ERJ open research THE BEST IN OPEN ACCESS BASIC, TRANSLATIONAL & CLINICAL RESPIRATORY RESEARCH

Early View

Original research article

AI-based software facilitates spirometry quality control in asthma and COPD clinical trials

Eva Topole, Sonia Biondaro, Isabella Montagna, Sandrine Corre, Massimo Corradi, Sanja Stanojevic, Brian Graham, Nilakash Das, Kevin Ray, Marko Topalovic

Please cite this article as: Topole E, Biondaro S, Montagna I, *et al*. AI-based software facilitates spirometry quality control in asthma and COPD clinical trials. *ERJ Open Res* 2022; in press (https://doi.org/10.1183/23120541.00292-2022).

TITLE

AI-based software facilitates spirometry quality control in asthma and COPD clinical trials

AUTHORS

Eva Topole[1], Sonia Biondaro[1], Isabella Montagna[1], Sandrine Corre[1], Massimo Corradi[2], Sanja Stanojevic[3], Brian Graham[4], Nilakash Das[5,6], Kevin Ray[6], Marko Topalovic[6]

AUTHOR AFFILIATIONS

[1] – Global Clinical Development, Chiesi Farmaceutici, S.p.A., Parma, Italy

[2] – Department of Medicine and Surgery, University of Parma, Italy

[3] – Department of Community Health and Epidemiology; Dalhousie University, Nova Scotia, Canada

[4] – Division of Respirology, Critical Care and Sleep Medicine, University of Saskatchewan, Canada

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

[6] – ArtiQ NV, Leuven, Belgium

CORRESPONDING AUTHOR:

Marko Topalovic, ArtiQ, marko.topalovic@artiq.eu

TAKE HOME MESSAGE:

In clinical trials, AI software can be used with high accuracy to evaluate the quality of spirometry data. This leads to increased consistency and repeatability and immediate feedback allowing a real time evaluation with the subject still at the site.

ABSTRACT

Rationale

Acquiring high-quality spirometry data in clinical trials is important, particularly when using FEV1 or FVC as primary endpoints. In addition to quantitative criteria, the ATS/ERS standards include subjective evaluation which introduces inter-rater variability and potential mistakes. We explored the value of AI-based software (ArtiQ.QC) to assess spirometry quality and compared it to traditional over-reading control.

Methods

A random sample of 2000 sessions (8258 curves) was selected from Chiesi COPD and Asthma trials (N=1000 per disease). Acceptability using the 2005 ATS/ERS standards was determined by over-reader review and by ArtiQ.QC. Additionally, three respiratory physicians jointly reviewed a subset of curves (N=150).

Results

The majority of curves (N=7267, 88%) were of good quality. The AI agreed with over-readers in 91% of cases, with 97% sensitivity and 93% positive predictive value. Performance was significantly better in the asthma group. In the revised subset, N=50 curves were repeated to assess intra-rater reliability, (Kappa: 0.83, 0.86 and 0.80). All reviewers agreed on 63% of 100 unique tests (Kappa = 0.5). When reviewers set the consensus (gold-standard), individual agreement with it was 88%, 94% and 70%. The agreement between AI and “gold-standard” was 73%, over reader agreement was 46%.

Conclusion

AI-based software can be used to measure spirometry data quality with comparable accuracy as experts. The assessment is a subjective exercise, with intra- and inter-rater variability even when the criteria are defined very precisely and objectively. By providing consistent results and immediate feedback to the sites, AI may benefit clinical trial conduct and variability reduction.

WORD COUNT ABSTRACT: 250 / 250

INTRODUCTION

Measures of lung function, typically using spirometry, are critical clinical outcomes used in respiratory clinical trials[1, 2]. The efficacy of treatments targeting the lungs is often assessed through two main spirometry derived parameters: the forced expiratory volume in first second (FEV_1) and the forced vital capacity (FVC)[3–5]. Since both of these indices are dependent on a maximal forced expiratory effort, spirometry requires skilled technicians to ensure the subjects' compliance and optimal performance of the test. Therefore, in the context of clinical trials, ensuring and assessing the quality of spirometry is of paramount importance to guarantee that the measured FEV_1 and FVC values are sufficiently reliable to assess the efficacy of a therapeutic.

To ensure valid data and reduce the variability in the clinical trials, it is normal to use central spirometry that incorporates standardized instrumentation, site training and centralised review (over reading (OR)) of all spirometry sessions [6, 7]. The OR is performed by independent reviewers who are registered respiratory experts trained to review spirometry in a clinical trial environment where the objective is to ensure consistent results across all patients, sites, and timepoints. These specialists typically use the criteria defined by an international task force appointed by the American Thoracic Society (ATS) and European Respiratory Society (ERS) to evaluate spirometry data quality, with an average turnaround time of 24-48 hrs for feedback to the investigational sites[8, 9]. This delay between subjects performing the spirometry session and the investigational site receiving feedback can delay study-specific decisions, such as subject randomisation, treatment changes, and may lead to repeat visits with additional spirometry sessions to be performed by the subjects, increasing burden on participants.

The ATS/ERS spirometry quality standards define both quantitative and qualitative criteria to ensure high quality standards for spirometry measurements. Quantitative criteria (e.g. expiration duration longer than a minimum time period, or back-extrapolated volume below a certain threshold) can be evaluated relatively easily by software. However, the qualitative components (e.g. artefact detection) require subjective assessment by the technicians acquiring the data[10–12]. Technician training, experience and other factors can lead to heterogeneity of the assessment[13]. Agreement between two reviewers assessing the same data has been measured as low as 52% in previous studies[14–17]. This inter-rater variability is important to reduce, especially in a clinical trial setting where spirometry data are being used for measuring the therapeutic effect of a pharmaceutical intervention.

In this investigation, we explored whether an artificial intelligence (AI) based software (ArtiQ.QC) can perform spirometry data quality assessment with an accuracy at least as high as manual over reading, bringing additional benefits of a shorter turnaround time and improved consistency.

MATERIALS AND METHODS

Study Objectives

The primary objective of the study was to determine whether an AI based software (ArtiQ.QC) can be used to perform spirometry data quality assessment with similar accuracy as expert over readers (OR). Secondary objectives were to assess the level of inter-rater variability and intra-rater reliability in spirometry data quality assessment, determine to what extent a single reviewer' judgment can be considered as a gold standard comparator, and determine whether AI based software (ArtiQ.QC) may outperform the over reader's quality assessment.

Data

Spirometry data from past clinical trials sponsored by Chiesi in the therapeutic areas of asthma and COPD were used for analysis[18–20]. These studies were registered with ClinicalTrials.gov, numbers NCT02579850, NCT02676076, and NCT02676089. Using a function for random selection of data, a completely randomly chosen sample of 2000 sessions was selected for analysis, with 1000 sessions from asthma and COPD subjects each. A total of 8258 spirometry curves (N=4085 COPD, N=4173 asthma) were used. All data selected for analysis were collected and assessed using the ATS/ERS 2005 spirometry standards[9]. Each curve was labelled by an expert over reader at the time of data collection according to the over-reader guidelines for each clinical trial, with possible outcomes of "Acceptable" or "Unacceptable". Over reader guidelines were aligned with the ATS/ERS 2005 standards, and stated that if a curve was "Usable" according to the ATS/ERS 2005 standards (i.e. have an acceptable start of test and are free from artefact, such as a cough) then the OR label should be "Acceptable". The whole session acceptability evaluation is not reported within this manuscript.

Study Design

In this non-interventional retrospective study, the AI based software (ArtiQ.QC, ArtiQ NV, Leuven, Belgium) was used to assess the quality of spirometry data using the ATS/ERS 2005 standards[9]. ArtiQ.QC is an artificial intelligence-based software which can apply the quantitative and qualitative criteria defined in the ATS/ERS 2005 standards to assess spirometry data quality. The qualitative criteria (i.e. artefact detection) are assessed using a previously published deep learning based approach[21]. ArtiQ.QC is a fully validated software, and in this study was used by Chiesi for testing in a clinical trial context.

An overview of the study design is shown in Figure 1. Firstly, ArtiQ.QC was used to assess data quality in all 8258 curves, and the performance of ArtiQ.QC was measured (accuracy, sensitivity, specificity, positive predictive value [PPV] and negative predictive value [NPV]) using the original OR labels as comparators. Secondly, ArtiQ.QC was recalibrated (without altering the AI) on 80% of the data selected for analysis (N=1600 sessions, equally split between asthma and COPD) and the performance of the recalibrated version determined on the remaining 20% of the data. The original OR labels were again used as comparators in this second performance evaluation to evaluate the extent to which ArtiQ.QC performance could be improved.

Subsequently, a selection of curves was manually assessed by three reviewers to determine "gold standard" labels (N=150 curves total, with N=100 unique curves, equally split between diseases; full breakdown of curves selected for review is provided in the Supplementary Material). Initially all curves were assessed separately by each reviewer. All curves where there was disagreement between the reviewers' applied labels were jointly assessed to eventually agree on a "gold standard" label for each curve. Reviewers were blinded to the original over reader and ArtiQ.QC labels, and used the criteria defined in the ATS/ERS 2005 standards to assess the spirometry data quality. Inter-rater variability and intra-rater reliability were measured based on reviewer labels. Finally, ArtiQ.QC and original over-reader labels were both compared to the "gold standard" labels provided by the joint consensus of reviewers.

Ensuring a fair comparison of ArtiQ.QC to Over Readers

Individual spirometry curves in the dataset used for analysis were labelled as either "Acceptable" or "Unacceptable" by the over readers. ArtiQ.QC generates labels for individual spirometry curves in accordance with the ATS/ERS 2005 standards, which permits "curves that have an acceptable start of test and are free from artefact, such as a cough" to be "usable curves" for the measurement of FEV_1 and FVC. For the purpose of comparing ArtiQ.QC and OR labels, an ArtiQ.QC label of "Usable" was converted to "Acceptable", consistent with the rules followed by over readers in the data source studies.

Analysis

A comparison of proportions test was used to evaluate whether significant differences in the baseline dataset characteristics existed. For performance evaluation of ArtiQ.QC, a 2x2 confusion matrix of labels assigned by ArtiQ.QC or original OR at a curve-level was created. Accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were measured from the confusion matrix. McNemar's test was used to test whether recalibration significantly improved performance of ArtiQ.QC. Cohen's Kappa was used to assess intra-rater reliability, and Fleiss's Kappa (the multi-rater generalisation of Cohen's Kappa) was used to assess inter-rater variability from the manual review. Accuracy, sensitivity, specificity, PPV and NPV were calculated for ArtiQ.QC and original OR labels using the expert reviewer labels as comparator. Performance of ArtiQ.QC and original OR was compared using McNemar's test.

RESULTS

Dataset Characteristics

The vast majority of curves (N=7267, 88%) were labelled as "Acceptable" by the original over reader assessment. More curves were labelled "Acceptable" in the asthma sub-group (N=3818/4173, 91.5%) compared to the COPD sub-group (N=3391/4085, 83%; $p < 0.0001$ comparison of proportions test).

Performance of AI software using original over reader labels as comparator

A summary of the performance of the AI based software using original over reader labels as the comparator is shown in Table 1. From all N=8258 curves, the AI software agreed with over-readers in 87% of cases, with 93% sensitivity and 92% positive predictive value (PPV). Performance was significantly better in the asthma sub-group (accuracy 90%, N=4173 curves) than the COPD sub-group (accuracy 84%, N=4085 curves; $p < 0.0001$ comparison of proportions test).

Performance of recalibrated AI software using original over reader labels as comparator

Tables 2 and 3 summarise the performance evaluation results of the recalibrated and baseline AI software, respectively, using original over reader labels as a comparator. Following recalibration of the AI software in 80% of the full dataset, 67 curves in the 20% test set changed label compared to the quality label assigned with the standard AI software. This resulted in a significant improvement in performance in the test data set ($p<0.0001$, McNemar's test). The parameters adjusted in the recalibration were the artefact probability and hesitation volume thresholds (see Supplementary Material for full details). Overall recalibrated accuracy was 91%, with sensitivity 97% and PPV 93% (N=1659 curves), compared to baseline accuracy of 87%, sensitivity 92%, and PPV 93%. The greatest improvement in performance was in the asthma sub-group (increase in accuracy from 89% to 94% in N=841 curves following recalibration; $p<0.001$ McNemar's test). A non-significant improvement in COPD performance was also seen (accuracy improved from 84% to 86% in N=818 curves; $p=0.13$ McNemar's test).

Manual Review: Reviewer Reliability and Variability

A set of N=150 curves were additionally manually reviewed by three reviewers. N=50 of these curves were repeated in a random fashion in order to assess intra-rater reliability. Cohen's Kappa values for the three reviewers were 0.83, 0.86 and 0.80, indicating that though reviewers mainly provided consistent labels when presented with the same data twice, some disagreement with their own evaluation was present.

From the N=100 unique curves, all reviewers agreed after their separate review sessions in 63% of curves (inter-rater Fleiss's Kappa = 0.5), indicating disagreement between reviewers of a magnitude consistent with previous studies[14–17]. Agreement between reviewers was higher in the subset of curves where the AI software and original over reading provided the same label (18/25, 75%, Fleiss's Kappa = 0.63) than in the subset where the AI software and original over reader labels differed (45/75, 60%, Fleiss's Kappa = 0.45). Despite using a set of objective criteria defined by the ATS/ERS 2005 standards, this result indicates that inter-rater disagreement exists between expert reviewers.

Performance of recalibrated AI software, original over readers, and each reviewer using "gold standard" labels as comparator

The N=37 curves where disagreement existed between reviewers were jointly reviewed to derive labels for all N=100 curves that all reviewers agreed on ("gold standard" labels). The percentage agreement with "gold standard" labels and Cohen's Kappa values for each reviewer, the reviewers' majority opinion, AI based software and the original over reader labels are shown in Table 4. Note that the lowest possible value for percentage agreement with "gold standard" labels for any individual reviewer is 63%, since 63/100 curves had the same label from the separate reviews. A percentage agreement of 89% for the majority opinion of reviewers from their separate review sessions indicates that the "gold standard" label was opposite to the reviewer majority opinion in 11 curves. Unsurprisingly, the agreement of individual reviewers with the "gold standard" labels is high since it is the reviewers' joint opinion that defines the "gold standard". The agreement between AI based software and “gold standard” label (73%) is similar to Reviewer 3 (70%). This agreement is also far superior to the original over reader labels (46%) in this subset of curves, though the high inter-rater variability in this subset may not generalise to the whole dataset.

DISCUSSION

This study demonstrates that AI based software can be used to measure spirometry data quality with similar accuracy as expert over readers. Additionally, this analysis confirmed that assessment of spirometry quality is a subjective exercise, with intra- and inter-rater variability even when the criteria are defined very precisely and objectively. When using the joint opinion of three expert reviewers as a “gold standard”, AI based software has a much higher agreement with the gold standard labels than original over reader labels. Additionally, by providing consistent results and immediate feedback to the sites, using ArtiQ.QC may benefit clinical trial conduct and reduce the variability in outcomes, though these benefits have not been considered in this retrospective study.

Assessing the performance of AI software using the original over reader labels as a comparator assumes that the original over reader labels are 100% accurate. However, the existence of inter- and intra-rater variability observed in this and other studies[1–5] suggests that this assumption is unlikely to be correct. Even when comparing three very experienced reviewers, and the exact criteria from the 2005 ATS/ERS standards were used to derive quality scores for each curve, disagreement between reviewers was still present in over a third of all 100 unique curves reviewed in this study.

The inter- and intra-rater variability present when spirometry data are manually reviewed means that a single reviewer’s opinion is not a reliable comparator for assessing the accuracy or reliability of an automated tool. Whilst the inter-rater agreement measured in this study was consistent with previous reports, this agreement may be influenced by experience, training and expertise of the reviewers[22–24]. This is reflective of the real-world scenario, however, where usually an only single reviewer assesses spirometry data quality. The joint opinion of the three reviewers may be considered a viable gold standard, but this approach is not realistic for implementation in clinical trials. Therefore, for AI based software to be useful its performance simply needs to match that of another manual reviewer. With percentage agreement of 73% with the “gold standard” labels in the subset of curves manually reviewed, the AI software performance is similar to reviewer 3 in this study. Compared to the original over reader labels (percentage agreement with the “gold standard” of 46%), it may be the case that AI software such as ArtiQ.QC could outperform over readers in assessing the quality of spirometry data in clinical trials.

Limitations

This study was conducted using legacy spirometry data acquired during the two clinical trials (one in asthma and one in COPD). The over-reading services for the two studies were provided by two different vendors potentially causing a difference in the level of agreement between AI algorithm and OR for each study. Additionally, algorithm was initially developed and trained on a diverse set of curves (from healthy subjects, different respiratory and other internal diseases), potentially leading to a lower accuracy than with the algorithm that would have been developed and trained on a set of curves within specific therapeutic area. Further work is required to explore the generalisability of these results to other disease areas, paediatric population and non-clinical trial settings[25, 26].

This study compares the performance value of the software that is trained to mimic experts and experts themselves, once the complete spirometry session is performed. However, the largest practical impact on the running clinical trials will be if the software would be used after each spirometry blow. This would reject immediately initial bad blows and help subjects and technicians to secure the acceptable quality of spirometry by the end of the session.

This study was conducted using data acquired using the 2005 ATS/ERS standards, and with over-reading performed according to those standards. New ATS/ERS standards have been published in 2019 which offer more guidance in quality assessment and change the way in which spirometry data must be acquired explicitly recommending a maximal inspiration following forced expiration[8]. This study serves as proof of concept that the AI methodology can be tailored to follow the 2005 spirometry standards to assess spirometry acceptability. Although the AI algorithms will be updated to comply with the revisions in the 2019 spirometry standards, the proven underlying AI methodology remains unchanged. With the new standards and AI software working in parallel, inter- and intra-rater variability in spirometry data quality assessment may decrease; this is the subject of ongoing work.

Conclusion

With this retrospective analysis on a subset of asthma and COPD spirometry data acquired according to the 2005 ATS/ERS standards, we confirmed that AI software can be used with high accuracy to evaluate the quality of spirometry data in clinical trials. AI software such as ArtiQ.QC may offer an alternative approach, or assistance, to manual over reading with the main advantage that AI software can provide immediate feedback allowing a real time evaluation with the subject still at the site. Additional advantages that AI software brings include increased consistency and repeatability of spirometry data quality assessment because the AI model is more objective than manual over reading. Further evaluation and testing of this technology are needed to understand its practical implications and accuracy according to the most recent 2019 ATS/ERS standards and in disease areas other than asthma and COPD. Prospective real-time testing would also allow to explore and evaluate additional potential benefits of this technology on reducing patient burden and enhancing quality of data in clinical trials. Next to this, such AI algorithms may contribute to increased reliability and enhanced validity of epidemiological studies where spirometry data is collected.

AUTHOR CONTRIBUTIONS

Conception and study design: ET, SB, IM, SC, SS, BG, ND, KR, MT

Acquisition, analysis or interpretation: ET, SB, IM, SC, MC, SS, BG, ND, KR, MT

Manuscript preparation and critical revision: ET, SB, IM, SC, MC, SS, BG, ND, KR, MT

CONFLICT OF INTEREST

ET, SB, IM, SC are all employees of Chiesi Farmaceutici S.p.A.

KR was an employee of ArtiQ NV, Belgium at the time the study was performed. MT is a founder of ArtiQ NV, Belgium.

MC received research funds and professional fees from Chiesi Farmaceutici S.p.A.

BG received professional fees from Chiesi Farmaceutici S.p.A., MGC Diagnostics, Vyaire Medical, and the Lung Association of Saskatchewan. BG has a patent application underway related to pulmonary function testing.

FINANCIAL SUPPORT STATEMENT

This research was funded by Chiesi Farmaceutici S.p.A.

TABLES

	Overall	Asthma sub-group	COPD sub-group
Num. Curves	8258	4173	4085
Accuracy	87%	90%	84%
Sensitivity	93%	93%	94%
Specificity	35%	52%	25%
PPV	92%	96%	87%
NPV	41%	40%	43%

Table 1: *Performance evaluation of baseline ArtiQ.QC using original over reader labels as the comparator in the full data set.*

	Overall	Asthma sub-group	COPD sub-group
Num. Curves	1659	841	818
Accuracy	87%	89%	84%
Sensitivity	92%	92%	94%
Specificity	24%	43%	15%
PPV	93%	96%	87%
NPV	30%	29%	32%

Table 2: *Performance evaluation of baseline ArtiQ.QC using original over reader labels as the comparator in the 20% test data set.*

	Overall	Asthma sub-group	COPD sub-group
Num. Curves	1659	841	818
Accuracy	91%	94%	86%
Sensitivity	97%	98%	98%
Specificity	22%	40%	14%
PPV	93%	96%	87%
NPV	58%	59%	57%

Table 3: *Performance evaluation of recalibrated ArtiQ.QC using original over reader labels as the comparator in the 20% test data set.*

	% Agreement with Gold Standard	Cohen's Kappa
Reviewer 1	88	0.73
Reviewer 2	94	0.85

Reviewer 3	70	0.42
Majority Opinion of Reviewers	89	0.75
AI software	73	0.47
Original Over Reader	46	-0.08

Table 4: *Percentage agreement and Cohen's Kappa value for agreement with the "gold standard" label, which was defined using the joint opinion of the three reviewers, from N=100 spirometry curves.*

FIGURES

Figure 1: Study design and flow

REFERENCES

1. CHMP EMA. Guideline on clinical investigation of medicinal products in the treatment of chronic obstructive pulmonary disease (COPD). EMA Amsterdam; 2012. p. 1–17.

2. EMA C. Guideline on the clinical investigation of medicinal products for the treatment of asthma. EMA Amsterdam; 2015. p. 1–21.

3. Busse WW, Morgan WJ, Taggart V, Togias A. Asthma outcomes workshop: overview. *J. Allergy Clin. Immunol.* Elsevier; 2012; 129: S1–S8.

4. Glaab T, Vogelmeier C, Buhl R. Outcome measures in chronic obstructive pulmonary disease (COPD): strengths and limitations. *Respir. Res.* Springer; 2010; 11: 1–11.

5. Cazzola M, MacNee W, Martinez FJ, Rabe KF, Franciosi LG, Barnes PJ, Brusasco V, Burge PS, Calverley PMA, Celli BR. Outcomes for COPD pharmacological trials: from lung function to biomarkers. *Eur. Respir. J.* Eur Respiratory Soc; 2008; 31: 416–469.

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

7. Malmstrom K, Peszek I, Botto A, Lu S, Enright PL, Reiss TF. Quality assurance of asthma clinical trials. *Control. Clin. Trials* Elsevier; 2002; 23: 143–156.

8. Graham BL, Steenbruggen I, Miller MR, Barjaktarevic IZ, Cooper BG, Hall GL, Hallstrand TS, Kaminsky DA, McCarthy K, McCormack MC. Standardization of spirometry 2019 update. An official American thoracic society and European respiratory society technical statement. *Am. J. Respir. Crit. Care Med.* American Thoracic Society; 2019; 200: e70–e88.

9. Miller MR, Hankinson J, Brusasco V, Burgos F, Casaburi R, Coates A, Crapo R, Enright P vd, Van der Grinten CPM, Gustafsson P. Standardisation of spirometry. *Eur. Respir. J.* Eur Respiratory Soc; 2005; 26: 319–338.

10. Müller-Brandes C, Krämer U, Gappa M, Seitner-Sorge G, Hüls A, von Berg A, Hoffmann B, Schuster

A, Illi S, Wisbauer M. LUNOKID: can numerical American Thoracic Society/European Respiratory Society quality criteria replace visual inspection of spirometry? *Eur. Respir. J.* Eur Respiratory Soc; 2014; 43: 1347–1356.

11. Townsend MC. The American Thoracic Society/European Respiratory Society 2019 Spirometry Statement and Occupational Spirometry Testing in the United States. *Am. J. Respir. Crit. Care Med.* American Thoracic Society; 2020; 201: 1010–1011.

12. Beeckman-Wagner L-AF, Freeland D. Spirometry quality assurance; common errors and their impact on test results. 2012; .

13. Borg BM, Hartley MF, Bailey MJ, Thompson BR. Adherence to acceptability and repeatability criteria for spirometry in complex lung function laboratories. *Respir. Care* Respiratory Care; 2012; 57: 2032–2038.

14. 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* IEEE; 2017; 22: 276–284.

15. 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.* Eur Respiratory Soc; 2015; 45: 1283–1292.

16. Tan WC, Bourbeau J, O’donnell D, Aaron S, Maltais F, Marciniuk D, Hernandez P, Cowie R, Chapman K, Sonia Buist A. Quality Assurance of Spirometry in a population-based study–predictors of good outcome in spirometry testing. *COPD J. Chronic Obstr. Pulm. Dis.* Taylor & Francis; 2014; 11: 143–151.

17. 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; .

18. Virchow JC, Kuna P, Paggiaro P, Papi A, Singh D, Corre S, Zuccaro F, Vele A, Kots M, Georges G. Single inhaler extrafine triple therapy in uncontrolled asthma (TRIMARAN and TRIGGER): two double-blind, parallel-group, randomised, controlled phase 3 trials. *Lancet* Elsevier; 2019; 394: 1737–1749.

19. Kots M, Georges G, Guasconi A, Vogelmeier C. S26 Effect of single-inhaler extrafine beclometasone/formoterol/glycopyrronium pMDI (BDP/FF/GB) compared with two-inhaler fluticasone furoate/vilanterol DPI+ tiotropium DPI (FLF/VIL+ TIO) triple therapy on health-related quality of life (HRQoL) in patients with COPD: The TRISTAR study. BMJ Publishing Group Ltd; 2021.

20. Papi A, Vestbo J, Fabbri L, Corradi M, Prunier H, Cohuet G, Guasconi A, Montagna I, Vezzoli S, Petruzzelli S. Extrafine inhaled triple therapy versus dual bronchodilator therapy in chronic obstructive pulmonary disease (TRIBUTE): a double-blind, parallel group, randomised controlled trial. *Lancet* Elsevier; 2018; 391: 1076–1084.

21. Das N, Verstraete K, Stanojevic S, Topalovic M, Aerts J-M, Janssens W. Deep-learning algorithm helps to standardise ATS/ERS spirometric acceptability and usability criteria. *Eur. Respir. J.* Eur Respiratory Soc; 2020; 56.

22. Enright P, Vollmer WM, Lamprecht B, Jensen R, Jithoo A, Tan W, Studnicka M, Burney P, Gillespie

S, Buist AS. Quality of spirometry tests performed by 9893 adults in 14 countries: the BOLD Study. *Respir. Med.* Elsevier; 2011; 105: 1507–1515.

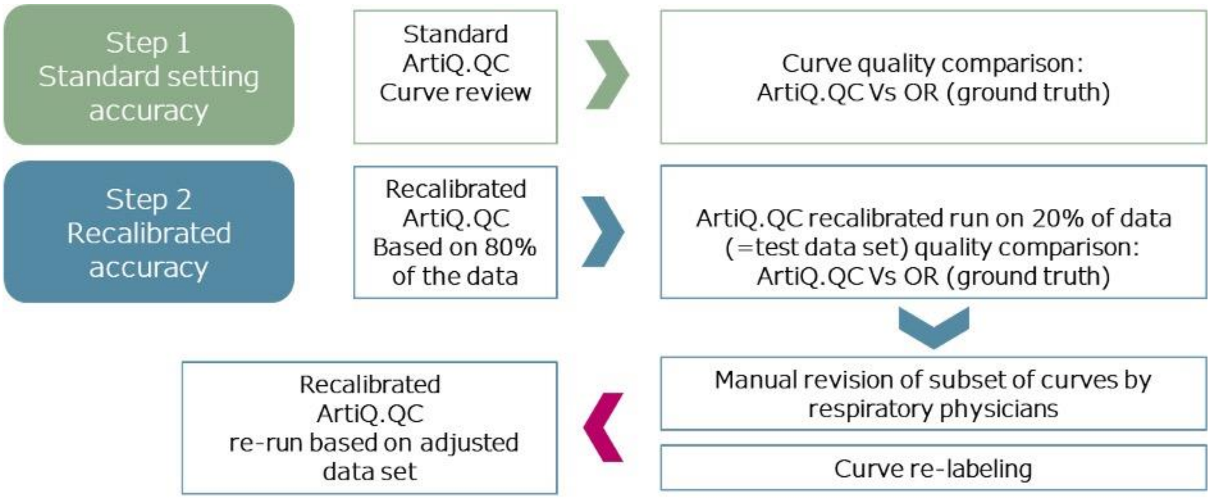
23. Enright PL, Beck KC, Sherrill DL. Repeatability of spirometry in 18,000 adult patients. *Am. J. Respir. Crit. Care Med.* American Thoracic Society; 2004; 169: 235–238.

24. Parsons R, Schembri D, Hancock K, Lonergan A, Barton C, Schermer T, Crockett A, Frith P, Effing T. Effects of the Spirometry Learning Module on the knowledge, confidence, and experience of spirometry operators. *NPJ Prim. care Respir. Med.* Nature Publishing Group; 2019; 29: 1–8.

25. Bonell C, Oakley A, Hargreaves J, Strange V, Rees R. Assessment of generalisability in trials of health interventions: suggested framework and systematic review. *Bmj* British Medical Journal Publishing Group; 2006; 333: 346–349.

26. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit. Heal.* Elsevier; 2020; 2: e489–e492.

Sandrine CORRE received support for the present manuscript from Chiesi, for which they are an employee. **Marko Topalovic** received support for the present manuscript from ArtiQ, for which they are a co-founder. **Eva Topole** received support for the present manuscript from Chiesi, for which they are an employee. **Isabella Montagna** received support for the present manuscript from Chiesi, for which they are an employee. **BIONDARO Sonia** received support for the present manuscript from Chiesi, for which they are an employee. **Sanja Stanojevic** has received consulting fees from Chiesi Farmaceuticals, outside the submitted work. **Brian L Graham** reports the following relationships outside the submitted work; received professional fees from Chiesi Farmaceutici S.p.A., MGC Diagnostics, Vyaire Medical, and the Lung Association of Saskatchewan. BG has a patent licensing agreement with Hans Rudolph Inc. **Kevin Ray** received support for the present manuscript from ArtiQ, for which they are an employee. **NILAKASH DAS** received support for the present manuscript from KU Leuven university for which they were a Post-doctoral researcher; patent pending for ArtiQ; stock options in lieu of patents held for ArtiQ. **Massimo Corradi** has a patent pending.



Step 1
Standard setting
accuracy

Standard
ArtiQ.QC
Curve review

Curve quality comparison:
ArtiQ.QC Vs OR (ground truth)

Step 2
Recalibrated
accuracy

Recalibrated
ArtiQ.QC
Based on 80%
of the data

ArtiQ.QC recalibrated run on 20% of data
(=test data set) quality comparison:
ArtiQ.QC Vs OR (ground truth)

Manual revision of subset of curves by
respiratory physicians

Curve re-labeling

Recalibrated
ArtiQ.QC
re-run based on adjusted
data set