LUNOKID: can numerical American Thoracic Society/European Respiratory Society quality criteria replace visual inspection of spirometry?

Christine Müller-Brandes[1,2,9], Ursula Krämer[3,9], Monika Gappa[2,4], Gabriele Seitner-Sorge[3], Anke Hüls[3], Andrea von Berg[5], Barbara Hoffmann[3], Antje Schuster[6], Sabina Illi[7], Matthias Wisbauer[6,8] and Dietrich Berdel[5]

Affiliations: [1]Dept of Anaesthesiology and Intensive Care Medicine, Medical School, Hannover, [2]Dept of Paediatric Pulmonology, Allergology and Neonatology, Medical School, Hannover, [3]IUF-Leibniz-Research Institute for Environmental Medicine, Düsseldorf, [4]Children's Hospital and Research Institute, Marien Hospital, Wesel, [5]Research Institute, Marien Hospital, Wesel, [6]University Children's Hospital, Heinrich-Heine University, Düsseldorf, [7]University Children's Hospital, Ludwig-Maximilians-Universität, Munich, and [8]Dept of Pedatric Cardiology and Pneumology, HELIOS Clinic Wuppertal, Wuppertal, Germany. [9]These authors contributed equally.

Correspondence: C. Müller-Brandes, Dept of Anaesthesiology and Intensive Care Medicine, Medical School, Carl-Neuberg-Str. 1, 30625 Hannover, Germany. E-mail: mueller-brandes.christine@mh-hannover.de

ABSTRACT   The gold standard for assessing quality of forced expiratory manoeuvres is visual inspection by an expert. American Thoracic Society/European Respiratory Society numerical quality criteria (NQC) include back-extrapolated volume (BEV), repeatability and forced expiratory time (FET). Equipment currently available provides feedback tempting the investigator to use NQC as pass–fail criterion.

To investigate whether using NQC instead of visual acceptability is a valid option, we analysed data from a multicentre national reference study in Germany of children aged 4–18 years. Spirometry was performed under field conditions. Receiver operating characteristic analysis was used to assess performance of BEV, repeatability, FET and a combination thereof in relation to visual acceptability.

We included data from 3133 healthy Caucasians in the analyses; 72% delivered at least two visually acceptable manoeuvres. Of these, 59% would have been rejected based on combined NQC, mainly because the FET criterion was not feasible. Specificity of the NQC was generally low (BEV 10%, repeatability 30% and FET 50%). Receiver operating characteristic analysis showed that a combination of the three measures could reach at best a sensitivity of 90% and specificity of 56%.

We conclude that visual control is mandatory and NQC may help obtain the best possible results, but a fixed cut-off for FET should be abandoned.

@ERSpublications
Visual control of spirometry is mandatory: quality criteria may help to get best results, but we should abandon FET http://ow.ly/uje5B

Introduction

Spirometry is the most frequently used method for assessing lung function in clinical routine and research. As the results are effort dependent, international standards based on numerical criteria have been defined to guide technicians in their effort to obtain the best possible curve and to ensure the quality of the data obtained [1, 2]. However, these criteria for quality control are not evidence based. Although the usefulness and validity of these recommendations have been questioned repeatedly, particularly in children and adolescents [3–12], they are widely applied and implemented in commercially available equipment. For quality control to be clinically relevant, however, criteria have to be feasible, sensitive and specific. They should aid the technician to obtain the best possible results and to support an expert's judgement.

The American Thoracic Society (ATS)/European Respiratory Society (ERS) [1, 2, 13] recommendations for quality control in spirometry include visual inspection to ascertain a satisfactory start with 1) rapid onset of expiration; 2) a clearly determined peak flow with a single distinct peak; 3) no manoeuvre artefacts, such as cough or glottis closure during expiration; and 4) no evidence of leak or early termination indicated by abrupt ending or truncation of the expiration, stipulating a plateau in the volume–time curve. Three quantitative numerical quality criteria (NQC) are listed to enable objective quality control, as follows. 1) Measurements of back-extrapolated volume (BEV); 2) forced expiratory time (FET) and repeatability, derived from inter-manoeuvre differences in forced expiratory volume in 1 s (FEV_1); and 3) forced vital capacity (FVC). Different cut-offs have been defined for different age groups [1, 2]. The ATS/ERS guidelines state that some curves may not be acceptable according to these NQC, but may still be partially usable. These criteria are computerised and commonly available in spirometers in contrast to several other suggested numerical criteria, including cessation of flow, plateau in the volume–time curve or curvilinearity [4, 10, 14, 15].

BEV is a measure of the quality of the start of the test and it may be expressed either as percentage of the FVC or as an absolute value. It is determined by the back-extrapolation method, where a new time zero is defined from the steepest part of the volume–time curve. The recommended cut-off for preschool children (aged <6 years) is 12.5% or 80 mL, and for older children and adults it is 5% or 150 mL, whichever is greater.

For young children (aged <6 years), no minimum FET has been recommended, but for children aged between 6 and 10 years, 3 s is defined as minimum duration of the forced expiration, and 6 s for children aged ≥10 years, unless the subject cannot or should not continue exhalation.

Repeatability is defined as the difference between the largest and next largest FEV_1 and FVC [1]. The cut-off points for repeatability depend on either FVC or age, and are 150 mL if FVC >1.0 L and 100 mL for an FVC <1.0 L or for preschool children.

Most software for currently available equipment provides feedback to the user whether these NQC criteria are met, and may dominate the individual decision about acceptability, regardless of the result of a potential visual inspection. We hypothesised that NQC parameters cannot replace visual inspection.

The aim of the present study was to investigate 1) whether the recommended NQC for spirometry (BEV, repeatability and FET) are feasible from preschool age through to adolescence; and 2) whether any of these criteria or a combination thereof can, when implemented in a computer system, replace visual quality control.

Material and methods

A detailed description of the methods is presented in the online supplementary material and has been published previously [16, 17].

Caucasian subjects aged 4–18 years were recruited for a German multicentre project to generate new reference data for spirometry from a randomised selection of kindergartens and schools (the LUNOKID study: Lungenfunktions-Normalwerte bei Kindern in Deutschland [Lung function Normal Values of Children in Germany]). The study had been approved by the local ethics committees.

After an initial training period, local technicians were trained and supervised monthly by the same quality control team to ensure standardised testing.

Prior to the actual measurement day, families were asked to complete a questionnaire. Only healthy subjects without acute infection on the day of testing, lower airway infection in the last 6 weeks, asthma or bronchitis, and from whom written consent had been obtained were included in the analysis.

Spirometry was carried out under field conditions in a sitting position using an EasyOne handheld device (ndd-Medizintechnik AG, Zurich, Switzerland) connected to a personal computer with online display of the manoeuvre. Nose clips were used if tolerated. After explaining and demonstrating the manoeuvre, maximal expiration was intensively encouraged to achieve the best possible results, aiming for three technically

acceptable curves in a maximum of nine trials. In very young children, recording of trials started once they were familiar with the situation and the procedure.

The first 100 data sets were reviewed by two observers (M. Gappa and C. Müller-Brandes) blinded to the identity of the child. After that, one experienced observer (C. Müller-Brandes) reviewed the data according to a standard protocol, as follows. First, all spirograms and flow–volume curves, as well as all the numerical results, were inspected for each single subject and then these manoeuvres were graded as ''visually acceptable'' (good start of test with a steep slope of the flow–volume curve at start of test, clearly distinguishable peak flow, no artefacts, and complete expiration judged by a plateau or no abrupt cessation during expiration), ''partially acceptable'' (good start of test and incomplete expiration) or ''not acceptable'' (start of test not acceptable and/or lack of a peak and/or artefact and/or expiration incomplete). This quality control process was monitored at least monthly (M. Gappa) throughout the study. Borderline acceptable manoeuvres with questionable visual acceptability (*e.g.* relatively slow start of test or no clear plateau, especially in younger children) were reviewed by both C. Müller-Brandes and M. Gappa. Examples of ''visually acceptable'', ''partially acceptable'', ''not acceptable'' and ''borderline'' curves can be found in online supplementary figures E1–E4.

The selection of the best curve followed international recommendations [2]. In order to directly investigate the association between visual acceptability and the ATS/ERS NCQ criteria for a single manoeuvre, FEV_1 and FVC were chosen from the same curve with the highest sum of FEV_1 and FVC and not from different curves, as recommended by the ATS. To assess repeatability, the curve with the second highest sum of FEV_1 and FVC was identified. Our selection method revealed no differences to the recommended selection procedures. It leads to almost identical mean curves of FVC or FEV_1 over age when applied to visual acceptable manoeuvres, as shown for males and females (online supplementary figs E5–E8).

Statistical analysis

Descriptive statistics were performed for all parameters. Medians and interquartile ranges are given to characterise the distribution.

Taking visual acceptability as gold standard, sensitivity and specificity were calculated for each numerical quality measurement and their combinations. The association between visual acceptability and the objective numerical quality measurements was described using logistic regression analysis. FET was standardised for age before inclusion into the models. Performance was quantified by the area under the receiver operating characteristic (ROC) curve (AUC). Details of the analysis can be found in the online supplementary material.

Results

From October 2007 until July 2009, 9410 families were contacted and 5427 returned a complete questionnaire. Figure 1 summarises the recruitment.

5415 of the children could be recruited for lung function testing and all of these manoeuvres were visually evaluated. Of the 5415 children, 77% were able to produce at least one acceptable curve. This percentage was higher in older children and adolescents (table 1).

5104 children and adolescents were Caucasian and had at least two curves providing all necessary lung function indices. Of these, 1971 children had to be excluded according to the protocol: 1224 (24%) were excluded because they had had doctor-diagnosed asthma or doctor-diagnosed bronchitis or had taken asthma medication on the day of the examination. Additionally, 747 (15%) were excluded because they had had an infection of the lower airways in the last 6 weeks or an infection of the upper airways on the day of investigation. The 3133 healthy children formed the basis for the further analysis. Of these, 72% (2262 out of 3133) performed at least two visually acceptable forced expiratory manoeuvres.

Table 2 shows the distribution of population characteristics for the entire group and in groups defined by visual acceptability. Sex did not influence visual acceptability, but younger children were less likely to produce a visually acceptable curve. The observed differences in height and weight between the groups are explained by differences in age.

Current objective numerical criteria

The feasibility of achieving current quality control criteria in a contemporary population of healthy children and adolescents is given in table 3. While BEV and repeatability criteria were met by the majority of children and adolescents, the required FET was achieved by only 18% of the subjects aged 10–18 years with a visually acceptable curve.

In the entire group of the children, 83% met the combination of both BEV and repeatability criteria. However, only 42% of the subjects who produced at least one visually acceptable curve simultaneously met

```
┌─────────────────────────────────────────────┐
│          Families addressed                  │
│              n=9410                           │
└─────────────────────────────────────────────┘
                      │
┌─────────────────────────────────────────────┐
│        Questionnaires returned               │
│              n=5427                           │
└─────────────────────────────────────────────┘
                      │
┌─────────────────────────────────────────────┐
│ Children recruited and all manoeuvres        │
│      visually evaluated                       │
│              n=5415                           │
└─────────────────────────────────────────────┘
                      │
┌─────────────────────────────────────────────┐
│ Lung function available (all necessary       │
│      indices)                                 │
│              n=5376                           │
└─────────────────────────────────────────────┘
                      │
┌─────────────────────────────────────────────┐
│      4–18-year-old Caucasian                 │
│              n=5104                           │
└─────────────────────────────────────────────┘
                      │
┌─────────────────────────────────────────────┐
│              Healthy                          │
│ (no acute infection, no asthma or            │
│ bronchitis, no lower airway                   │
│ infection during the last 6 weeks)           │
│              n=3133                           │
└─────────────────────────────────────────────┘
```
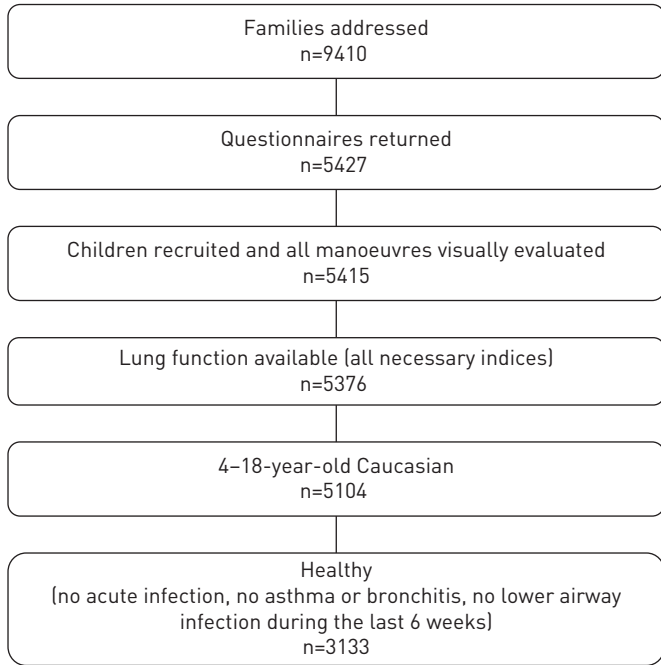
FIGURE 1 Schematic representation of recruitment in the LUNOKID study.

all three of the numerical ATS/ERS criteria. This means that 58% of visually accepted curves would have been rejected if only computerised criteria were considered.

Sensitivity of BEV and repeatability were good compared with visual acceptability as the gold standard. However, specificity was low. Of the visually unacceptable manoeuvres, only 10% were also unacceptable based on BEV, 30% were based on repeatability and 36% were based on both criteria. Taking all three quality control parameters together, specificity increased to 67% (only 33% falsely accepted), but sensitivity was reduced to 42%. That is because the FET criterion was difficult to meet and was generally not feasible in the >10 years age group.

The abrupt changes in NQC at a specified age was also shown to have an impact on feasibility, sensitivity and specificity, as can be seen in table 3.

Age dependency of quantitative ATS/ERS quality measurements

The age dependency in males of BEV, FET and the difference between the two highest FVCs in groups defined by NQC and visual acceptability is illustrated in figure 2. The graphs for females can be found in online supplementary fig. E9. The graphs for age dependency of the two highest FEV_1 measurements are similar to the FVC graphs (not illustrated).

FET strongly depends on age. The median$\pm$SD FET was 2.4$\pm$1.5 s in children aged 4–<6 years, 4.3$\pm$1.5 s in children aged 6–<10 years and 4.2$\pm$1.6 s in subjects aged 10–18 years. Up to the age of ~8 years FET increased linearly with age and, afterwards, a plateau is reached. BEV as a percentage of FVC (% FVC)

TABLE 1 Percentage of visually acceptable manoeuvres per child in age groups[#]

Age years	Subjects	Acceptable manoeuvres per child			
		0	1	2	⩾3
4–<6	1052	457 (43)	117 (11)	120 (11)	358 (34)
6–<10	1510	322 (21)	170 (11)	201 (13)	817 (54)
10–<15	1599	254 (16)	124 (8)	143 (9)	1078 (67)
15–<19	1228	191 (16)	103 (8)	105 (9)	829 (68)
Total[¶]	5415	1228 (23)	519 (10)	572 (11)	3096 (57)

Data are presented as n or n (%). [#]: n=5415 children with lung function testing from the LUNOKID study; [¶]: includes children aged slightly <4 or >18 years.

	Visually acceptable	Visually not acceptable	p-value[¶]	Total
Subjects n (%)	2262 (72.2)	871 (27.8)		3133 (100)
Male %	48	45	0.1407	48
Age years	11.1 (7.6)	7.7 (7.9)	<0.0001	10.3 (8.0)
Height cm	148.0 (39.0)	128.0 (46)	<0.0001	143.0 (43.0)
Weight kg	41.0 (33.0)	27 (30)	<0.0001	37.0 (33.0)

TABLE 2 Population characteristics in groups defined by visual acceptability[#]

Data are presented as median (interquartile range), unless otherwise stated. [#]: n=3133 healthy children from the LUNOKID study; [¶]: differences between groups using Wilcoxon rank sum test.

decreased slightly with age, whereas the measure for repeatability (absolute difference between highest and second highest FVC) increased with age. Thus, repeatability is lower for adolescents than for children.

Comparison of quantitative ATS/ERS quality measurements with visual acceptability: ROC analysis
Visual acceptability is significantly related to the objective quality measurements (table 4). Children with a visually acceptable forced expiratory manoeuvre had significantly lower BEV, smaller differences between the two highest FVCs and the highest FEV_1, and higher FET than children without a visually acceptable manoeuvre.

The ROC curve for the full model, including BEV, FET and repeatability to predict visual acceptability, is shown in figure 3 along with the ROC curves for the single quality measurements. The coefficients of the prediction equation can be found in the online supplementary material. The AUC is 0.81 for the full model indicating good overall concordance.

Using FET alone, at best, 75% correct classification of visual acceptability can be reached, with a sensitivity of 95% and specificity of 24%. The best cut-off value for FET is 1.68 s. Using BEV alone, at best, correct classification can be reached in 79% with a sensitivity of 95% and specificity of 36%. This is reached with a BEV value of 4.6%. Using repeatability (differences between FEV_1 or FVC) alone only marginally increases the percentage correctly classified (73.8% compared with 72.2% when simply judging every curve as visually acceptable). At best, by combining BEV, FET and repeatability, correct classification is achieved in 80.6% with a sensitivity of 90% and specificity of 56%.

Discussion
Our data from a large group of healthy subjects aged 4–18 years demonstrate that quality criteria based on BEV% of FVC, FET and repeatability cannot replace visual acceptability when assessing quality of forced expiratory manoeuvres. This is true with regard to the current international quality criteria, where the FET criterion is not met by the majority of subjects in our study, as well as for a new combination of these parameters where, despite good sensitivity, specificity is poor.

The feasibility of meeting the international numeric ATS/ERS recommendations for quality criteria based on BEV% of FVC, FET and repeatability has been the topic of a number of studies [2–10]. However, despite these discussions, the guidelines are still in place.

In our study, the main reason to reject current international standards was that FET was shorter than recommended. The current recommendations include different cut-offs for FET according to age. While for children below school age, no minimal FET is defined, a minimal FET of 3 s in children aged <10 years and 6 s in older subjects has been suggested. We observed quite a different pattern for different ages: While FET increased linearly up to the age of ∼8 years, there was a plateau with median FET of 4.2 s in older children. If current recommendations were followed, the majority of manoeuvres (60%) would be rejected, even if rated visually acceptable. These results are in line with previous reports, even though subjects included in these studies often had more experience in performing lung function tests [3, 4, 18, 19]. ARETS *et al.* [3] found that 90% of their population (experienced in lung function testing and most with current pulmonary problems) were able to produce a FET of at ⩾1.8 s in children aged 12–15 years and 2.1 s in children aged 8–11 years. GOCHICOA-RANGEL *et al.* [20] found 87.2% of their young population (median age 6.7 years) could reach a FET of ⩾3 s. Longer FET correlated with age, experience and history of asthma and wheeze [3, 21]. The age dependency was explained by increasing lung volumes during growth, as well as the changing relationship between airway calibre and lung size and, hence, rate of lung emptying [22, 23]. In our population, male subjects achieved higher values for FET than females. This sex difference has been explained by males having a bigger lung volume compared with females and, thus, smaller airways in

TABLE 3 Feasibility, sensitivity and specificity of numerical American Thoracic Society/European Respiratory Society criteria met in groups defined by age and visual acceptability

	Age years	Criterion	Fulfilled criteria (feasibility)#	Visually acceptable¶		Visually unacceptable+	
				Fulfilled criteria (sensitivity)	Did not fulfil criteria	Fulfilled criteria	Did not fulfil criteria (specificity)
BEV	<6	12.5% FVC or 80 mL	628/637 (98.6)	342/342 (100.0)	0/342 (0.0)	286/295 (96.9)	9/295 (3.1)
	≥6	5% FVC or 150 mL	2417/2496 (96.8)	1916/1920 (99.8)	4/1920 (0.2)	501/576 (87.0)	75/576 (13.0)
	All		3045/3133 (97.2)	2258/2262 (99.8)	4/2262 (0.2)	787/871 (90.4)	84/871 (9.6)
Repeatability difference FEV1 and FVC	<6	10% or 0.1 L	480/637 (75.4)	298/342 (87.1)	44/342 (12.9)	182/295 (61.7)	113/295 (38.3)
	≥6	0.15 L	2160/2496 (86.5)	1731/1920 (90.2)	189/1920 (9.8)	429/576 (74.5)	147/576 (25.5)
	All		2640/3133 (84.3)	2029/2262 (89.7)	233/2262 (10.3)	611/871 (70.1)	260/871 (29.9)
FET	<6		637/637 (100)	342/342 (100.0)	0/342 (0.0)	295/295 (100.0)	0/295 (0.0)
	6–10	3 s	606/865 (70.1)	479/617 (77.6)	138/617 (22.4)	127/248 (51.2)	121/248 (48.8)
	10–19	6 s	258/1631 (15.8)	240/1303 (18.4)	1063/1303 (81.6)	18/328 (5.5)	310/328 (94.5)
	All		1501/3133 (47.9)	1061/2262 (46.9)	1201/2262 (53.1)	440/871 (50.5)	431/871 (49.5)
Combination of repeatability and BEV	<6		476/637 (74.7)	298/342 (87.1)	44/342 (12.9)	178/295 (60.3)	117/295 (39.7)
	≥6		2111/2496 (84.6)	1728/1920 (90.0)	192/1920 (10.0)	383/576 (66.5)	193/576 (33.5)
	All		2587/3133 (82.6)	2026/2262 (89.6)	236/2262 (10.4)	561/871 (64.4)	310/871 (35.6)
Combination of repeatability, BEV and FET	<6		476/637 (74.7)	298/342 (87.1)	44/342 (12.9)	178/295 (60.3)	117/295 (39.7)
	≥6		767/2496 (30.7)	659/1920 (34.3)	1261/1920 (65.7)	108/576 (18.8)	468/576 (81.2)
	All		1243/3133 (39.7)	957/2262 (42.3)	1305/2262 (57.7)	286/871 (32.8)	585/871 (67.2)

Data are presented as n/N (%). BEV: back-extrapolated volume; FEV1: forced expiratory volume in 1 s; FVC: forced vital capacity; FET: forced expiratory time. #: n=3133 healthy children from the LUNOKID study; ¶: n=2262; +: n=871.
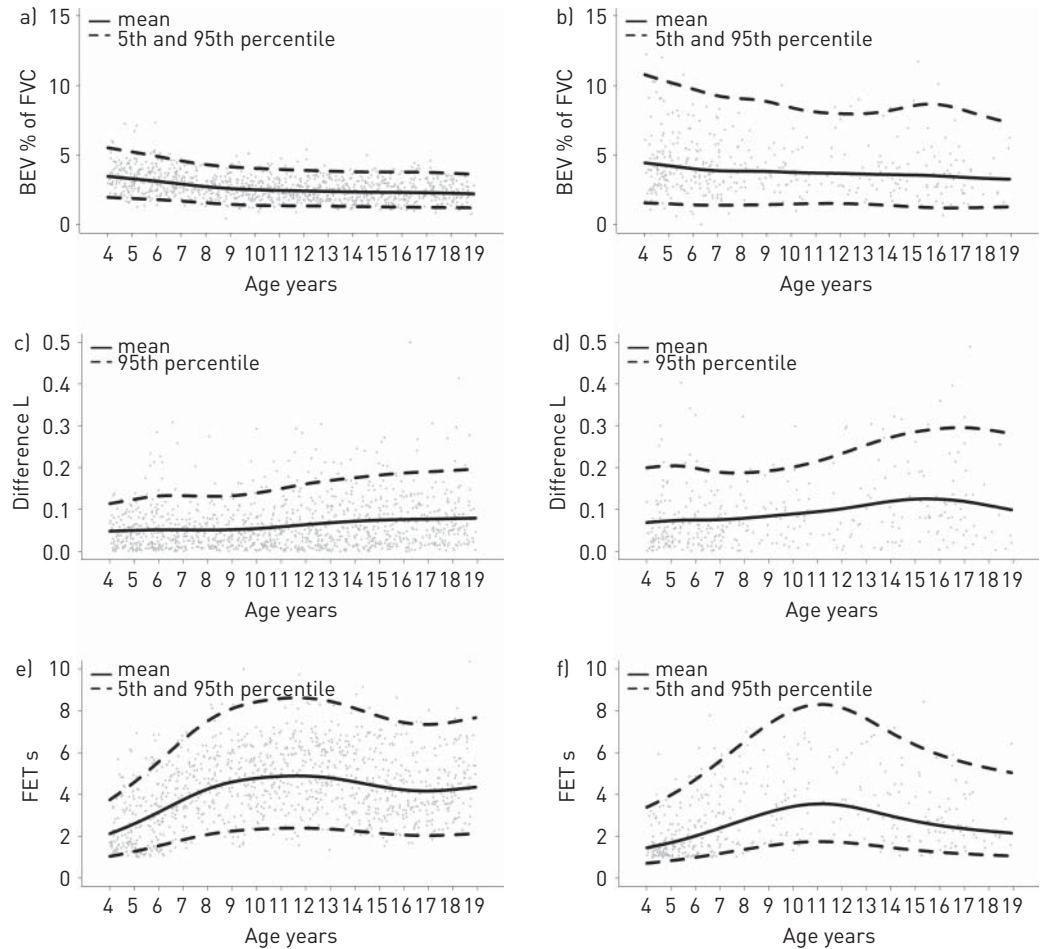
FIGURE 2 Age dependency of a) and b) back-extrapolated volume (BEV), c) and d) repeatability (shown as the difference between the two highest forced vital capacities (FVCs)) and e) and f) forced expiratory time (FET) in males defined by visual acceptability, where a), c) and e) are visually acceptable and b), d) and f) are visually unacceptable.

relation to lung volume [3, 21, 22, 24–27]. The growth spurt of the lung is known to occur out of phase with that of the body, which is even more pronounced in males than in females [28, 29].

Although the actual ATS/ERS recommendations for FET are not feasible in a paediatric population, length of expiratory time is positively associated with visual acceptability. However, the specificity of the best cut-off is low. Taken together, both the age dependency of FET and the low specificity of any cut-off value for

TABLE 4 Characteristics of American Thoracic Society/European Respiratory Society quality measurements in groups defined by visual acceptability[#]

	Visually acceptable	Visually unacceptable	p-value[¶]	Total
Subjects n	2262	871		3133
BEV L	0.07 (0.05)	0.08 (0.07)	<0.0001	0.07 (0.05)
BEV % FVC	2.70 (1.16)	3.88 (2.41)	<0.0001	2.92 (1.47)
Difference FEV1 L	0.04 (0.05)	0.05 (0.09)	<0.0001	0.04 (0.06)
Difference FVC L	0.04 (0.06)	0.06 (0.09)	<0.0001	0.05 (0.07)
FET s	3.95 (2.39)	2.30 (2.06)	<0.0001	3.54 (2.63)

Data are presented as median (interquartile range), unless otherwise stated. BEV: back-extrapolated volume; FVC: forced vital capacity; FEV1: forced expiratory volume in 1 s; FET: forced expiratory time. [#]: n=3133 healthy children from the LUNOKID study; [¶]: differences between groups using Wilcoxon rank sum test.
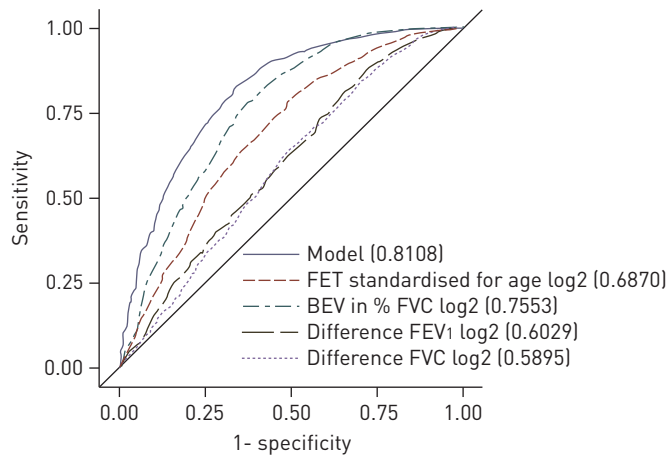
FIGURE 3 Receiver operating characteristic (ROC) curves of different combinations of quality measurements to predict visual acceptability. FET: forced expiratory time; BEV: back-extrapolated volume; FEV1: forced expiratory volume in 1 s; FVC: forced vital capacity. See the online supplementary material for further explanation.

FET compared with the gold standard of visual acceptability do not support current international recommendations for the use of FET as a quality criterion.

The majority (97%) of our subjects were able to meet the ATS/ERS criterion of BEV to assess the start-of-test part of the forced expiratory manoeuvre. Age dependency was hardly visible. BEV is associated with visual acceptability. A lower BEV increases the chance of a manoeuvre being visually acceptable. ROC analysis gave a "best" value for BEV in our population close to the recommended 5%.

Other studies reported similar feasibility rates of BEV, emphasising that young age and neuromuscular problems may yield lower feasibility rates [3–5, 30].

Compared with BEV, the ATS/ERS criterion for repeatability was slightly more difficult to meet (84%). Particularly in adolescent males, the ATS/ERS criterion of repeatability was difficult, even if their curves were visually acceptable. This again may be explained by setting a fixed cut-off point of 150 mL if FVC >1 L, independent of the measured actual lung volume. However, our data suggest that the variance between measurement increases with age and increasing lung volumes advocate the use of percentage rather than an absolute value for repeatability. This is particularly relevant in male adolescents with higher lung volumes, where visually acceptable curves with high FVCs may not be repeatable according to current NQC. The ROC curves for our healthy population show that the association between repeatability and visual acceptability is lower than the association between FET or BEV with visual acceptability. Repeatability is a quality criterion that is largely independent of visual acceptability. Subjects may perform a test poorly over and over again and produce only one visual acceptable curve, thus not meeting the criteria for repeatability. Reports concerning rate of repeatability vary from 55% to 98% depending on age and experience of the subjects [3–5, 31].

In order to find better cut-off values for the criteria to identify acceptable manoeuvres, a ROC analysis was performed including all quantitative measurements separately, as well as a combination thereof. A combination of all three numerical quality control parameters (BEV% of FVC, FET and repeatability) reached good overall concordance with good sensitivity but still poor specificity. Application of this formula would lead to misinterpretation of almost half of the curves as acceptable although they were not visually acceptable.

As expected, the specificity of the numerical parameters taken separately was low compared with the complex approach taken to assess visual acceptability. However, even an optimised combination of all three NQC cannot replace visual inspection. An as yet unidentified parameter different from BEV% of FVC, FET or repeatability, or a combination thereof would be necessary to replace visual control using computer software. Currently, visual quality control by an experienced lung function specialist remains superior. However, in everyday practice, an experienced observer will not always be available and inaccurate interpretation of spirometric curves might lead to false therapeutic decisions [31]. Training of technicians who are responsible for achieving the best possible results in an individual should focus on NQC, bearing in mind that unacceptable curves might still be usable, a decision that can be delegated fully to neither the computer nor the technicians. Nevertheless, better additional measures, including detection of artefacts such as cough, glottis closure or hesitation, and preferably avoiding arbitrary cut-off points, have to be identified that will be able to support rather than replace the clinical expert's judgement. Such measures

could focus, for example, on the pattern of expiratory flows, such as any point in comparison to the preceding flow with none topping the peak flow, thereby excluding "twin peak flows" or coughing, which produces a suddenly higher flow. More difficult to detect are cases that are "borderline" with regard to start and end of test. Especially when children and adolescents are included, growth and development have to be considered such that future quality control measures should opt for percentage and avoid absolute values. FET should be abandoned completely except for choosing the timed forced expiratory volume appropriate for the particular individual (*i.e.* $FEV_{0.75}$ or $FEV_{0.5}$ in young children, where lung size and relationship to airways explain short expiratory times and make the use of FEV_1 inappropriate). FET could potentially be replaced by defining a plateau, but a more complicated computerised description of the final part of the forced expiratory flow–volume curve is likely to be necessary in younger subjects with rapid lung emptying.

Potential limitations of our study

In order to select a healthy population for the development of reference values, all subjects with lower airway infection during the last 6 weeks, as well as infection on the day of testing, asthma or bronchitis, were excluded according to protocol. As three-quarters of the study took place during the winter, this led to a fairly high exclusion rate.

The rate of feasibility in our study was lower than in some of the previous studies. We think it is unlikely to be explained by our selection method, as we found no difference in mean FVC or FEV_1 over age when applied to visually acceptable manoeuvres (online supplementary figs E5–E8). In order to compare visual quality control of a single manoeuvre with the three NQCs we performed, however, we did not include "partly" acceptable curves (*i.e.* curves with only a good start of test); this decreased the rate of usable FEV_1, but not FVC and FET values. Several study groups have previously demonstrated that paediatric spirometry can reliably be performed outside the lung function laboratory [12]. All our study teams had received appropriate (central) training and had received regular supervision throughout the study. However, the present study was performed in spirometry-naïve subjects under field conditions. Even the youngest children in kindergarten attended the study without their parents. Distractions could have led to a reduced attention span. In addition, quality control was strict in order to obtain solid reference data, which would sustain critical review from international groups.

To our knowledge, this is the biggest study questioning international recommendations for quality control in spirometry in children and adolescents. We consider the current data relevant because measurements were performed under field conditions in subjects with no prior experience in lung function testing. Therefore, it is likely that the curves included in our analysis reflect the entire spectrum of quality.

Conclusions

From the data presented, we conclude that NQC criteria based on BEV% of FVC, FET and repeatability cannot replace visual inspection when assessing the quality of forced expiratory manoeuvres. Our study enhances current knowledge that, in contrast to BEV and repeatability, achieving the ATS/ERS recommended FET is neither feasible nor reasonable, and the use of FET should be abandoned. The low specificity of any combination of the three computer-based quality control measurements suggests that currently recommended criteria cannot replace visual inspection by an expert; up to 50% of visually unacceptable curves would be accepted if only computerised quality control was relied upon, thus emphasising that these criteria are only meant as drivers to aid a technician to obtain the best possible results, and not as pass–fail criterion. It remains a challenge for the respiratory community to develop better parameters and algorithms to increase validity of computerised quality control for paediatric spirometry.

References

1   Beydon N, Davis SD, Lombardi E, *et al.* An official American Thoracic Society/European Respiratory Society statement: pulmonary function testing in preschool children. *Am J Respir Crit Care Med* 2007; 12: 1304–1345.

2   Miller MR, Hankinson J, Brusasco V, *et al.* Standardisation of spirometry. *Eur Respir J* 2005; 2: 319–338.

3   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; 4: 655–660.

4   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; 10: 1152–1159.

5   Gaffin JM, Shotola NL, Martin TR, *et al.* Clinically useful spirometry in preschool-aged children: evaluation of the 2007 American Thoracic Society Guidelines. *J Asthma* 2010; 7: 762–767.

6   Loeb JS, Blower WC, Feldstein JF, *et al.* Acceptability and repeatability of spirometry in children using updated ATS/ERS criteria. *Pediatr Pulmonol* 2008; 10: 1020–1024.

7   Nystad W, Samuelsen SO, Nafstad P, *et al.* Feasibility of measuring lung function in preschool children. *Thorax* 2002; 12: 1021–1027.

8   Turner SW, Craig LC, Harbour PJ, *et al.* Spirometry in 5-year-olds – validation of current guidelines and the relation with asthma. *Pediatr Pulmonol* 2007; 12: 1144–1151.

9   Vilozni D, Barak A, Efrati O, *et al.* The role of computer games in measuring spirometry in healthy and "asthmatic" preschool children. *Chest* 2005; 3: 1146–1155.

10   Eigen H, Bieler H, Grant D, *et al.* Spirometric pulmonary function in healthy preschool children. *Am J Respir Crit Care Med* 2001; 163: 619–623.

11   Lum S, Stocks J. Forced expiratory manoeuvres. *In:* Frey U, Merkus PJFM, eds. Paediatric Lung Function. *Eur Respir Monogr* 2010; 47: 46–65.

12   Kirkby J, Welsh L, Lum S, *et al.* The EPICure study: comparison of pediatric spirometry in community and laboratory settings. *Pediatr Pulmonol* 2008; 12: 1233–1241.

13   Miller MR, Hankinson J, Brusasco V, *et al.* Standardisation of lung function testing: the authors' replies to readers' comments. *Eur Respir J* 2010; 6: 1496–1498.

14   Neve V, Matran R, Baquet G, *et al.* Quantification of shape of flow–volume loop of healthy preschool children and preschool children with wheezing disorders. *Pediatr Pulmonol* 2012; 9: 884–894.

15   Zheng CJ, Adams AB, McGrail MP, *et al.* A proposed curvilinearity index for quantifying airflow obstruction. *Respir Care* 2006; 1: 40–45.

16   Berdel D, Beckmann C, von BA, *et al.* Erhebung von Lungenfunktionsnormalwerten (Spirometrie) bei Kindern und Jugendlichen in Deutschland: Die LUNOKID-Studie [New spirometric reference data for children and adolescents in Germany: the LUNOKID study]. *Atemw-Lungenkrkh* 2010; 36: 395–404.

17   Huls A, Kramer U, Gappa M, *et al.* Neue spirometrische Referenzwerte für Kinder und Jugendliche in Deutschland ünter Berucksichtigung der Grösse und nichtlinearer Alterseffekte: Die LUNOKID-Studie [New spirometric reference values for children and adolescents in Germany considering height and non-linear age effects: the LUNOKID-study]. *Pneumologie* 2013; 3: 141–149.

18   Crenesse D, Berlioz M, Bourrier T, *et al.* Spirometry in children aged 3 to 5 years: reliability of forced expiratory maneuvers. *Pediatr Pulmonol* 2001; 1: 56–61.

19   Tomalak W, Radlinski J, Latawiec W. Jakość badania spirometrycznego u dzieci 10-letnich i mlodszych w świetle zaleceń standaryzacyjnych [Quality of spirometric measurements in children younger than 10 years of age in the light of the recommendations]. *Pneumonol Alergol Pol* 2008; 6: 421–425.

20   Gochicoa-Rangel L, Vargas-Dominguez C, Garcia-Mujica ME, *et al.* Quality of spirometry in 5-to-8-year-old children. *Pediatr Pulmonol* 2013; 48: 1231–1236.

21   Enright PL, Linn WS, Avol EL, *et al.* Quality of spirometry test performance in children and adolescents: experience in a large field study. *Chest* 2000; 3: 665–671.

22   Merkus PJ, ten Have-Opbroek AA, Quanjer PH. Human lung growth: a review. *Pediatr Pulmonol* 1996; 6: 383–397.

23   Quanjer PH, Stanojevic S, Stocks J, *et al.* Changes in the FEV_1/FVC ratio during childhood and adolescence: an intercontinental study. *Eur Respir J* 2010; 6: 1391–1399.

24   Stanojevic S, Wade A, Cole TJ, *et al.* Spirometry centile charts for young Caucasian children: the Asthma UK Collaborative Initiative. *Am J Respir Crit Care Med* 2009; 6: 547–552.

25   Stanojevic S, Wade A, Stocks J, *et al.* Reference ranges for spirometry across all ages: a new approach. *Am J Respir Crit Care Med* 2008; 3: 253–260.

26   Zapletal A, Chalupova J. Forced expiratory parameters in healthy preschool children (3–6 years of age). *Pediatr Pulmonol* 2003; 3: 200–207.

27   Becklake MR, Kauffmann F. Gender differences in airway behaviour over the human life span. *Thorax* 1999; 12: 1119–1138.

28   Quanjer PH, Stocks J, Polgar G, *et al.* Compilation of reference values for lung function measurements in children. *Eur Respir J* 1989; 2: Suppl. 4, 184S–261S.

29   Degroodt EG, Quanjer PH, Wise ME, *et al.* Changing relationships between stature and lung volumes during puberty. *Respir Physiol* 1986; 2: 139–153.

30   Vilozni D, Berkun Y, Levi Y, *et al.* The feasibility and validity of forced spirometry in ataxia telangiectasia. *Pediatr Pulmonol* 2010; 10: 1030–1036.

31   Gillette C, Loughlin CE, Sleath BL, *et al.* Quality of pulmonary function testing in 3 large primary care pediatric clinics in rural North Carolina. *NC Med J* 2011; 2: 105–110.