ERJ Express. Published on December 23, 2014 as doi: 10.1183/09031936.00116814

Use of forced vital capacity and forced expiratory volume in 1 second quality criteria for determining a valid test

John L. Hankinson[1], Bill Eschenbacher[2], Mary Townsend[3], Janet Stocks[4] and Philip H. Quanjer[5,6]

Affiliations: [1]Hankinson Consulting, Inc., Athens, GA, USA. [2]Cincinnati VA Medical Center, Cincinnati, OH, USA. [3]MC Townsend Associates, Pittsburgh, PA, USA. [4]Respiratory, Critical Care and Anaesthesia Section (Portex Unit), UCL Institute of Child Health, London, UK. [5]Dept of Pulmonary Diseases, Erasmus University Medical Centre, Rotterdam, The Netherlands. [6]Dept of Paediatrics, Division of Respiratory Medicine, Erasmus University Medical Centre – Sophia Children's Hospital, Rotterdam, The Netherlands.

Correspondence: John L. Hankinson, Hankinson Consulting, 1860 Barnett Shoals Rd, Suite 103, PMB 505 Athens, GA 30605, USA. E-mail: john@hankconsulting.com

ABSTRACT The 2005 American Thoracic Society (ATS)/European Respiratory Society (ERS) spirometry guidelines define valid tests as having three acceptable blows and a repeatable forced vital capacity (FVC) and forced expiratory volume in 1 s (FEV_1). The aim of this study was to determine how reviewer and computer-determined ATS/ERS quality could affect population reference values for FVC and FEV_1.

Spirometry results from 7777 normal subjects aged 8–80 years (NHANES (National Health and Nutrition Examination Survey) III) were assigned quality grades A to F for FVC and FEV_1 by a computer and one reviewer (reviewer 1). Results from a subgroup of 1466 Caucasian adults (aged 19–80 years) were reviewed by two additional reviewers. Mean deviations from NHANES III predicted for FVC and FEV_1 were examined by quality grade (A to F).

Reviewer 1 rejected (D and F grade) 5.2% of the 7777 test sessions and the computer rejected ~16%, primarily due to end-of-test (EOT) failures. Within the subgroup, the computer rejected 11.5% of the results and the three reviewers rejected 3.7–5.9%. Average FEV_1 and FVC were minimally influenced by grades A to C allocated by reviewer 1.

Quality assessment of individual blows including EOT assessments should primarily be used as an aid to good quality during testing rather than for subsequently disregarding data. Reconsideration of EOT criteria and its application, and improved grading standards and training in over-reading are required. Present EOT criteria results in the exclusion of too many subjects while having minimal impact on predicted values.

@ERSpublications
ATS/ERS end of test criteria for spirometry exclude too many subjects while having minimal impact on predicted values http://ow.ly/DQRCD

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

Received: June 26 2014 | Accepted after revision: Oct 23 2014

Conflict of interest: Disclosures can be found alongside the online version of this article at erj.ersjournals.com

Introduction

The first step in deciding if a subject's spirometry results should be included in a study is by assessing the quality and validity of their test session. The 2005 American Thoracic Society (ATS)/European Respiratory Society (ERS) statement [1] defines a valid test during test performance as one that has three acceptable blows and a repeatable forced vital capacity (FVC) and forced expiratory volume in 1 s (FEV1). The ATS/ERS state that "It is desirable to use a computer-based system that provides feedback to the technician. […] The repeatability criteria are used to determine when more than three acceptable FVC manoeuvres are needed; these criteria are not to be used to exclude results from reports or to exclude subjects from a study" [1]. However, it is unclear when quality is insufficient for acceptance of results into research studies. There are no specific recommendations as to how strictly these criteria should be applied during test inclusion decisions, or the extent to which a reviewer should over-ride strict or computerised decisions regarding test validity. Which tests (subjects) are included in a study may impact study results, particularly in widely used reference values studies. Thus, amendments to quality criteria during the past two decades could potentially impact on spirometry reference standards. The NHANES (National Health and Nutrition Examination Survey) III spirometry data were collected with the aim of each subject trying to exhale for at least 6 s, with a minimum of five blows and a repeatable FVC and FEV1 within 5% or 100 mL. By contrast, the 1987 ATS criteria defined an end-of-test (EOT) plateau as a "<40 mL increment over 2 s" [2]. This was manually determined using an expanded EOT display using no change in volume for 1–2 s and/or a good expiratory effort. The ATS EOT criteria were subsequently amended, first to "no change in volume over 1 s", based on the spirometer's limit of detection volume of ⩽0.030 L [3], and then to "no change in volume (<0.025 L) for ⩾1 s, and the subject has tried to exhale for >6 s in subjects aged >10 years" [1]. The extent to which available spirometry software has implemented EOT and other quality assessment algorithms is varied and unclear.

Therefore, the aims of the present study were as follows. 1) To assess what proportion of spirometry test results from healthy nonsmokers in the NHANES III study [4] would be included in a study if quality control decisions had been based on computerised application of the 1994 ATS [3] and 2005 ATS/ERS [1] EOT definitions. 2) To assess the extent to which different grading systems might affect published reference data. 3) To compare computer-generated quality grades for FEV1 and FVC with those from a reviewer, as well as comparing quality grades between three reviewers.

Since test quality is an integral part of clinical test interpretation and therefore differs considerably from its application when used in the research context, we limited the scope of our investigations to healthy research subjects.

Study design

Spirometry curves and results from the NHANES III reference value study [2] were re-analysed to determine to what extent observed values for FVC and FEV1 are affected by quality control using published grading systems [5–7]; modified here to accommodate ATS and ATS/ERS recommendations (table 1) when applied by a computer and reviewers.

Since "normal" subjects were used to derive the reference equations for NHANES III and there was one test per subject, the impact of alterations in quality control criteria and relative grades can be measured by any deviation from the original predicted values. Consequently, predicted values [2] for the entire dataset

TABLE 1 Computer algorithms for assigning quality criteria grades for forced vital capacity (FVC)

Grade	Acceptable blows n	Repeatable	EOT criteria met[#,¶]?
A	⩾3	⩽150 mL	Yes
B	2	⩽150 mL	Yes
C	⩾2	⩽200 mL	Yes
D	⩾2	⩽250 mL	Yes
F	1	NA	Yes

EOT: end-of-test; NA: not applicable. [#]: FVC only; [¶]: during the original analysis of the National Health and Nutrition Examination Survey dataset, EOT was defined according to the 1987 American Thoracic Society (ATS) recommendations as no change in volume (*i.e.* <40 mL over 2 s) or exhalation for >15 s [2]. In the current review, analysis was repeated after amending this to the 1994 ATS recommendations (*i.e.* <25 mL for ⩾1 s [3] and, in subjects aged >10 years, exhalation for >6 s [1]).

DOI: 10.1183/09031936.00116814

were compared across computer-determined quality grades, using ⩽25 mL EOT criterion, with those from one reviewer (reviewer 1).

In a separate analysis, a subset of Caucasian adults was reviewed by two additional experienced reviewers to assess the agreement between reviewers. Apart from being told that "D" and "F" grades would result in a test (subject) being excluded from a study, no specific instructions or training were provided prior to these reviews.

Methods

Test results from 7777 healthy, nonsmoking subjects (58% male, 41% Caucasian, 31% aged <18 years) from the NHANES III study [4] were assigned quality grades (A through F) for FVC and FEV_1, separately as previously described [5–7], using both computer algorithms (table 1) and grades allocated by reviewer 1. Data from 348 normal subjects previously excluded from the NHANES reference dataset [4], due to unreliable results as judged by two original reviewers, were included in this re-analysis. Data from all 7777 subjects (table E1) were reviewed and graded both by computer and reviewer 1 according to the different quality criteria. Test results from the subset of all healthy Caucasian adults (n=1466) were reviewed by two additional reviewers. Predicted or lower limit of normal values were not provided to the reviewers.

When assigning test session quality grades, the computer visual display of all the subject's blows were available to the independent reviewers (*i.e.* all spirograms and flow–volume curves, including any that did not meet quality criteria), together with: 1) the corresponding values for FVC, FEV_1, FEV_1/FVC and peak expiratory flow; 2) their variability (*i.e.* how much individual values differed from the largest (best) values); and 3) a series of green or red bars for each of these outcomes to indicate whether or not acceptability criteria had been met according to the computer, and whether the largest FVC and/or FEV_1 were repeatable. Criteria for acceptability included: no artifacts, no abrupt termination, no glottis closure or cough within the first second of the test, no early termination, no leaks and no large back extrapolated volume, as well as a maximal continuous effort. The computer algorithm was set to determine if the EOT criterion was met, *i.e.* a volume change of <25 mL for ⩾1 s, and whether forced expiratory time (FET) was ⩾6 s. The computer grades and acceptability criteria, as shown in table 1, were provided as a guide but could be modified if judged appropriate by each reviewer following inspection of each blow, *e.g.* inspection of the tail of the curve appeared satisfactory even if it did not meet the computerised criteria of <25 mL in 1 s or FET ⩾6 s . In contrast to the computer, reviewers used visual inspection to judge a satisfactory EOT rather than basing this decision on a specific value for FET. Using all the information above, FVC and FEV_1 were scored separately; FVC and FEV_1 values could be used from blows where EOT criteria had not been met.

Analysis

FVC and FEV_1 were expressed as the difference (mL) between results observed in this study and the NHANES predicted values (ΔFVC and $ΔFEV_1$) [4]. The mean (95% confidence intervals) for ΔFVC and $ΔFEV_1$ were then calculated by quality grade. Mean differences should equal zero, with higher or lower values suggesting a bias according to quality grades. As we were assessing the influence of quality control on reference values obtained with the whole dataset [4], a subject's test session was included in the analysis even if not all EOT criteria were met. Therefore, any differences from predicted were not due to exclusion of blows where the FVC did not meet EOT criteria as such blows could still be used when deriving the subject's best FVC and FEV_1.

While the ATS/ERS standardisation document states that manoeuvres that do not meet EOT criteria should not be used to satisfy the requirements of three acceptable blows [1], this requirement was ignored for the purposes of this study so that its potential effect on results could be explored.

Statistical analyses were performed using SigmaPlot (Version 12.5; Systat Software, Inc., San Jose, CA, USA). ANOVA and linear regression were used to test for within-reviewer trends between ΔFVC and $ΔFEV_1$ and quality control grades. A p-value <0.05 was considered significant.

Results

Review of entire dataset

FVC

As expected, the rejection rate based on FVC quality alone increased with increasing quality and EOT requirements (table 2). When re-analysing the entire dataset, reviewer 1 rejected (D or F grade) 5.2% of the FVC results (4.5% for males, 6.5% for females), compared to 16% for the computer when using <25 mL EOT criterion (fig. 1). Requiring a quality grade of "A" would exclude >25% of tests, based on either reviewer or computer results (fig. 1). While the computer rejected a far higher percentage of tests (16% being graded F primarily due to failure to achieve an appropriate EOT, based on the 25 mL plateau

TABLE 2 Computer rejection rates based on application of original quality control criteria used in the NHANES III study and application of the 2005 ATS/ERS end-of-test (EOT) criteria[#]

Acceptable blows[¶] n	Other quality control criteria	Rejected n (%)	
		Age ⩾19 years[+]	Age <19 years[§]
NHANES III EOT criteria			
2[f]	Repeatable FVC	399 (8.7)	171 (5.3)
3	Repeatable FVC	427 (9.3)	183 (5.7)
ATS/ERS EOT criteria [1]			
⩾2	⩾6 s exhalation[##]	668 (14.6)	580 (18.1)
2	EOT plateau ⩽25 mL over 1 s	655 (13.6)	420 (13.1)
⩾2	Plateau ⩽25 mL and >6 s exhalation	758 (16.6)	656 (20.4)
⩾2	Expiratory time required to achieve a plateau (*i.e.* ⩽25 mL in 1 s) <6 s[¶¶]	2458 (53.8)	2383 (74.3)

Data are presented as n (%). NHANES: National Health and Nutrition Examination Survey; ATS: American Thoracic Society; ERS: European Respiratory Society; FVC: forced vital capacity. [#]: in addition to the 348 subjects excluded from the original NHANES reference dataset due to <2 acceptable blows, a further 222 had to be excluded on re-analysis, either because the original flow–volume curves could no longer be retrieved or because best FVC varied by >300 mL from the next best. [¶]: for the purpose of this analysis, the number of acceptable blows ignores the ATS/ERS statement regarding the need for EOT criteria to be met before blows can be considered "acceptable" unless specified by "other criteria". For example, of the 4568 subjects aged >19 years who were assessed, 4169 (91.3%) subjects achieved at least two repeatable FVC (irrespective of EOT criteria) whereas only 3411 (85.4%) of these subjects also achieved a plateau and an expiratory time of at least 6 s. [+]: n=4568. [§]: n=3209. [f]: test must have two acceptable blows before any criteria for a repeatable test can be calculated. [##]: 6 s exhalation was determined using total expiratory time or the time that the subject continued exhalation regardless of whether a plateau was achieved. [¶¶]: time required to reach a plateau with ⩽25 mL volume change was ⩾6 s, *i.e.* in 54% and 74% of subjects aged ⩾19 years and <19 years, respectively, the actual expiratory time required to reach a plateau (defined as a volume change <25 mL in 1 s) was <6 s.

criterion for any of the blows) than reviewer 1, the reviewer used additional information from inspecting the shape and consistency of curves, allocating fewer grade A but more grades B and C than the computer. For the test shown in figure 2, the computer assigned an FVC grade of "F" primarily due to less than two acceptable blows due to the technical lack of a plateau in the volume–time curve, while reviewer 1 assigned grade "B" since the FVC would only have increased by ~70 mL had exhalation continued for 15 s [4].

There were no differences in reviewer rejection rates between ethnic groups (5.5%, 5.6% and 5.9% rejected by reviewer 1 for Caucasian, Mexican–American and African–American subjects, respectively).

FVC differed significantly from predicted for grades A, C, D and F for reviewer 1 (fig. 3). The results were minimally, but statistically (ANOVA p=0.03), different for grades A–C with the mean (95% confidence interval) ΔFVC across these grades being 13.5 (3.1–23.9) mL. There was only a slight trend for FVC decreasing between grades A–C, with grade C providing results that were essentially equivalent to grades A and B. The ΔFVC from predicted for reviewer-allocated grades D and F (−309 and −518 mL, respectively) were much larger than those associated with computerised grading (−78 and −158 mL, respectively; ANOVA p<0.001) (fig. 3). Similarly, while FVC did decrease with poor computer-determined grades, ΔFVC differed significantly from zero only for grades A and F.

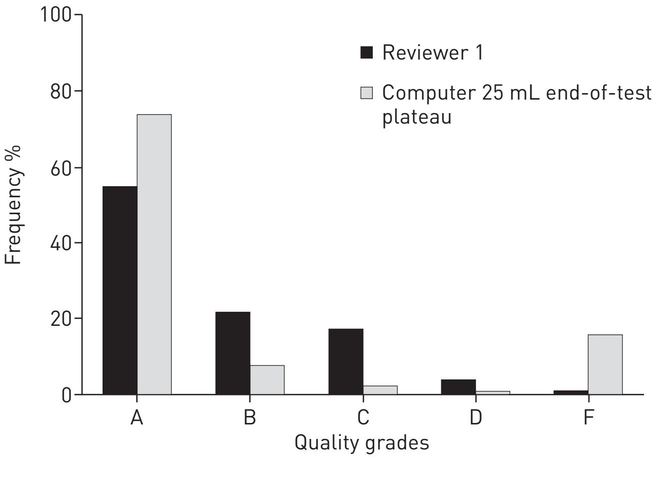
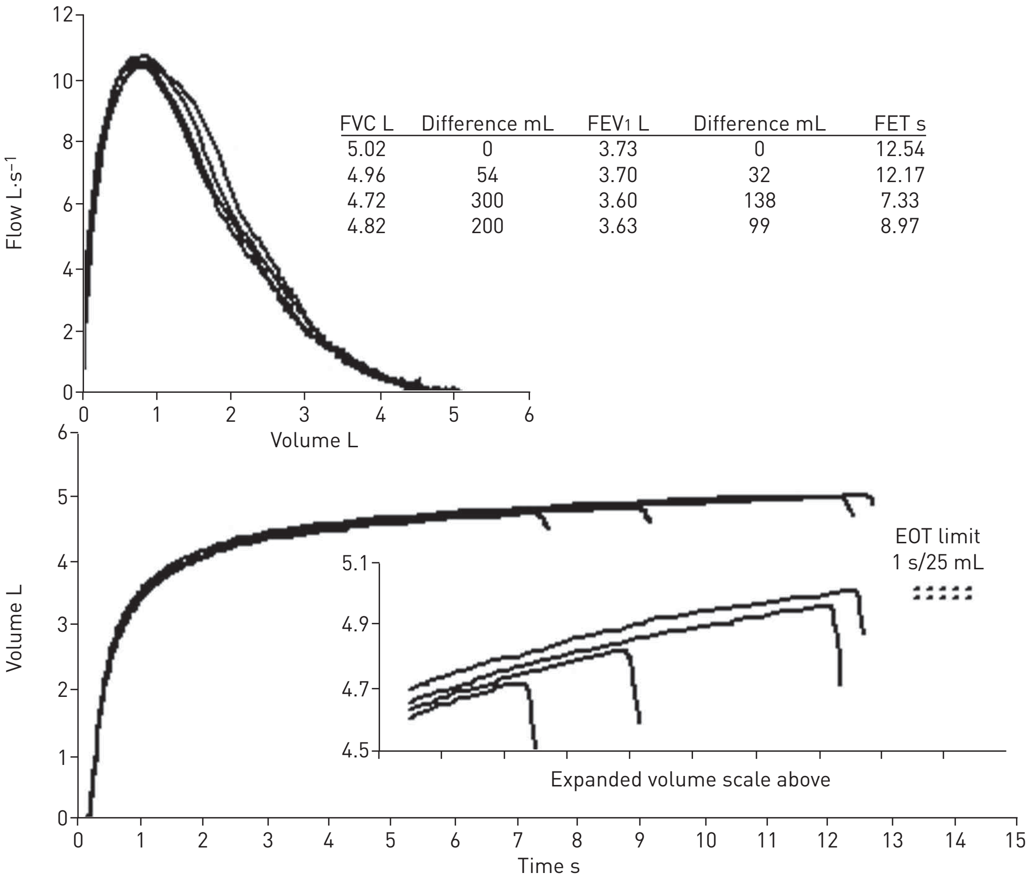


FIGURE 1 Distribution of forced vital capacity quality grades by reviewer 1 and the computer. The computer rejected a far higher percentage of tests (16% being graded F due to failure to achieve an appropriate end-of-test, based on the 25 mL plateau criteria for at least two blows) than reviewer 1. The reviewer used additional information from inspecting the shape and consistency of the curves, allocating fewer grade A but more grades B and C than the computer.

DOI: 10.1183/09031936.00116814



FVC L	Difference mL	FEV1 L	Difference mL	FET s
5.02	0	3.73	0	12.54
4.96	54	3.70	32	12.17
4.72	300	3.60	138	7.33
4.82	200	3.63	99	8.97

FIGURE 2 Sample curve showing end-of-test (EOT) or plateau failure as determined by the computer. FVC: forced vital capacity; FEV1: forced expiratory volume in 1 s; FET: forced expiratory time.

The total FET for all 7777 subjects (fig. 4a) was >6 s for virtually all subjects because this had been the EOT testing goal during test performance. However, in 54% of subjects aged ⩾19 years and 74% of subjects aged <19 years, the expiratory time actually required to reach a plateau defined as a volume change of <25 mL over 1 s was <6 s. (table 2). The <25 mL EOT plateau requirement resulted in FVC differences from predicted that increased with age to ~85 mL in older subjects (fig. 4b) because the criterion terminates the manoeuvre earlier.

FEV1

The FEV1 grades were higher than FVC since they were not affected by EOT issues and relatively few poor (D or F) FEV1 manoeuvres were identified by either reviewer 1 or the computer (3.1% and 3.9%, respectively) (fig. 5). There was a large decrease in FEV1 from grade A to F with ΔFEV1 being significantly below zero for both grade D and F results allocated by reviewer 1 and for computer-determined grade F. Reviewer 1 had more grades B and C than the computer (21.5–6.45%), primarily because, even if back extrapolated volume did not exceed 150 mL or 5% of FVC, lower grading occurred if the peak flow was suboptimal, with the exception of one study [8]. Suboptimal effort has been shown to affect flows and hence the FEV1 [9–12].

Subset analysis

Among the 1466 Caucasian adults that comprised the subset, the computer rejected two to three times as many subjects as all reviewers (table 3). Figure 6a shows the differences between the observed and predicted FVC results by quality grade. Only the ΔFVC results for D and F grades from reviewer 1 and the F grades from reviewer 2 were statistically different from zero. However, the trend of positive difference associated with grade A to more negative differences with grade F found in figure 3 is repeated for all reviewers and the computer.

Reviewers 1 and 2 found a significantly lower FEV1 for grade F (mean (95% CI) −243.6 (−428.0– −59.2) mL and −263.1 (−416.3– −109.9) mL, respectively). An association between ΔFVC and ΔFEV1

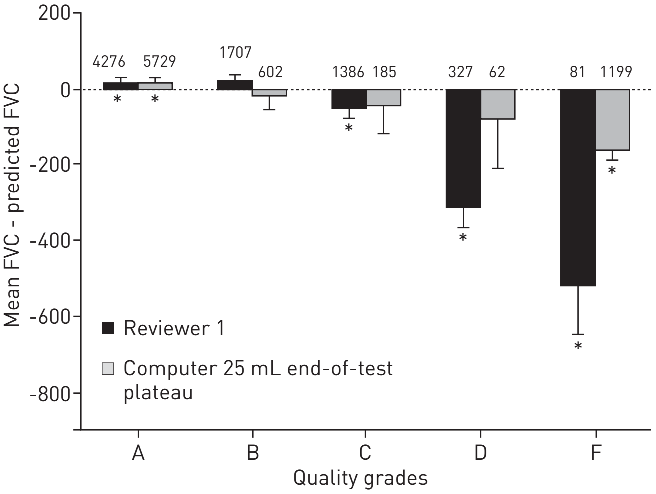


FIGURE 3 Differences between predicted and observed forced vital capacity (FVC) according to quality control grades allocated by reviewer 1 and the computer when applied to the entire dataset. The quality control grades are presented in table 1. n=7777. *: p<0.05.

with decreasing quality grade was found by reviewer 1 (all subjects: ANOVA $p<0.001$ and $p=0.04$; Caucasian adults: ANOVA $p=0.04$ and $p=0.03$, for ΔFVC and ΔFEV$_1$, respectively). This trend was also observed for ΔFVC when using the 25 mL EOT criterion ($p=0.03$), but not for reviewer 3.

Differences between reviewers

Differences in rejection rates between reviewers were greater for FVC than FEV$_1$ grades (3.7–5.9% and 1.9–2.1%, respectively) (table 3); thus, reflecting the more complex EOT judgments needed for FVC. Mean values for rejected FVC test results were significantly lower than predicted for all reviewers and the computer. FEV$_1$ rejected test rates were similar for all reviewers and the computer (fig. E1). Only reviewer 1 and the computer had values for rejected tests that were significantly lower than predicted.

Discussion

The main finding in this study was the proportion of tests regarded as being of unsatisfactory quality when assessed by strict application of the ATS/ERS guidelines using computer algorithms, corroborating

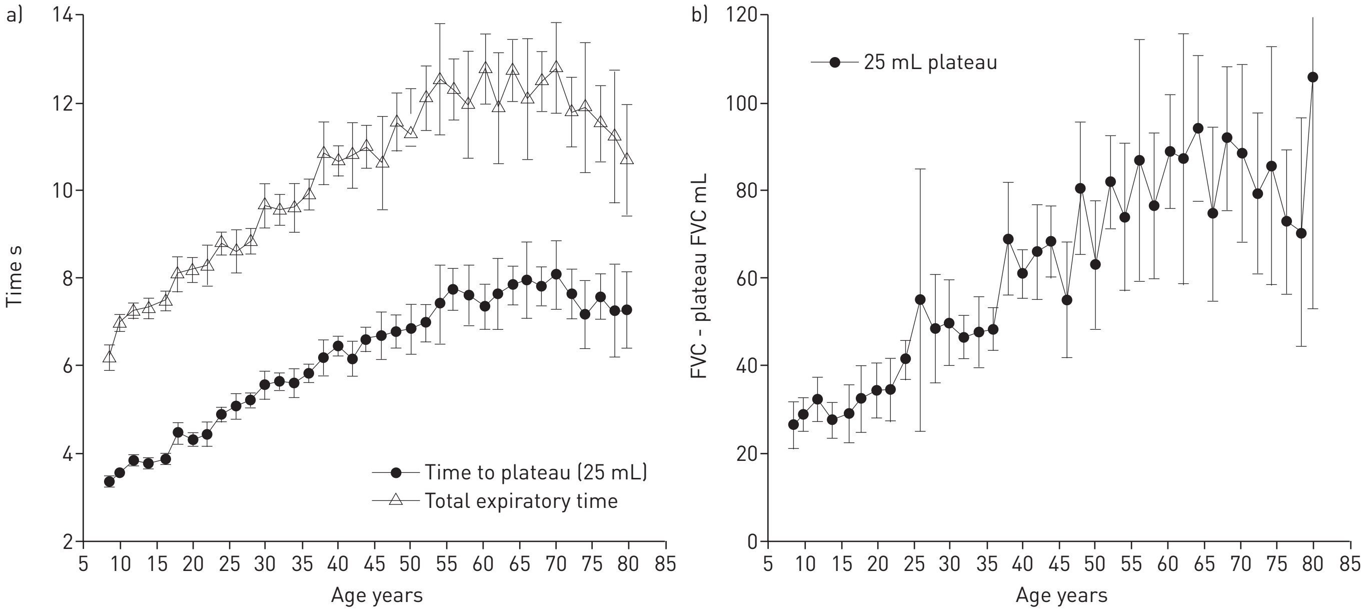


FIGURE 4 a) Total expiratory time was consistently longer at any age than the time required to reach an end-expiratory plateau of 25 mL. Note that in subjects aged <37 years the average time to reach a 25 mL plateau is <6 s. b) Applying a 25 mL over 1 s criteria for the plateau leads to an age-related underestimation of the maximum forced vital capacity (FVC).

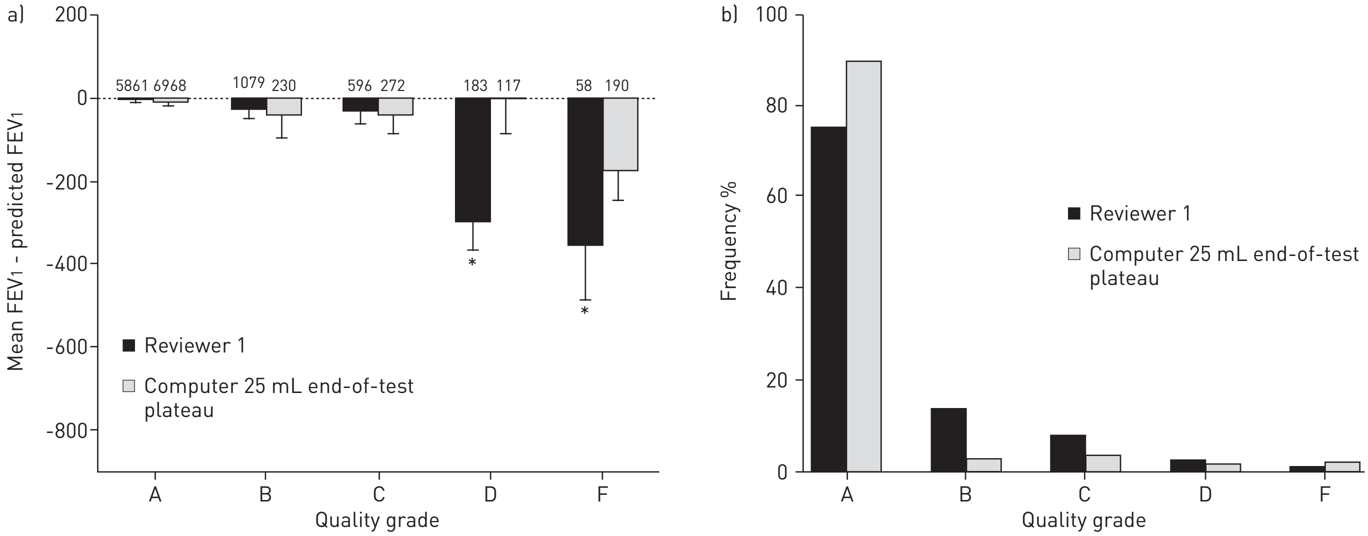


FIGURE 5 a) Differences between measured and predicted forced expiratory volume in 1 s (FEV1) for the entire dataset, according to quality control grades allocated by reviewer 1 and the computer. The numbers above the bars represent the number of subjects. Reviewer 1 had fewer grade A and more grades B and C than the computer. b) Distribution of FEV1 quality grades by reviewer 1 and the computer. *: $p<0.05$.

findings in children [13]. Including data rejected by the computer had no significant effect on overall results, whereas the reviewers were more discriminative. The large number of tests that would be rejected, particularly in younger subjects, due to failure to complete a 6-s exhalation but which could be included if a test was terminated as soon as an expiratory plateau had been obtained (table 2, fig. 4a), strongly suggests that the 2005 ATS/ERS EOT recommendations [1] need to be reconsidered. Our results also emphasise that quality assessment regarding the acceptability of individual blows including EOT assessments should be primarily used as an aid to good quality during testing rather than an excuse to subsequently disregard data. Of particular note is that use of the >25 mL EOT criterion resulted in an FVC error (fig. 4b), increasing with age up to a maximum error of 100 mL. Although these differences may not be clinically significant, such differences do indicate EOT quality may influence FVC results and are age dependent, even when using the current ATS/ERS recommended 25 mL EOT criterion.

There are several lessons to be learned from this study. The NHANES III study produced high-quality data because it adhered to a strict protocol [14], which included rigorous training of technicians and refresher sessions, well-defined test conditions, and care and maintenance of equipment, as well as online feedback

TABLE 3 Relative rates of rejection (D or F grade) and mean differences[#] in rejected values according to reviewers and computerised assessment for healthy subjects and Caucasian adults

Entire dataset	**FVC**		**FEV1**	
	Rejected	**ΔFVC[¶] mL**	**Rejected**	**ΔFEV1[¶] mL**
Healthy subjects[+]				
Reviewer 1	5.2	−351 (−402–−300)	3.1	−312 (−372–−251)
Computer 25 mL	16.0	−154 (−181–−128)	3.9	−110 (−165–−55)
Caucasian adults[§]				
Reviewer 1	3.7	−428 (−591–−265)	2.1	−212 (−383.4–−40.2)
Reviewer 2	3.8	−167 (−331–−4)	1.9	−92 (266.7–83.6)
Reviewer 3	5.9	−97 (−233–39)	2.0	−40 (−266.7–83.6)
Computer 25 mL	11.5	−104 (−191–−16)	3.1	−12 (−191.3–−16.1)

Data are presented as % or mean (95% CI). FVC: forced vital capacity; FEV1: forced expiratory volume in 1 s. [#]: calculated as the mean difference from predicted values observed for tests rejected according to the various types of analysis; [¶]: FVC was rejected by the computer (grades D and F) based on the 2005 American Thoracic Society/European Respiratory Society recommendations (*i.e.* end-of-test did not attain <25 mL for ⩾1 s in subjects aged >10 years, if exhalation was <6 s); [+]: n=7777; [§]: n=1466.

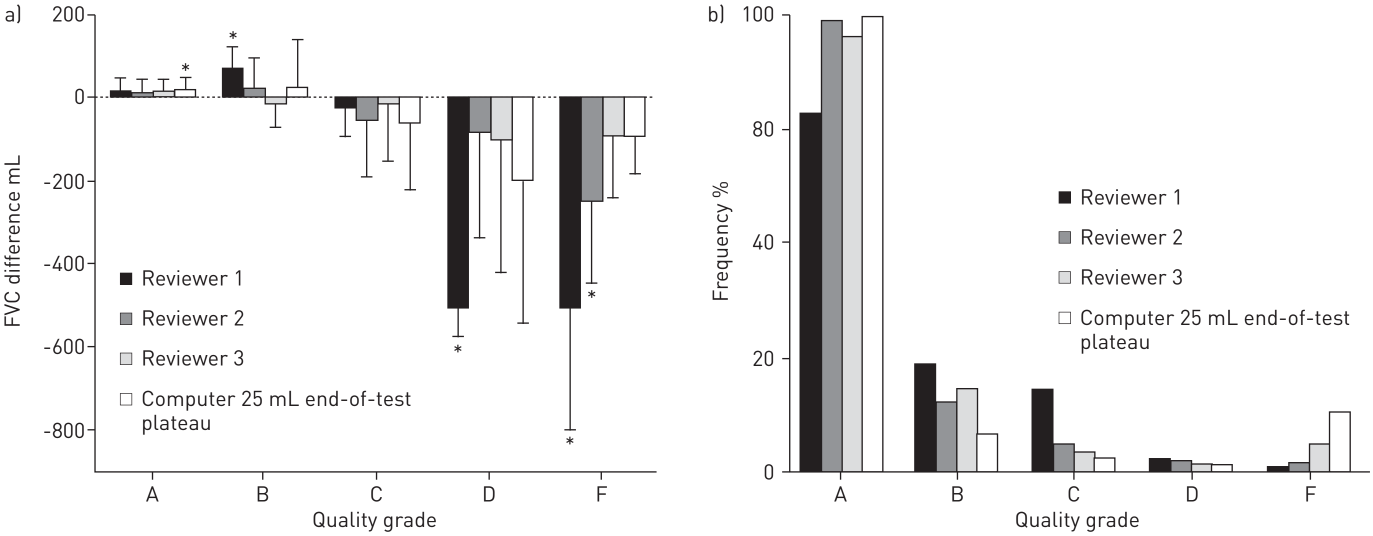


FIGURE 6 a) The difference between measured and predicted forced vital capacity (FVC) in 1466 Caucasian adults varied between reviewers and the computer, and increases the lower the quality grade. b) The computer rejected several measurements (grades D and F) which had little impact on predicted FVC, whereas reviewers were more discriminative, rejecting a low percentage of measurements with a considerable impact on FVC (a and b). See also figure E1for similar data relating to analysis of forced expiratory volume in 1 s in this subset. *: p<0.05.

of the quality of each manoeuvre and test. The maximum expiratory time limit was 20 s and a minimum of five blows were obtained. Yet a computer algorithm that applied the 2005 ATS/ERS recommendations (table 2) identified 18.2% of measurements as being of insufficient quality. A particularly important finding of this study was the large proportion of subjects, particularly younger subjects, who did not require 6 s to reach the 2005 ATS/ERS defined plateau volume of 25 mL (table 2). The fact that 53.8% of healthy subjects aged ⩾19 years and 74.3% of subjects aged <19 years would have their FVC results rejected for lack of exhaling for 6 s strongly suggests that the 2005 ATS/ERS EOT criteria should be reconsidered and that computerised quality control should not be relied upon in isolation. Clearly, the 2005 ATS/ERS quality goals, while providing an appropriate incentive during test performance, should not be used to determine whether a subject's results are included in a study. Several other studies underpin this suggestion [15–17], with Müller-Brandes *et al.* [13] concluding that a fixed cut-off for FET should be abandoned.

Enright *et al.* [15] found that ∼80% of their subjects met the 2005 ATS/ERS spirometry goals of three acceptable blows and a repeatable test to within 150 mL (grade A or B) [1]. The authors used an EOT requirement of 40 mL in the last second instead of the ATS/ERS recommended 25 mL, closer to the EOT used by reviewer 1 who rejected ∼5.2% of the NHANES III tests due to FVC failures. Since the 50th and 75th percentiles for the EOT volume in the study by Enright *et al.* [15] were 23 mL and 38 mL, respectively, somewhere between 25% and 50% of their subjects would not have met the ATS/ERS requirements for an EOT volume of 25 mL; similar to our results using a computer determined EOT. In another study of patients aged 20–89 years, 33% of test results needed to be excluded, mainly because the 6 s criterion could not be satisfied [16], whereas among children aged 8–11 years only 13.3% could exhale for 6 s [17].

While aiming for a 3 s exhalation in children aged <10 years, and 6 s beyond that age serves as a good incentive during test performance, many people simply cannot expel any gas from their lungs for that long and FET should not be used as a criterion for disqualifying a manoeuvre. Signal drift and errors in the spirometry software may also lead to inaccurate recording of expiratory duration and EOT [18]. If a volume plateau is not reached during an exhalation extended to up to 15 s, despite a good effort (fig. 2), it seems unjustified to qualify the results as unsatisfactory. This situation is most likely to arise in subjects with airways obstruction, where the largest FVC from a 15-s manoeuvre with good effort will still provide clinically useful information. Measurements of lung function may be used for various purposes including clinical diagnosis or management, as outcome measures in research studies or for deriving predicted values, these do not necessarily require the same accuracy and precision. FVC and FEV1 quality grades should allow the use of different grading limits according to underlying purpose, or allow investigation of the effects of eliminating subjects with low-quality control grades from a study. We found that grades D and F were associated with lower test results whether allocated by reviewers or the computer, with these differences being highly significant for some reviewers.

DOI: 10.1183/09031936.00116814

Inter-reviewer differences

Differences in grading between reviewers were generally small (fig. 6b), with all reviewers showing decreasing test results with progressively poorer quality control grades. Reviewer 1 routinely reviews or grades tests for research studies where the concern is to prevent poor-quality tests from influencing the results and to provide quality feedback to technicians. Thus, reviewer 1 may be biased towards obtaining the most accurate expected value rather than determining whether the test results are "clinically" useful. Reviewer 2 runs spirometry training courses, conducts reviews for occupational clients and provides feedback to technicians. Therefore, reviewer 2 is also concerned with high levels of accuracy as job placement and costly unnecessary referrals may result from falsely low values. By contrast, reviewer 3 conducts reviews for clinical interpretation where the objective is detection or classification of disease and where false-negative results are more of a concern. Thus, for reviewer 3, the tendency is to be more concerned with test repeatability and particularly EOT issues, which can significantly impact the FEV_1/FVC due to early termination resulting in a false negative classification of patients with the obstructive lung disease pattern.

One potential limitation of this study is that results are based on NHANES III data which were collected using a volume-type spirometer, whereas flow-type spirometers are more frequently used nowadays. However, since the errors associated with incorrect registration of zero flow that may occur in flow-type spirometers do not occur when using volume-type spirometers, the use of a volume-type spirometer may in fact be an advantage when investigating EOT quality issues.

Due to time constraints, the subset used for comparisons between reviewers was limited to Caucasian adults. However, as reviewer 1 did not find differences in quality between ethnic groups, this is unlikely to have affected the results.

Differences between reviewers might have been decreased slightly had an initial formal training period been undertaken to ensure internal consistency. We did not study the extent to which within-reviewer variability contributed to overall variability. Regardless, the differences between the reviewers were small and probably reflect the experience of our reviewers as discussed above. The largest differences were between reviewers and the computer, with the computer rejecting at least twice the number of tests.

Conclusion

The results from this study emphasise that quality assessment regarding the acceptability of individual blows including EOT assessments should be primarily used as an aid to good quality during testing rather than a reason to subsequently disregard data. Although important when used as a testing goal, application of the 2005 ATS/ERS EOT criteria will reject many tests judged acceptable by experienced reviewers [1]; specifically, the need for a fixed minimum FET needs to be reconsidered. In adults, provided that at least two acceptable blows are obtained wherein values of FVC and FEV_1 are within 200 mL, slight deviations in quality grading are unlikely to significantly impact study results or clinical interpretation. Although differences in quality grading between experienced reviewers were minimal, the background of the reviewer or the context of the review may influence the grading system in detecting poor quality tests. Computer quality assessments appear to reject far more tests than human reviewers, primarily due to the application of EOT requirements. Based on our results, strict computer-generated quality grading will probably have a minimal effect on predicted values, but will result in far fewer subjects being included in a study. Similarly, as application of quality criteria may not always improve clinical interpretations, visual inspection [13] will always be required to grade slightly lower quality tests and needs to be standardised as far as possible.

Acknowledgements

We thank M.R. Miller and the European Respiratory Society anonymous reviewers for their critical and constructive comments, which were instrumental in improving the presentation of our findings and clarifying the scope of our research.

References

1 Miller MR, Hankinson J, Brusasco V, *et al.* Standardisation of spirometry. *Eur Respir J* 2005; 26: 319–338.
2 Standardization of spirometry – 1987 update. Statement of the American Thoracic Society. *Am Rev Respir Dis* 1987; 136: 1285–1298.
3 Standardization of spirometry, 1994 update. American Thoracic Society. *Am J Respir Crit Care Med* 1995; 152: 1107–1136.
4 Hankinson JL, Odencrantz JR, Fedan KB. Spirometric reference values from a sample of the general U.S. population. *Am J Respir Crit Care Med* 1999; 159: 179–187.
5 Enright PL, Johnson LR, Connett JE, *et al.* Spirometry in the Lung Health Study. 1. Methods and quality control. *Am Rev Respir Dis* 1991; 143: 1215–1223.
6 Enright PL. How to make sure your spirometry tests are of good quality. *Respir Care* 2003; 48: 773–776.
7 Ferguson GT, Enright PL, Buist AS, *et al.* Office spirometry for lung health assessment in adults: a consensus statement from the National Lung Health Education Program. *Chest* 2000; 117: 1146–1161.

8 Hegewald MJ, Lefor MJ, Jensen RL, *et al.* Peak expiratory flow is not a quality indicator for spirometry: peak expiratory flow variability and FEV$_1$ are poorly correlated in an elderly population. *Chest* 2007; 131: 1494–1499.

9 Ingram RH, Schilder DP. Effect of thoracic gas compression on the flow–volume curve of the forced vital capacity. *Am Rev Respir Dis* 1966; 94: 56–63.

10 Sadoul P. Mesure de la capacité vitale et des débits maximaux. *In*: Denolin H, Sadoul P, Orie NGM, eds. L'exploration fonctionnelle pulmonaire, 2me partie. Paris, Flammarion, 1971.

11 Van de Woestijne KP, Afschrift M. Airway dynamics during forced expiration in patients with chronic obstructive lung disease. *In*: Orie NGM, Van der Lende R, eds. Bronchitis III. Assen, Royal Van Gorcum, 1970; pp. 195–206.

12 Krowka MJ, Enright PL, Rodarte JR, *et al.* Effect of effort on measurement of forced expiratory volume in one second. *Am Rev Respir Dis* 1987; 136: 829–833.

13 Müller-Brandes C, Krämer U, Gappa M, *et al.* LUNOKID: can numerical American Thoracic Society/European Respiratory Society quality criteria replace visual inspection of spirometry? *Eur Respir J* 2014; 43: 1347–1356.

14 Centers for Disease Control and Prevention. Third National Health and Nutrition Examination Survey III. Spirometry Procedure Manual. Rockville, Westat Inc., 1988.

15 Enright PL, Skloot GS, Cox-Ganser JM, *et al.* Quality of spirometry performed by 13,599 participants in the World Trade Center Worker and Volunteer Medical Screening Program. *Respir Care* 2010; 55: 303–309.

16 Swanney MP, Jensen RL, Crichton DA, *et al.* FEV$_6$ is an acceptable surrogate for FVC in the spirometric diagnosis of airway obstruction and restriction. *Am J Respir Crit Care Med* 2000; 162: 917–919.

17 Arets HG, Brackel HJ, van der Ent CK. Forced expiratory manoeuvres in children: do they meet ATS and ERS criteria for spirometry? *Eur Respir J* 2001; 18: 655–660.

18 Aurora P, Stocks J, Oliver C, *et al.* Quality control for spirometry in preschool children with and without lung disease. *Am J Respir Crit Care Med* 2004; 169: 1152–1159.