Early View

Original article

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

Nilakash Das, Kenneth Verstraete, Sanja Stanojevic, Marko Topalovic, Jean-Marie Aerts, Wim Janssens

Please 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; in press (<https://doi.org/10.1183/13993003.00603-2020>).

This manuscript has recently been accepted for publication in the *European Respiratory Journal*. It is published here in its accepted form prior to copyediting and typesetting by our production team. After these production processes are complete and the authors have approved the resulting proofs, the article will move to the latest issue of the ERJ online.

Title Page

Title

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

Authors

Nilakash Das¹, Kenneth Verstraete¹, Sanja Stanojevic², Marko Topalovic^{1, 3}, Jean-Marie Aerts⁴
and Wim Janssens¹

Author' affiliations

¹ Laboratory of Respiratory Diseases and Thoracic Surgery, Department of Chronic Diseases Metabolism and Ageing, Katholieke Universiteit Leuven, Leuven, Belgium

² Translational Medicine, Division of Respiratory Medicine, Hospital for Sick Children, Toronto, Canada

³ ArtiQ NV, Leuven, Belgium

⁴ Division Animal and Human Health Engineering, Department of Biosystems, Katholieke Universiteit Leuven, Leuven, Belgium

Corresponding author

Wim Janssens

O&N1 Herestraat 49 bus 706

3000 Leuven, Belgium

Email: wim.janssens@uzleuven.be

Take home message for Twitter

We developed a deep learning models to standardize ATS/ERS spirometric acceptability and usability criteria. This approach reduces the inter-technician variability and provides instant feedback to the user.

Abstract

Rationale

While ATS/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 inter-technician variability. We propose a deep learning approach called convolutional neural network (CNN), to standardize spirometric manoeuvre acceptability and usability.

Methods and methods

In 36,873 curves from the national health and nutritional examination survey (NHANES) USA 2011-12, 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 FEV1. 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=3,738), 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<1s (MEFVC pattern within first second of exhalation) and plateau in volume-time were most important in determining acceptability, while MEFVC<1s entirely determined usability.

Conclusion

The CNNs identified relevant attributes in spirometric curves to standardize 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.

Abstract

Keywords

Spirometry; Manoeuvre acceptability; Deep learning; Quality control;

Text

Introduction

The validity of spirometric indices such as forced expiratory volume in 1 second (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]. Further, ATS/ERS also provides recommendations on the usability of a manoeuvre when it does not meet a satisfactory EOT but still contains a valid FEV₁.

The evaluation of forced expiratory manoeuvre acceptability involves quantitative criteria, namely back-extrapolated volume (BEV), time of forced expiration (tFE) and existence of a plateau (EOP) in volume-time (VT) curve. The current guidelines also 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, non-maximal effort, obstructed mouthpiece etc. [1, 2]. For usability (i.e. whether an FEV₁ can be reported), the criteria are based on an acceptable BEV and the technician's judgement to interpret if the measurement of FEV₁ 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 also 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 standardized 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, tFE 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 employed features were not only difficult to interpret, but also fail to capture all possible artefacts making them less robust to apply in practice. Deep learning models like convolutional neural network (CNN) have become the dominant method in computer vision tasks today, 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 previous feature driven approaches. Since spirometry review is accomplished by visual inspection of the MEFVC and VT curves [2], we hypothesize that application of 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 2011-12 [12]. We also describe methodology that allows interpretation of the model to provide feedback to the user.

Materials and methods

Study subjects

NHANES 2011-12 spirometry component

Spirometry data were collected in NHANES 2011-12. Eligible participants between the ages of 6 and 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 2011-12 dataset contained 6,696 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-second interval during forced expiration [14]. Technicians had assigned an effort-rating variable 'SPAQEFF' to individual curves as: **A**-the curve quality attributes were acceptable (N=19,828); **B**- the curve had a large time to peak flow or a non-repeatable peak flow (N=5,890); **C**- the curve had either less than 6 seconds of exhalation or no plateau (N=8,605); **D**- the curve had either a cough or large extrapolated volume (N=2,550). Further, a variable 'SPAACC' indicated that curves with ratings A, B and C (N=34,323) were used for obtaining FEV₁ or FVC . Only curves with rating D were rejected in NHANES 2011-12.

Study design

In this retrospective analysis of the NHANES 2011-12 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 EOT). Curves with effort-ratings A, B or C were considered as the gold standard for usable manoeuvres (i.e. reportable FEV₁).

Model development

Processing of input

We first 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-by-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(A) 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 tFE (difference between time at residual volume and t_0) > 6s criteria. We also defined existence of plateau (EOP) as no change in volume (≤ 0.025 L) in VT curve for at least 1 s of expiration [1]. Finally, we calculated time to peak flow (tPEF) from t_0 . All input processing was done using Python scripts [15].

Development and test data

We split NHANES spirometry sessions (N=7,154) randomly into 80% for training (5726 sessions or 29,452 curves), 10% for validation (711 sessions or 3,683 curves) and 10% for testing (710 sessions or 3,738 curves). We additionally 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 1(B). 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 (FC) network. The final output of the model is the probability of manoeuvre acceptability, with a value above 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=2,022) along with 700 curves each from A, B and C rating 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 greater than 0.5.

Model interpretability

We used a game-theory based concept called Shapley value (SV) 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 FEV₁ measurement, we considered the pattern of MEFVC from maximum inspiration until 1 second after exhalation begins (MEFVC<1s). This was achieved by dividing the MEFVC pixel matrix horizontally into two-pixel regions, MEFVC<1s and MEFVC>1s, at the point of FEV₁ (Supplement S1). Then, we calculated the Shapley values of pixels associated with MEFVC<1s and MEFVC>1s along with BEV criteria, tFE>6s, EOP and

tPEF using a technique called Shapley additive explanation (SHAP) [19]. A positive SV is interpreted as evidence supporting a model's prediction while a negative SV is counter-evidence. The magnitude of SV denotes the strength of the contribution.

Statistical analysis

We used one-way ANOVA to analyse differences in means between effort-ratings. χ^2 tests was used to compare proportions across the effort ratings. We reported the model's performance using accuracy, sensitivity, specificity, positive predicted value (PPV), negative predicted value (NPV) and area under a receiver operating characteristics 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 (tFE>6s or EOP) 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 SVs for individual curves and summarised the SV 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 FEV₁ and FEV₁/FVC after applying repeatability criteria or by selecting highest FEV₁ and FVC whenever repeatability failed. All results were reported on the test set (N=3,738 curves or 710 sessions). The significance level was 0.05 and values were expressed as mean $\pm$ standard deviation 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 2011-12 (e.g. A_D). Just over half (54%) satisfied ATS/ERS acceptability criteria whereas 93% were considered useful (i.e. reportable FEV₁). 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. Further, tPEF was highest in B-rated curves (0.13 ± 0.07 s, N=5,890) while EOP was achieved only in 8% of the C-rated curves (N=8,605), agreeing with their respective label description. Finally, BEV criteria and EOP were satisfied in less than half of the rejected curves (D-rated, N=2,550).

Model evaluation

In the test set (N=3,738 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). We also observed high AUROC in both the cases (0.93 and 0.98, figure S1).

Comparison against ATS/ERS quantitative criteria

A simple ATS/ERS rule-based algorithm that predicted manoeuvre acceptability and usability when BEV criteria, tFE>6s and EOP was 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.

Model interpretability

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

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

Finally, we show a manoeuvre predicted as unusable ($P(\text{usability})$ =0.16) with an evidence plot for usability in figure 3C. The highest counter-evidence comes from $\text{MEFVC}<1\text{s}$ (SV=-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 $\text{tFE}>6\text{s}$, but their evidence was comparatively very low.

Global interpretability

In figure 4A, we summarised the SV evidence towards prediction of manoeuvre acceptability in the test set (N=3,738 curves). $\text{MEFVC}<1\text{s}$ (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. Further, 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 SV for MEFVC<1s (0.142 (0.051)) and a EOP (0.096 (0.001)) were the strongest contributors in a sub-group that was predicted as acceptable (N=2021, supplement figure S2 (a)), while a negative SV of MEFVC<1s (-0.336 (0.079)) was the strongest contributor in curves predicted as discarded (N=548, figure S2(b)).

On examining a sub-group of curves that were predicted as unacceptable but usable (N=1,211, figure S2(c)), 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=1,211), still contained a usable FEV₁. We also observed strong counterevidence from MEFVC<1s (-0.172 (0.378)), which was due to presence of B-rated curves (25% of N=1,211) that contained a non-repeatable or a large time to peak flow.

Effect on distribution of best FEV₁ and FEV₁/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%).

Further, the best FEV₁ after checking repeatability criteria, was not statistically different ($p>0.05$) in CNN and ATS/ERS approach from gold standard ($2.76\pm0.83L$). However, significant differences ($p<0.001$) in FEV₁/FVC were observed between the gold standard ($82\pm7.7\%$) and ATS/ERS approach, but not ($p>0.05$) between gold standard and CNN approach. Finally, we noted small differences in LLN (=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 specific (85%) when evaluated on a test set (N=3,738 curves). The acceptability model was then recalibrated to identify usable curves (reportable FEV₁) with a high sensitivity (92%) and specificity (96%). Model interpretation results show that MEFVC<1s (MEFVC pattern within first second of exhalation) 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 mandate a satisfactory start and EOT. They also adhered to ATS/ERS recommendations on manoeuvre usability that mandate an undisturbed FEV₁. Thus, this automated approach for spirometry quality control that combines both the visual experience of technicians and ATS/ERS quantitative rules, helps in standardizing the application of ATS/ERS guidelines of within-manoeuve 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. Our application could also 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% vs 87% for acceptability and 66% vs 92% for usability). Moreover, unlike the rules-based approach the distribution of best FEV1 and FEV1/FVC 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, sub-maximal 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 is 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 characterization of artefacts was available such as poor

peak flow, non-satisfactory end-of-test (EOT) or cough. At this stage, these artefacts are represented in negative Shapley values of MEFVC $< \text{or} > 1$ second. 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 VT curves) is optimally reflecting the clinical setting. In fact our CNN captures the clinical decision process as it extracts visual features from a 32x32 image of MEFVC and considers ATS/ERS quantifiable criteria in automating manoeuvre acceptability. Moreover, the conversion of raw flow-volume data to 32x32 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 standardization 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 2011-12 was an epidemiological study and mostly contained healthy individuals. Since the shape of MEFVC is affected by the presence of emphysema, restrictive disease or extra-thoracic obstruction, future validation studies in disease cohorts should incorporate this aspect. Finally, we would also like to point out that a 32x32 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 hand-written character recognition, which also involved sparse input, has inspired our modelling choice[27]. While a CNN efficiently utilizes a low resolution image that is not adequate for a visual inspection, human vision is still vastly superior when it comes to generalization.

We are living in an era in which technology significantly influences medicine and medical applications. Our study perfectly reflects that technologic wave, which aims to improve quality of tests, 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.D, K.V, S.S, M.T, J.M.A and W.J. Acquisition, analysis or interpretation: N.D, K.V, S.S, M.T, J.M.A and W.J. Manuscript preparation and critical revision: N.D, K.V, S.S, M.T, J.M.A and W.J. Funding obtained by N.D and W.J

Conflict of interest

WJ and MT are founders of ArtiQ NV, Belgium

Support statement

ND is supported by Research Foundation-Flanders (FWO) strategic basic fellowship 2018-2019. WJ is a senior clinical investigator of the Flemish research foundation. The study is supported by an AstraZeneca grant in respiratory pathophysiology.

References

1. Miller MR, Hankinson J, Brusasco V, Burgos F, Casaburi R, Coates A, Crapo R, Enright P, van der Grinten CPM, Gustafsson P, Jensen R, Johnson DC, MacIntyre N, McKay R, Navajas D, Pedersen OF, Pellegrino R, Viegi G, Wagner J. Standardisation of spirometry. *Eur. Respir. J.* 2005. p. 319–338.
2. Crapo O, Hankinson JL, Irvin C, MacIntyre NR, Voter KZ, Wise RA, Graham B, O'Donnell C, Paoletti P, Roca J, Veigi G. Standardization of spirometry: 1994 Update. *Am. J. Respir. Crit. Care Med.* 1995.
3. Graham BL, Steenbruggen I, Miller MR, Barjaktarevic IZ, Cooper BG, Hall GL, Hallstrand TS, Kaminsky DA, McCarthy K, McCormack MC, Oropez CE, Rosenfeld M, Stanojevic S, Swanney MP, Thompson BR. Standardization of Spirometry 2019 Update. An Official American Thoracic Society and European Respiratory Society Technical Statement. *Am. J. Respir. Crit. Care Med.* [Internet] American Thoracic Society - AJRCCM; 2019; 200: e70–e88 Available from: <https://doi.org/10.1164/rccm.201908-1590ST>.
4. Velickovski F, Ceccaroni L, Marti R, Burgos F, Gistau C, Alsina-Restoy X, Roca J. Automated Spirometry Quality Assurance: Supervised Learning from Multiple Experts. *IEEE J. Biomed. Heal. Informatics* 2018; .
5. Hankinson JL, Eschenbacher B, Townsend M, Stocks J, Quanjer PH. Use of forced vital capacity and forced expiratory volume in 1 second quality criteria for determining a valid test. *Eur. Respir. J.* 2015; .
6. Tan WC, Bourbeau J, O'donnell D, Aaron S, Maltais F, Marciniuk D, Hernandez P, Cowie R, Chapman K, Sonia Buist A, Sin D, Mark Fitzgerald J. Quality assurance of

- spirometry in a population-based study-predictors of good outcome in spirometry testing. *COPD J. Chronic Obstr. Pulm. Dis.* 2014; 11: 143–151.
7. Pérez-Padilla R, Vázquez-García JC, Márquez MN, Menezes AMB, PLATINO Group. Spirometry quality-control strategies in a multinational study of the prevalence of chronic obstructive pulmonary disease. *Respir. Care* 2008; .
 8. Eaton T, Withy S, Garrett JE, Mercer J, Whitlock RML, Rea HH. 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, Velickovski F, Lluch-Ariet M, Roca J, Caminal P. Algorithm for automatic forced spirometry quality assessment: technological developments. *PLoS One* 2014; .
 10. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights Imaging* 2018.
 11. LeCun Y, Bengio Y, Hinton G, Y. L, Y. B, G. H. Deep learning. *Nature* 2015; 521: 436–444.
 12. Centers for Disease Control and Prevention. Data Documentation, Codebook, and Frequencies (Pre and Post-Bronchodilator). *Natl. Heal. Nutr. Exam. Surv.* 2011-12 Hyattsville, MD (USA); 2014.
 13. Centers for Disease Control and Prevention (CDC); National Center for Health Statistics (NCHS). Respiratory Health Spirometry Procedures Manual. *Natl. Heal. Nutr. Exam. Surv.* 2011-12 Hyattsville, MD (USA); 2011.
 14. Centers for Disease Control and Prevention (CDC); National Center for Health Statistics

- (NCHS). Data Documentation, Codebook, and Frequencies (Spirometry - Raw curve data). Natl. Heal. Nutr. Exam. Surv. 2011-12 Hyattsville, MD (USA); 2014.
15. Klein B, Klein B. NumPy. *Einführung Python 3* 2014.
 16. Chollet F. Keras: The Python Deep Learning library. *Keras.io* 2015; .
 17. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 2010.
 18. Shapley LS. A value for n-person games. *The Shapley value* 2009.
 19. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* 2017.
 20. R Core Team. R: A language and environment for statistical computing. <http://www.R-project.org/>. R Found. Stat. Comput. Vienna, Austria 2017.
 21. Culver BH, Graham BL, Coates AL, Wanger J, Berry CE, Clarke PK, Hallstrand TS, Hankinson JL, Kaminsky DA, MacIntyre NR, McCormack MC, Rosenfeld M, Stanojevic S, Weiner DJ. Recommendations for a standardized pulmonary function report. An official American Thoracic Society technical statement. *Am. J. Respir. Crit. Care Med.* 2017; .
 22. Ramos Hernández C, Núñez Fernández M, Pallares Sanmartín A, Mouronte Roibas C, Cerdeira Domínguez L, Botana Rial MI, Blanco Cid N, Fernández Villar A. Validation of the portable Air-Smart Spirometer. *PLoS One* 2018; .
 23. Burgos F, Disdier C, De Santamaria EL, Galdiz B, Roger N, Rivera ML, Hervàs R, Durán-Tauleria E, Garcia-Aymerich J, Roca J. Telemedicine enhances quality of forced spirometry in primary care. *Eur. Respir. J.* 2012; .

24. National Institute for Occupational Safety and Health (NIOSH). Spirometry Quality Assurance: Common Errors and Their Impact on Test Results. Natl. Inst. Occup. Saf. Heal. 2012.
25. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI-Explainable artificial intelligence. *Sci. Robot.* 2019; .
26. O'Mahony N, Campbell S, Carvalho A, Harapanahalli S, Hernandez GV, Krpalkova L, Riordan D, Walsh J. Deep Learning vs. Traditional Computer Vision. *Adv. Intell. Syst. Comput.* 2020.
27. LeCun Y, Cortes C, Bottou L, Jackel L. Comparison of Learning Algorithms for Handwriting Digit Recognition. *Int. Conf. Artif. Neural Networks* 1995; .

Tables

Table 1

Baseline characteristics of spirometry curves and their effort-ratings in NHANES 2011-12.

Effort rating	A (N=19,828)	B (N=5,890)	C (N=8,605)	D (N=2,550)	p
Rating description	All curve quality attributes were acceptable	Large time to PEF or a non-repeatable PEF	Less than 6 s of exhalation or no plateau	Cough or large extrapolated volume	
EOP (%)	97	96	8	49	<0.001
tFE > 6s (%)	91	87	10	45	<0.001
BEV criteria (%)	100	99	99	44	<0.001
tPEF (s)	0.07 ± 0.02	0.13 ± 0.07	0.10 ± 0.07	0.21	<0.001

Abbreviations: NHANES 2011-12=National health and nutritional examination survey in the United States in the year 2011-12; EOP=Existence of plateau in volume-time curve defined as no change in volume (≤ 0.025 L) for at least 1s of expiration; tFE > 6s= time of forced expiration, defined as difference between time at residual volume and time 0 obtained by back-extrapolation, is greater than 6 seconds; BEV criteria=back-extrapolated volume is lesser than maximum of 5% of FVC or 0.15 L; tPEF=time to peak flow from time 0

Table 2

The performance of CNN and a rule-based ATS/ERS algorithm at predicting manoeuvre acceptability and usability in the test set (N=3,738 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

Abbreviations: CNN=Convolutional neural network; ATS/ERS= American thoracic society/European respiratory society; PPV=Positive predictive value; NPV=Negative predictive value

Figure legends

Figure 1

Title: Processing of input

Legend: Flow-volume data is converted into 32x32 pixel matrices of maximal expiratory flow-volume loop (MEFVC, supplement). The white and the black regions contain a pixel value of 1 and 0 respectively. Further, American thoracic society/European respiratory society (ATS/ERS) recommendations are used to compute criteria (True/False) associated with back-extrapolated volume (BEV), time of forced expiration > 6s (tFE>6s) and existence of plateau (EOP) from volume-time curve, along with time to peak flow (tPEF).

Figure 2

Title: Convolutional neural network architecture

Legend: The pixel matrix of maximal expiratory flow-volume loop (MEFVC) is passed through a convolutional layer composed of three convolutional blocks (Supplement). The convolved output, that represents features from the MEFVC, along with back-extrapolated volume (BEV) criteria, time of forced expiration > 6s (tFE>6s), existence of plateau (EOP) and time to peak flow (tPEF), are fed into a simple neural network. The final output is the probability of manoeuvre acceptability, which if greater than 0.5 is considered acceptable.

Figure 3

Title: Interpreting the convolutional neural network (CNN)

Legend: The evidence, as quantified by Shapley values, from maximal expiratory flow volume from maximum inspiration until 1s after exhalation begins (MEFVC<1s), MEFVC after 1s of exhalation (MEFVC>1s), back-extrapolated volume (BEV) criteria, time of forced expiration > 6s (tFE>6s), existence of plateau (EOP) and time to peak flow (tPEF) in examples

of curves predicted as (a) acceptable with satisfactory start and end of test ($P(\text{acceptability})=0.90$), (b) unacceptable but usable ($P(\text{acceptability})=0.25$, $P(\text{usability})=0.91$) and (c) unacceptable and unusable ($P(\text{acceptability})=0.02$, $P(\text{usability})=0.16$).

Abbreviations: $P(\text{acceptability})$ =Probability of acceptability, $P(\text{usability})$ = Probability of usability

Figure 4

Title: Global interpretability

Legend: Forrest plots summarising evidence, quantified by Shapley values, from maximal expiratory flow volume curve pattern within first second of exhalation ($\text{MEFVC}<1\text{s}$), MEFVC after first second of exhalation ($\text{MEFVC}>1\text{s}$), back-extrapolated volume (BEV) criteria, time of forced expiration $> 6\text{s}$ ($t_{\text{FE}}>6\text{s}$), existence of plateau (EOP) and time to peak flow (t_{PEF}), in determining curve acceptability (a) and usability (b) in test set ($N=3,738$ curves).

32 by 32 pixel matrix

- BEV criteria
- EOP
- tFE>6s
- tPEF (s)

Convolution block

Convolution block

Convolution block

Convolutional layer

Convolutional layer output

BEV criteria

EOP

tFE > 6s

tPEF (s)

Manoeuvre acceptability

(a)

(b)

(c)

S1-Model development

This section contains additional information on the model development section

Processing raw data

The raw curve data (raw_data) in NHANES 2011-12 was available as a sequence of numeric values representing change in volume in millilitres over a 0.01-second interval during forced expiration. We obtain the time, volume and flow using python library numpy (np) as follows[1]:

```
time = np.arange(0, len(raw_data))*0.01
volume = (np.cumsum(raw_data))/1000
flow= np.diff(vol)/ np.diff(time)
```

where np.arrange (returns evenly spaced values in an interval), np.cumsum (cumulative sum) and np.diff (first order difference) are numpy-based functions. Here, time, volume and flow are 1D arrays

Once the data was processed, we removed any trailing forced inspiratory signal at the end of forced expiration.

Generating pixel matrixes

We generated the pixel matrices of the maximal expiratory flow-volume loop (MEFC) with an aspect ratio of two units of flow for each unit of volume in accordance with the display guidelines of ATSERS for flow-volume curves. We do this as follows:

We first determine the maximum range (yf) to be represented in the pixel matrix

```
yf = max(flow) - min(flow)
```

if $yf < 2 \times (\max(\text{volume}) - \min(\text{volume}))$, then $yf = 2 \times (\max(\text{volume}) - \min(\text{volume}))$

Then, we initialise an identity matrix of dimension pxl x pxl. In our study, pxl=32

```
pxl_matrix = np.ones((pxl,pxl))
```

Then, we map the flows and volumes to corresponding pixel location

```
I_pxl = (2*((pxl-1)/yf) * volume)
```

```
J_pxl = (((pxl-1)/yf) * flow)
```

```
pxl_matrix [J_pxl, I_pxl] = 0
```

Convolutional neural network (CNN) architecture

We use a convolutional neural network that takes the 32x32 MEFVC matrix as input. We first describe the building blocks of our model[2].

Building blocks

a) Convolutional block

A convolutional block (Figure S1) consists of the following three operations in sequence:

- i) Convolution (CONV): A CONV layer consists of a set of learnable filters, which are spatially small but extends through the depth of the input volume. For example, a typical filter on the first convolutional block may have a size of 3x3x1 (i.e. 3 pixels width and height and 1 because the images have a single channel). During a forward pass, we slide the filter over the width and height of the input volume and produce a 2-dimensional activation map that gives the responses of that filter at every spatial position. Similarly, a set of K filters will produce K two-dimensional activation maps, which we stack along the depth dimension and produce the output volume. Intuitively, we can interpret that the network will learn a set of K filters that activate when they see some type of visual feature such as an edge of some orientation on the first layer, or eventually entire honeycomb or wheel-like patterns on later blocks/layers of the network.

Mathematically, a discrete 2D convolution operation can be written as follows:

$$y[m,n,d] = \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} f[u,v] \cdot k_d[u-m, v-n] + b_d, d = 1,2..K_i$$

where $y[m,n, d]$ is the result of the convolutional operation at column m and row n for the d^{th} filter, $f[i, j]$ represents a depth slice at column i and row j of the input volume, k_d is called a filter or kernel which is a square weight matrix of size F_i that is translated over the input volume and b_d is a bias. To summarize, a CONV layer at the i^{th} convolutional block:

- Accepts an input volume X_{1i} of size $W_{1i} \times H_{1i} \times D_{1i}$
- Performs a 2D convolution operation with the hyper-parameters some of which are fixed by us:
 - Number of filters (K_i)
 - The filter width and height ($F_i = 3$)
 - The filter stride ($S_i = 1$)
 - The amount of zero padding ($P_i = 0$)

Results in a total of $(F_i \cdot F_i \cdot D_{1i}) \cdot K_i$ weights and K_i biases

- Produces an output volume X_{2i} of size $W_{2i} \times H_{2i} \times D_{2i}$ where
 - $W_{2i} = (W_{1i} - F_i + 2P_i)/S_i + 1$
 - $H_{2i} = (H_{1i} - F_i + 2P_i)/S_i + 1$
 - $D_{2i} = K_i$

- ii) Rectified linear unit (ReLU): The output volume from the previous step is passed through the ReLU activation function that introduces non-linearity as follows:

$$y = \max(0, x)$$

In other words, the activation is simply thresholded at zero. We can summarize the computations in this step as follows:

- Accepts X_{2i} as input
- Applies the ReLU activation function
- Produces an output volume X_{3i} of size $W_{2i} \times H_{2i} \times D_{2i}$

- iii) Max pooling (MAXPOOL): The function of the pooling layer is to progressively reduce the spatial size of the representation to reduce the amount of parameters and computation in the network, and hence to also control overfitting. The MAXPOOL Layer operates by considering a $MP_i \times MP_i$ region with a stride SP_i over a given 2D slice and outputs a single value, which is the maximum in that region. It operates independently on every depth slice of the input and it has no learnable parameters. We can summarize the computations in this layer as follows:

- Accepts X_{3i} as input
- Performs MAXPOOL that requires the hyper-parameters:
 - The spatial extent ($MP_i = 2$)
 - The stride ($SP_i = 2$)
- Produces an output volume X_{4i} of size $W_{3i} \times H_{3i} \times D_{3i}$ where
 - $W_{3i} = (W_{3i} - MP_i)/SP_i + 1$
 - $H_{3i} = (H_{3i} - MP_i)/SP_i + 1$
 - $D_{3i} = D_{2i}$

b) Fully connected (FC) network (neural network)

A FC network (Figure 6) consists of a standard feed-forward neural network with one hidden layer and a single neuron output layer. The activation of the hidden layer 1 can be expressed as follows:

$$A^1 = \text{ReLU}(W^1 \cdot A^0 + B^1)$$

where W^1 is a $n^1 \times n^0$ weight matrix where n^1 denotes the number of neurons in layer 1, A^0 are the activations from the previous layer of length $n^0 \times 1$ which in our case would be a flattened output volume, B^1 is the bias and ReLU is an activation function defined as:

$$\text{ReLU}(x) = \max(0, x)$$

The computation in the final layer is written as:

$$\hat{Y} = \text{sigmoid}(W^f \cdot A^1 + B^f)$$

where $\hat{Y}$ is a scalar value between 0 and 1, W^f is a weight matrix of size $1 \times n^1$ and B^f is the single bias term in the final layer. The sigmoid function is defined as:

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

To summarize, the FC network:

- Accepts a vector of length $n^0 \times 1$
- Performs a feed-forward computation involving the hyperparameters:
 - The number of hidden layer neurons n^1
 Results in a total of $(n^0 \cdot n^1 + n^1)$ weights and $B^1 + 1$ biases
- Produces a scalar value $\hat{Y}$

Figure S2 Activations and weight matrices
W in a FC network

Model training

The cross-entropy error (Loss) was used to quantify the loss and the cost function J was described as the average cross-entropy error over the training set. Mathematically,

$$Loss(Y_i, \hat{Y}_i) = -\{Y \log \hat{Y} + (1 - Y) \log (1 - \hat{Y})\}$$

$$J(W) = \frac{\sum_{i=1}^{m_{train}} Loss(Y_i, \hat{Y}_i)}{N_{train}} + \lambda \sum W.W$$

In the above equation, Y_i refers to the label of the i^{th} training example which is either 1 (acceptable) or 0 (rejected), $\hat{Y}_i$ refers to the probability calculated by the model corresponding i^{th} training example, W represents the weights of the entire network and λ is the L2 regularization parameter used to prevent overfitting. W is updated using gradient descent:

$$W := W - \alpha \frac{\partial J}{\partial W}$$

where α is the learn rate. In our case, we use batch gradient descent with a batch size of n_{batch} with Adam optimization scheme[3] to train the model. The number of epochs used are n_{epochs} resulting in a total of $(m_{train}/n_{batch}) \times n_{epochs}$ iterations.

Hyperparameter tuning

Parameters that are not calculated by gradient descent are called hyperparameters. The following hyper-parameters were considered for optimization in our study:

- The learn rate (α)
- Batch size (n_{batch})
- Number of epochs (n_{epochs})
- L2 regularization (λ)
- Dropout rate between flattened volume and first hidden layer (dr)
- No of filters in the first convolution block of both branches (K_1)
- No of filters in the second convolution block of both branches (K_2)
- No of filters in the third convolution block of both branches (K_3)

For each set of hyper-parameters, the model is trained and prediction probabilities are calculated on the validation set. A threshold of 0.5 is used to assign a prediction label i.e. a manoeuvre with a probability of greater than 0.5 is assigned an acceptable label (1) and with a probability of lesser than 0.5 is assigned a rejected label (0).

The cost function J_{hp} for optimizing the hyper-parameters of the CNN is an average of validation sensitivity and specificity.

$$Sensitivity = \frac{TP}{TP + FN}, \quad Specificity = \frac{TN}{TN + FP},$$

$$J_{hp} = -0.5 (Sensitivity + Specificity)$$

where TP is true positives, FN is false negatives, TN is true negatives and FP is false positives. We using a Bayesian search scheme [4] with 100 evaluations to solve the following optimization problem:

$$\min_{\alpha, n_{batch}, n_{epochs}, \lambda, dr, F_1, F_2, F_3} J_{hp}(\alpha, n_{batch}, n_{epochs}, \lambda, dr, F_1, F_2, F_3)$$

Model interpretability

A direct interpretation of CNN is difficult; therefore we used a game-theoretic approach called Shapley values to interpret our model's output[5]. The Shapley value of a feature value is its

contribution to the output, weighted and summed over all possible feature value combinations. The formula for Shapley value is:

$$\Phi_j = \sum_{S \in \{x_1, \dots, x_p\} \setminus \{x_j\}} \frac{|S|! (p - |S| - 1)!}{p!} (f(S \cup \{x_j\}) - f(S))$$

Where $x_1, \dots, x_p$ represents the set of p input features for function f , x_j is the desired feature, Φ_j is the Shapley value of feature j , S is a combination of inputs without including x_j (done by replacing value of x_j with a background value from the dataset), $S \cup \{x_j\}$ includes feature x_j

To calculate Shapley values in our study, we first divided the MEFVC pixel matrix into 2 regions: MEFVC<1s containing the MEFVC pixels from the point of maximum inspiration until 1 second after time 0, and MEFVC>1s containing the rest remaining curve. This division can be observed in figure s3 where a vertical line on the 20th pixel divides the matrix into two regions. Then using a technique called Shapley additive explanation (SHAP) [6], which allows a faster but approximate Shapley value calculation, we estimated how MEFVC<1s, MEFVC>1s, BEV criteria, tFE>6s, EOP and tPEF contributed to the model's output. To simulate the effect of background value, we replaced these inputs by randomly drawing instances from our training dataset. For MEFVC<1s and MEFVC>1s, we replaced all the pixels for corresponding regions from a randomly drawn pixel matrix. In this example, this would amount to replacing all pixels between vertical lines 0 and 20 for MEFVC<1s and between vertical lines 20 and 32 for MEFVC>1s.

References

1. Klein B, Klein B. NumPy. *Einführung Python 3* 2014.
2. LeCun Y, Bengio Y, Hinton G, Y. L, Y. B, G. H. Deep learning. *Nature* 2015; 521: 436–444.
3. Kingma DP, Ba JL. Adam: a Method for Stochastic Optimization. *Int. Conf. Learn. Represent. 2015* 2015; : 1–15.
4. Bergstra J, Yamins D, Cox DD. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. *12th PYTHON Sci. CONF. (SCIPY 2013)* [Internet] 2013; : 13–20 Available from:

<http://hyperopt.github.io/hyperopt/%5Cnhttps://github.com/jaberg/hyperopt%5Cnhttp://www.youtube.com/watch?v=Mp1xnPfE4PY>.

5. Shapley LS. A value for n-person games. *The Shapley value* 2009.
6. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* 2017.

Supplement figure S1

Title: Receiver operating characteristics (ROC) analysis

Legend: ROC curves for convolutional neural network (CNN) and ATS/ERS rule-based criteria in determining (a) acceptability and (b) usability in the test set (N=3,738 curves). The area under curve (AUC) for CNN was significantly higher ($p < 0.0001$) in both the cases.

Supplement figure S2

Title: Global interpretability in sub-groups

Legend: Box plots summarising evidence for acceptability, 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 > 6s (tFE>6s), existence of plateau (EOP) and time to peak flow (tPEF), in a sub-group of curves predicted as (a) acceptable (N=2021), (b) discarded (N=548) and (c) unacceptable but usable (N=1211), in the test set.

(a)

(b)

(a)

(b)

(c)

Supplement Table S1

Distribution of best FEV₁ and FEV₁/FVC after applying ATS/ERS repeatability criteria in curves accepted by gold-standard evaluation (visual evaluation by technicians), CNN and ATS/ERS rules-based model.

	Gold-standard	CNN	ATS/ERS
FEV₁ (L)	2.76±0.83	2.76±0.84 (p=0.98)	2.77±0.84 (p=0.08)
FEV₁ LLN (L)	1.39	1.39	1.39
FEV₁/FVC (%)	82.05 ±7.67	82.07 ±7.73 (p=0.68)	81.77±7.68 (p<0.01)
FEV₁/FVC LLN (%)	69.43	69.35	69.13

Abbreviations: CNN=Convolutional neural network; ATS/ERS= American thoracic society/European respiratory society; LLN=Lower limit of normal (Mean-1.645 SD); FEV₁=Forced expiratory volume in 1s; FVC=Forced vital capacity