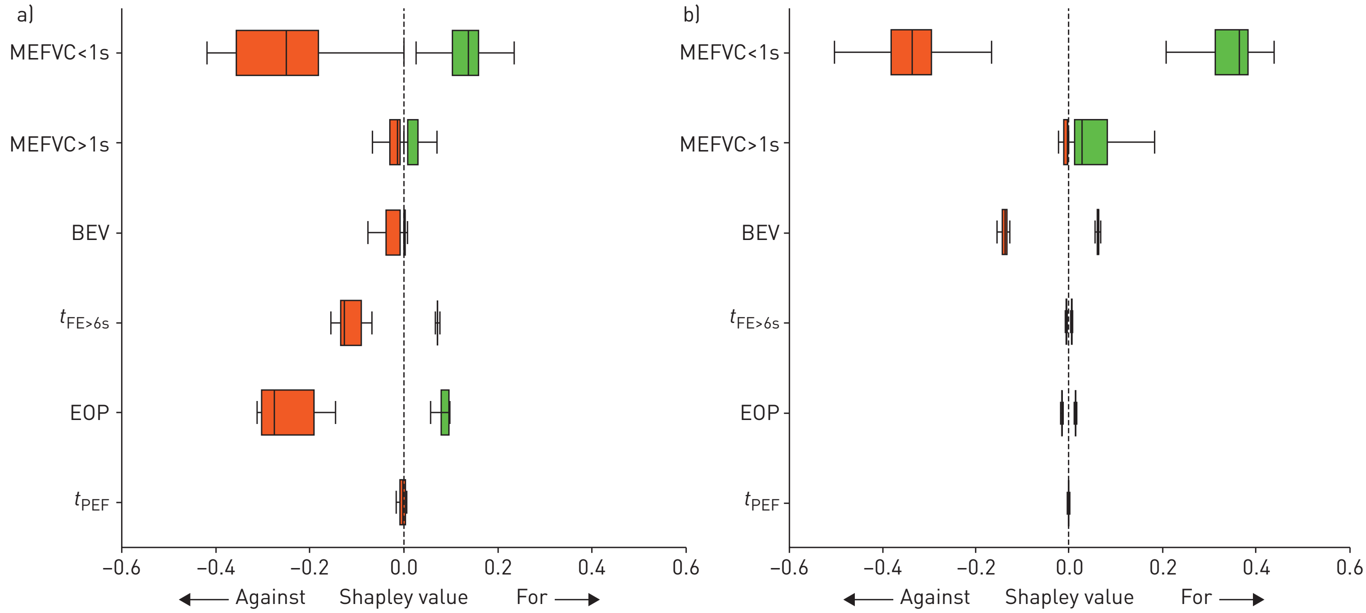

FIGURE 4 Global interpretability. Forest plots summarising evidence, quantified by Shapley values, from maximal expiratory flow volume curve pattern within first second of exhalation (MEFVC<1s), MEFVC after first second of exhalation (MEFVC>1s), back-extrapolated volume (BEV) criteria, time of forced expiration >6 s ($t_{FE>6\ s}$), existence of plateau (EOP) and time to peak flow (t_{PEF}), in determining curve a) acceptability and b) usability in the test set (n=3738 curves).

mandate a satisfactory start and EOT. In addition, they adhered to ATS/ERS recommendations on manoeuvre usability that mandate an undisturbed FEV_1. Thus, this automated approach for spirometry quality control that combines both the visual experience of technicians and ATS/ERS quantitative ruleshelps in standardising the application of ATS/ERS guidelines of within-manoeuvre acceptability and usability criteria.

The most interesting aspect of the CNN approach is that while it adheres to the ATS/ERS quantitative rules, it also adds consistency in interpreting the protocols that currently suffer from variability arising from visual evaluation [4, 5, 8]. A real-life application would involve determining acceptability and usability of individual manoeuvres by the CNN approach followed by checking the ATS/ERS repeatability criteria and finally reporting the best test results and session grades [21]. By collaborating with spirometry manufacturers, the application can provide manoeuvre feedback in real time, and thereby assist technicians and primary care physicians in daily practice to comply better with ATS/ERS standards. In mobile spirometry [22], our system can remotely verify acceptability requirements and further provide feedback to the user. Additionally, our application could be used as an independent module during retrospective evaluation of spirometry curves (*e.g.* clinical and epidemiological studies) and in cases where spirograms are remotely reviewed by humans [23].

Our study is the first to analyse MEFVC as a pixel matrix (image) using deep learning. This approach leads to a significantly superior performance ($p<0.0001$) when compared to a rule-based algorithm containing just ATS/ERS quantifiable criteria (accuracy=78% *versus* 87% for acceptability and 66% *versus* 92% for usability). Moreover, unlike the rules-based approach, the distribution of best FEV_1 and FEV_1/FVC ratio after repeatability criteria does not change between gold standard and the CNN, implying its implementation will not change clinical outcomes involving cut-offs for abnormality diagnosis. However, a demographic-based analysis may be required to completely establish these facts. We further noted that MEFVC<1s was dominant over BEV in determining acceptability and usability. This is because MEFVC<1s could capture a host of anomalous patterns associated with cough in first second, variable effort, obstructed mouthpiece, submaximal blast, *etc.* in addition to the hesitation pattern associated with excessive extrapolated volume [24]. Finally, our approach exploits the efficient pattern recognition capabilities of deep learning and also generates clear explanations providing a dual benefit of accuracy and explainability, which is often a trade-off [25]. Our Shapley value explanations provide a hybrid system that quantifies the visual or qualitative evidence from MEFVC pattern and the objective evidence from ATS/ERS criteria.

While the current paradigm of determining manoeuvre acceptability and usability may be enough for clinical practice, further determining the artefact type in a manoeuvre could have underscored the power of artificial intelligence in achieving a detailed level of interpretation. In this study, only a rudimentary characterisation of artefacts was available such as poor peak flow, nonsatisfactory EOT or cough. At this stage, these artefacts are represented in negative Shapley values of MEFVC<1s or MEFVC>1s. In the future, we believe that algorithms, which could also point to the presence of specific artifacts (hesitation, variable effort, cough, *etc.*) in addition to acceptability and usability, will further improve real-time feedback to the end-user.

A direct comparison with past methods may not be fair [4, 9], as their sample sizes were quite small, contained different labels and involved manual selection and tuning of features [11, 26]. By contrast, the current CNN identifies the requisite features associated with a valid MEFVC without the explicit need for manual programming and further takes advantage of the availability of a large labelled dataset. Although other choices of input data (*e.g.* sequential flow data) and deep-learning frameworks (*e.g.* using recurrent networks) may work, our choice of input and modelling (a technician inspecting MEFVC and volume–time curves) optimally reflects the clinical setting. In fact, our CNN captures the clinical decision process, as it extracts visual features from a 32×32 image of MEFVC and considers ATS/ERS quantifiable criteria in automating manoeuvre acceptability. Moreover, the conversion of raw flow–volume data to 32×32-pixel matrices occurs algorithmically, without any intermediate step of saving image files (supplement S1).

A major drawback of our study included a lack of data on forced inspiration. An insufficient inspiration can severely affect FVC, a fact that was not stressed in ATS/ERS 2005 spirometry guidelines. The 2019 update of ATS/ERS standardisation of spirometry has incorporated rules on the inspiratory manoeuvre before and after forced expiration [3], and this will need to be included in future iterations of the algorithm when forced inspiration data are available. Another potential drawback of this study was a lack of diversity in our population. NHANES 2011–2012 was an epidemiological study and mostly contained healthy individuals. Since the shape of MEFVC is affected by the presence of emphysema, restrictive disease or extrathoracic obstruction, future validation studies in disease cohorts should incorporate this aspect. Finally, we would also like to point out that a 32×32-pixel matrix is an inefficient way of representing to input, as it is a sparse matrix (contains lot more ones than zeros). The past success of CNN on handwritten character recognition, which also involved sparse input, inspired our modelling choice [27]. While a CNN efficiently utilises a low-resolution image that is not adequate for a visual inspection, human vision is still vastly superior when it comes to generalisation.

We are living in an era in which technology significantly influences medicine and medical applications. Our study perfectly reflects that technological wave, which aims to improve quality of tests and care, and therefore patient outcomes. Our technology emulates the experience of skilled spirometry readers into a computer algorithm that could bring the necessary quality checks to all areas of spirometry use.

Author contributions: Conception and study design: N. Das, K. Verstraete, S. Stanojevic, M. Topalovic, J-M. Aerts and W. Janssens. Acquisition, analysis or interpretation: N. Das, K. Verstraete, S. Stanojevic, M. Topalovic, J-M. Aerts and W. Janssens. Manuscript preparation and critical revision: N. Das, K. Verstraete, S. Stanojevic, M. Topalovic, J-M. Aerts and W. Janssens. Funding obtained by N. Das and W. Janssens.

Support statement: N. Das is supported by Research Foundation-Flanders (FWO) strategic basic fellowship 2018–2019. W. Janssens is a senior clinical investigator of the Flemish research foundation. The study is supported by an AstraZeneca grant in respiratory pathophysiology. Funding information for this article has been deposited with the Crossref Funder Registry.

Conflict of interest: N. Das has a patent spirometry evaluation method (application number 1914446.8 UK patent office) pending. K. Verstraete has nothing to disclose. S. Stanojevic has nothing to disclose. M. Topalovic is a co-founder of a spin-off company ArtiQ. J-M. Aerts has not nothing to disclose. W. Janssens reports grants from AstraZeneca and Chiesi, fees for lectures and advisory boards from AstraZeneca and Chiesi, outside the submitted work; has a patent on pulmonary function loops pending; and is co-founder of ArtiQ, a spin-off company of KU Leuven.

References

1 Miller MR, Hankinson J, Brusasco V, *et al.* Standardisation of spirometry. *Eur Respir J* 2005; 26: 319–338.
2 American Thoracic Society. Standardization of spirometry: 1994 update. *Am J Respir Crit Care Med* 1995; 152: 1107–1136.
3 Graham BL, Steenbruggen I, Miller MR, *et al.* Standardization of spirometry 2019 update. An official American Thoracic Society and European Respiratory Society technical statement. *Am J Respir Crit Care Med* 2019; 200: e70–e88.
4 Velickovski F, Ceccaroni L, Marti R, *et al.* Automated spirometry quality assurance: supervised learning from multiple experts. *IEEE J Biomed Heal Inform* 2018; 22: 276–284.
5 Hankinson JL, Eschenbacher B, Townsend M, *et al.* Use of forced vital capacity and forced expiratory volume in 1 second quality criteria for determining a valid test. *Eur Respir J* 2015; 45: 1283–1292.
6 Tan WC, Bourbeau J, O'Donnell D, *et al.* Quality assurance of spirometry in a population-based study – predictors of good outcome in spirometry testing. *COPD* 2014; 11: 143–151.

7 Pérez-Padilla R, Vázquez-García JC, Márquez MN, *et al.* Spirometry quality-control strategies in a multinational study of the prevalence of chronic obstructive pulmonary disease. *Respir Care* 2008; 53: 1019–1026.

8 Eaton T, Withy S, Garrett JE, *et al.* Spirometry in primary care practice: the importance of quality assurance and the impact of spirometry workshops. *Chest* 1999; 116: 416–423.

9 Melia U, Burgos F, Vallverdú M, *et al.* Algorithm for automatic forced spirometry quality assessment: technological developments. *PLoS One* 2014; 9: e116238.

10 Yamashita R, Nishio M, Do RKG, *et al.* Convolutional neural networks: an overview and application in radiology. *Insights Imaging* 2018; 9: 611–629.

11 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; 521: 436–444.

12 Centers for Disease Control and Prevention (CDC). Data Documentation, Codebook, and Frequencies (Pre and Post-Bronchodilator). National Health and Nutrition Examination Survey 2011–2012. Hyattsville, CDC, 2014.

13 Centers for Disease Control and Prevention (CDC); National Center for Health Statistics (NCHS). Respiratory Health Spirometry Procedures Manual. National Health and Nutrition Examination Survey 2011–2012. Hyattsville, CDC, 2011.

14 Centers for Disease Control and Prevention (CDC); National Center for Health Statistics (NCHS)CDC. 2014. – Data Documentation, Codebook, and Frequencies (Spirometry Raw curve data). National Health and Nutrition Examination Survey 2011–2012.

15 Klein B. NumPy. *In:* Klein B. Einführung in Python 3. Munich, Carl Hanser Verlag, 2014; pp. 323–344.

16 Chollet F. 2015. Keras: The Python Deep Learning library. https://keras.io

17 Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng* 2010; 22: 1345–1345.

18 Shapley LS. A value for n-person games. *In:* Roth AE, ed. The Shapley Value. Cambridge, Cambridge University Press, 2009.

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

20 R Core Team. R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing, 2017. www.R-project.org/

21 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.

22 Hernández CR, Fernández MN, Sanmartín AP, *et al.* Validation of the portable Air-Smart Spirometer. *PLoS One* 2018; 13: e0192789.

23 Burgos F, Disdier C, De Santamaria EL, *et al.* Telemedicine enhances quality of forced spirometry in primary care. *Eur Respir J* 2012; 39: 1313–1318.

24 National Institute for Occupational Safety and Health (NIOSH). Spirometry Quality Assurance: Common Errors and Their Impact on Test Results. Washington, DC, NIOSH, 2012.

25 Gunning D, Stefik M, Choi J, *et al.* XAI – explainable artificial intelligence. *Sci Robot* 2019; 37: eaay7120.

26 O'Mahony N, Campbell S, Carvalho A, *et al.* Deep Learning vs. Traditional Computer Vision. *In*: Arai K, Kapoor S, eds. Advances in Computer Vision. CVC 2019. Advances in Intelligent Systems and Computing, Vol. 943. Cham, Springer, 2020.

27 LeCun Y, Cortes C, Bottou L, *et al.* Comparison of Learning Algorithms for Handwriting Digit Recognition. International Conference on Artificial Neural Networks, 1995.