Reliable and Valid Assessment of Clinical Bronchoscopy Performance

Lars Konge  Klaus Richter Larsen  Paul Clementsen  Henrik Arendrup  Christian von Buchwald  Charlotte Ringsted
Centre for Clinical Education, University of Copenhagen and Capital Region of Denmark, Copenhagen, Denmark

Key Words
Assessment · Bronchoscopy · Education · Procedural training

Abstract

Background: There have been several attempts to systematically assess performance in bronchoscopy. Earlier validation studies have used bronchoscopy simulators, not real-life performance in patients. Objectives: The aim of this study was to explore the reliability and validity of an assessment tool aimed for the use in a clinical setting. Methods: Five junior residents, 5 senior residents and 9 consultants performed 3 bronchoscopies each. All 57 bronchoscopies were video-recorded and assessed blindly and independently by two bronchoscopy experts using the new assessment tool. Results: The interrater reliability was high, with Cronbach’s α = 0.86. Assessment of 3 bronchoscopies by a single rater had a generalizability coefficient of 0.84. The correlation between experience and performance was good (Pearson correlation = 0.76). There were significant differences between the groups for all aspects of the assessment, but post hoc tests showed different discriminative abilities. Conclusions: This new tool for assessing clinical bronchoscopy performance has a high interrater reliability. One rater assessing performance of 3 bronchoscopies ensures sufficient reliability. The assessment tool demonstrated sufficient construct validity.

Introduction

Assessment of competence is playing an increasing role in education and certification of physicians’ competence. There are four important reasons to use reliable and valid assessment methods: to certify the competence of future practitioners, discriminate among candidates for advanced training, provide motivation and direction for learning (feedback), and judge the adequacy of training programs [1].

The growing concern for patient safety has increased focus on the question when basic competence is reached and a physician is ready to perform procedures on his or her own in an efficient and safe way. Studies indicate that, on average, Americans only receive about half the recommended medical care [2] and that errors cause between 44,000 and 98,000 deaths every year in American hospitals [3]. In the program ‘Maintenance of Certification’, the American Board of Surgery has recently added a part about Assessment of Performance in Practice. This is a shift from relying on an individual’s qualifications based on time in training and clinical experience toward measuring competence. The American Board of Surgery recognizes that defining and measuring surgeon competence may continue to be difficult and encourages physicians to engage in this effort [4].

In areas as diverse as gynecology, vascular surgery, urology, open surgery, laparoscopy, endoscopy and endo-
Flexible bronchoscopy is a complex and potentially hazardous procedure that requires technical skills and good hand-eye coordination. Inexperienced bronchoscopists have a higher complication rate [12] and a lower yield of positive biopsy material [13]. International guidelines dictate that basic competency is established when a trainee has performed at least 100 procedures in a supervised setting [14, 15]. However, it may be questioned whether trainees after completing the prescribed number of procedures are in fact proficient in bronchoscopy. Indeed, studies using newly developed tools for assessment of performance in bronchoscopy, upper endoscopy and colonoscopy demonstrate that current case recommendations do not represent what is actually needed to become proficient [16, 17]. These findings underscore the need to establish when basic competence in bronchoscopy is reached.

Earlier attempts to assess bronchoscopy skills vary from nonvalidated assessments by either a single thoracic surgeon [18] or an experienced nurse [19] to the ambitious Bronchoscopy Skills and Tasks Assessment Tool created by bronchoscopy experts at the University of California [20]. This tool requires a trained assessor to directly observe the trainee and hence includes a risk of rater bias [16, 21]. Validity of this tool was studied using performance on a bronchoscopy simulator and demonstrated significant differences when comparing complete novices (who had never operated a bronchoscope) to fellows and attendings. Unfortunately, there was no statistical difference in the Bronchoscopy Skills and Tasks Assessment Tool scores of fellows and attendings, indicating a ceiling effect [21]. Hence, for the purpose of assessing clinical bronchoscopy performance of trainees of different levels we designed a new assessment tool.

The aim of this study was to explore the reliability and validity of this new tool for assessing clinical bronchoscopy performance. The research questions were as follows:

– What is the interrater reliability of the assessment tool?

– What is the optimal number of raters and cases assessed in order to achieve a reliable assessment of performance?

– What is the construct validity of the assessment tool in terms of being able to differentiate between physicians with varying experience?

– How do individual elements of performance anchored to the assessment tool relate to stages of experience?

**Materials and Methods**

**Development of the Assessment Tool**

The assessment tool was developed by a team of 5 physicians – 3 clinical experts (P.C., K.R.L., Dr. Martin Iversen), 1 expert in educational measurement (C.R.) and 1 person having experience in both fields (L.K.). The group of clinical experts agreed on including three key elements of bronchoscopy competence in the assessment form: technical ability (the ability to move around the bronchial tree in a gentle and competent way), demonstrating anatomical knowledge (performing a systematic and thorough procedure with visualization of all 19 bronchial segments), and ability to describe the procedure and the findings. Due to the complex nature of the procedure we chose to use a rating scale instead of a checklist and provided anchors for the different rating options [22]. To improve reliability we included as many items in the assessment form as possible within practical limits [23]. This made each item rather specific and hence we limited the number of response categories to three. An exception was the rating of the description of the procedure, where we included 5 categories. To further investigate the ability to correctly describe anatomical findings we included an already validated rating scale, the Thompson Bronchitis Index [24]. The first draft of the assessment form was carefully revised according to a pilot study assessing three individuals’ bronchoscopy performance. The final form is shown in table 1.

**The Validation Process**

**Sample**

The necessary sample sizes for validation studies are debatable [23]. We chose to follow the recommendation of Cicchetti and Fleiss [25] who argue that in most situations with two raters, a sample size of $2k^2$, where k is the number of points on the scale, should be sufficient. Using a three-point scale we should include at least $2 \times 3^2 = 18$ participants. For validation purposes we included physicians with varying bronchoscopy experience: junior residents, senior residents and consultants.

**Assessment Procedure**

Before the assessment procedure the participating physicians received written information which explained the issues that were going to be assessed. Participants were asked to give their demographic information and signed informed consent. Participants were instructed to make a standard, systematic examination of the bronchial tree and to freeze the picture at six different locations: in the right upper lobe, in the middle lobe, in the basal part of the right lower lobe, in the apical part of the left upper lobe, in the lingula, and in the basal part of the left lower lobe. This was aimed to aid the assessment of anatomical knowledge. A list of the six locations was available throughout the procedure. The physi-
cians were informed to complete the diagnostic bronchoscopy and return to the trachea before any sampling procedures were performed. After the bronchoscopy each physician filled out the Thompson Bronchitis Index and wrote a standard description of the procedure and findings. The bronchoscopies were recorded using either a MediCap™ USB200 medical recorder (MediCapture), a Sony DVD recorder (Sony Corp.) or a Verbatim™ Mediasation HD DVR network multimedia recorder 1 TB (Verbatim) depending on the institution.

Each physician performed bronchoscopy on 3 different patients, which made a total of 57 datasets each consisting of a video recording, a completed Thompson Bronchitis Index and a description of the procedure.

The recordings were trimmed on both ends so they started with the insertion of the bronchoscope and ended when the bronchoscope returned to the trachea. The procedure descriptions were rewritten on blank paper omitting the name of the patient, the institution and the performing physician.

Table 1. The bronchoscopy assessment tool

<table>
<thead>
<tr>
<th>Assessment form</th>
<th>Score: 0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scope insertion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Administration of LA</td>
<td>Missing or inadequate administration of LA</td>
<td>Acceptable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passage through vocal cords</td>
<td>Needs several attempts to pass vocal cords</td>
<td>Acceptable</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Perfect administration of LA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Right side</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right upper lobe</td>
<td>False identification</td>
<td>Uncertain but correct identification</td>
<td>Secure and correct identification</td>
<td></td>
</tr>
<tr>
<td>Middle lobe</td>
<td>False identification</td>
<td>Uncertain but correct identification</td>
<td>Secure and correct identification</td>
<td></td>
</tr>
<tr>
<td>Right lower lobe</td>
<td>False identification</td>
<td>Uncertain but correct identification</td>
<td>Secure and correct identification</td>
<td></td>
</tr>
<tr>
<td><strong>Left Side</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left upper lobe – apical</td>
<td>False identification</td>
<td>Uncertain but correct identification</td>
<td>Secure and correct identification</td>
<td></td>
</tr>
<tr>
<td>Lingula</td>
<td>False identification</td>
<td>Uncertain but correct identification</td>
<td>Secure and correct identification</td>
<td></td>
</tr>
<tr>
<td>Left lower lobe</td>
<td>False identification</td>
<td>Uncertain but correct identification</td>
<td>Secure and correct identification</td>
<td></td>
</tr>
<tr>
<td><strong>Scope movement</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collisions</td>
<td>Repeated scope collisions</td>
<td>Few collisions</td>
<td>No unnecessary scope collisions</td>
<td></td>
</tr>
<tr>
<td>Red-out</td>
<td>Unacceptable length of time in ‘red-out’</td>
<td>Minimal time in ‘red-out’</td>
<td>No time in ‘red-out’</td>
<td></td>
</tr>
<tr>
<td>Centering</td>
<td>Frequent pointing away from lumen</td>
<td>Scope centered most of the time</td>
<td>Scope always centered</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Thompson Bronchitis Index</strong></td>
<td>Normal</td>
<td>Light red</td>
<td>Red</td>
<td>Beefy red</td>
</tr>
<tr>
<td></td>
<td>Blunting of airway bifurcations</td>
<td>Loss of normal airway wall indentations</td>
<td></td>
<td>Airway occluded</td>
</tr>
<tr>
<td>Secretions</td>
<td>Normal</td>
<td>Strands of clear mucus</td>
<td>Glob of mucus</td>
<td>Airway occluded</td>
</tr>
<tr>
<td>Friability</td>
<td>Normal</td>
<td>Punctate submucosal hemorrhages with scope trauma</td>
<td>Linear submucosal hemorrhages with scope trauma</td>
<td>Frank bleeding with scope trauma</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Score: 0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td><strong>Description of procedure</strong></td>
<td>The description is</td>
<td>Unacceptable</td>
<td>Poor</td>
<td>Acceptable</td>
</tr>
</tbody>
</table>

Rating of Performance

Two raters (K.R.L. and P.C.) scored each performance individually. Each rater received a hard drive with the recordings and a folder with numbered assessment forms as well as associated descriptions of the bronchoscopy. The raters viewed each recording and filled out the first 11 items of the assessment form and the Thompson Bronchitis Index before reading the description of the procedure and filling out the last item. The ability to correctly describe findings was assessed by comparing the Thompson Bronchitis Index filled out by the physician to that of the expert (used as a gold standard): total agreement scored 2 points, one point off scored 1 point and more than one point off scored zero points. Finally, scores for technical abilities, anatomical knowledge and ability to describe the procedure as well as findings were summed up, and the total score was divided by the duration of the procedure to calculate score per minute (the bronchoscopy quality score).
Statistical Analysis

Interrater reliability was analyzed by intraclass correlation and reported by Cronbach’s $\alpha$ for the total score and the three subgroups of items. To examine optimal balance between number of bronchoscopies performed and number of raters used to get a reliable measure of bronchoscopy performance we used generalizability analyses as described by Brennan [26] in 2001. We adopted a two-stage process: in the first stage we obtained estimated variance components for a G study (two-facet, balanced design with test persons crossed with bronchoscopies crossed with raters, that is $p \times b \times r$). The second stage involved using these estimated variance components in the context of a D study design to estimate generalizability coefficients (a reliability-like coefficient defined by Cronbach et al. [27] in 1972) according to number of bronchoscopies performed and raters used. We aimed at a generalizability coefficient above 0.8.

Construct validity was explored in two ways. Once by correlating the experience of the physicians to their total score, time and bronchoscopy quality score using the Pearson correlation. Furthermore, we used analysis of variance with Bonferroni post hoc tests to compare the performance of the three groups (the junior residents, the senior residents and the consultants).

Statistical analysis was performed using a statistical software package (PASW, version 18.0; SPSS Inc.). Differences were considered to be statistically significant when the $p$ value was $<0.05$. The G and D studies were performed using the GENOVA software ver. 3.1 (Crick and Brennan).

Ethics

Consultants supervised all bronchoscopies performed by residents, and no ethical approval was needed according to the Danish National Committee on Biomedical Research Ethics.

Results

The study sample included 19 physicians: 5 junior residents, 5 senior residents and 9 consultants. Table 2 shows the demographic data of the physicians (age, sex and number of previously performed bronchoscopies).

<table>
<thead>
<tr>
<th>Sex</th>
<th>Age</th>
<th>Bronchoscopies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Female</td>
<td>2012-2013</td>
</tr>
<tr>
<td>Junior residents (n = 5)</td>
<td>35 (27–43)</td>
<td>8 (3–12)</td>
</tr>
<tr>
<td>Senior residents (n = 5)</td>
<td>38 (34–41)</td>
<td>75 (50–120)</td>
</tr>
<tr>
<td>Consultants (n = 9)</td>
<td>51 (42–60)</td>
<td>1,506 (250–3,000)</td>
</tr>
<tr>
<td>Total (n = 19)</td>
<td>43 (27–60)</td>
<td>735 (3–3,000)</td>
</tr>
</tbody>
</table>

Table 2. Demographics of the participating physicians

Fig. 1. D study: the effect on the generalizability coefficient of assessing more procedures or adding more raters.

**Reliability**

The interrater reliability concerning the total scores in the 57 procedures was good, Cronbach’s $\alpha = 0.86$. Reliability was highest for technical abilities, Cronbach’s $\alpha = 0.86$, and anatomical knowledge. Cronbach’s $\alpha = 0.83$, and somewhat lower concerning the descriptions of the procedures, Cronbach’s $\alpha = 0.67$. Figure 1 shows the generalizability coefficient according to the number of procedures assessed and the number of raters used. The results of the G and D study demonstrate that assessment of three procedures using a single rater secures a generalizability coefficient above 0.80 (0.84) and is better than using two raters assessing one case.

**Validity**

Construct validity was supported by a positive correlation between the experience of the physicians (number of performed bronchoscopies, logarithmic) and the average total scores (across the two raters), $r = 0.71$, and a negative correlation between the experience of the physicians and
the time spent on the procedures, \( r = -0.72 \). The correlation between the experience of the physicians and the bronchoscopy quality scores (score per minute) is plotted in figure 2, \( r = 0.76 \). Analysis of variance demonstrated significant differences between the three groups (junior residents, senior residents and consultants) for total score \( (p < 0.001) \), time spent \( (p < 0.001) \), bronchoscopy quality score \( (p < 0.001) \), technical abilities \( (p < 0.001) \), anatomical knowledge \( (p < 0.001) \), and ability to describe procedure and findings \( (p = 0.035) \). Table 3 shows comparisons between the individual groups using post hoc tests (Bonferroni). The variables’ total score, time and quality scores sufficient discriminated between the three groups. The subelement ‘technical ability’ discriminated significantly between senior residents and consultants, and the subelement ‘anatomical knowledge’ discriminated between junior and senior residents. However, the subelement ‘description of procedure and findings’ did not discriminate between any of the three groups. The probability density curves for technical abilities, anatomical knowledge, and ability to describe procedure and findings are plotted in figure 3a–c.

**Discussion**

There are several studies dealing with assessment of bronchoscopy skills [16, 28–31], but this is the first study reporting on reliability and validity of blinded assessments of video recordings of bronchoscopies performed on real patients. The interrater reliability was high and our generalizability analyses confirmed this finding and showed that it is okay to use a single rater, given this is a trained rater. However, it is important to realize that decisions regarding bronchoscopy skills cannot be based on assessing one single bronchoscopy performance. Differences in performance related to patient, milieu, state of the physician and so on require the assessment of three bronchoscopies as a minimum.

Our assessment tool deals with the physicians’ technical ability to maneuver the scope, his or her anatomical knowledge, and the ability to describe the procedure and the findings in a correct and meaningful way. The best way to differentiate between trainees and experts is by assessing the technical abilities (fig. 3a), whereas the anatomical knowledge better differentiates trainees at different stages of learning (fig. 3b). The ability to describe the procedure and the findings was almost the same regardless of bronchoscopy experience (fig. 3c), and this accords well with a previous study that found no correlation be-

---

**Table 3.** Comparisons between the individual groups concerning total score, time spent, quality score (score/min), technical abilities, anatomical knowledge, and ability to describe procedure and findings (Bonferroni post hoc tests)

<table>
<thead>
<tr>
<th></th>
<th>JR mean ± SD</th>
<th>SR mean ± SD</th>
<th>JR vs. SR</th>
<th>C mean ± SD</th>
<th>JR vs. C</th>
<th>SR vs. C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total score</td>
<td>16.6 ± 5.02</td>
<td>22.6 ± 4.40</td>
<td>0.001*</td>
<td>26.2 ± 4.10</td>
<td>&lt;0.001*</td>
<td>0.044*</td>
</tr>
<tr>
<td>Time, min</td>
<td>10.9 ± 3.62</td>
<td>7.4 ± 3.05</td>
<td>0.001*</td>
<td>4.8 ± 1.33</td>
<td>&lt;0.001*</td>
<td>0.008*</td>
</tr>
<tr>
<td>Quality score</td>
<td>1.8 ± 1.00</td>
<td>3.46 ± 1.3</td>
<td>0.018*</td>
<td>6.0 ± 2.01</td>
<td>&lt;0.001*</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>Technical abilities</td>
<td>5.5 ± 2.03</td>
<td>4.9 ± 2.22</td>
<td>0.218</td>
<td>7.6 ± 1.96</td>
<td>&lt;0.001*</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>Anatomical knowledge</td>
<td>6.2 ± 3.42</td>
<td>9.5 ± 1.65</td>
<td>0.001*</td>
<td>10.4 ± 1.63</td>
<td>&lt;0.001*</td>
<td>0.634</td>
</tr>
<tr>
<td>Ability to describe procedure and findings</td>
<td>6.8 ± 1.81</td>
<td>8.2 ± 1.81</td>
<td>0.073</td>
<td>8.1 ± 1.65</td>
<td>0.059</td>
<td>1.000</td>
</tr>
</tbody>
</table>

JR = Junior residents; SR = senior residents; C = consultants. * Statistically significant difference.

---

**Fig. 2.** Correlation between experience and performance.
between bronchoscopy experience and performance on a multiple-choice test on bronchoscopy [30]. These findings suggest that the diagnostic bronchoscopy procedure is learned as a three-stage process: first the theoretical background is acquired – this does not require any practical training, but can be done by reading about the procedure and observing senior colleagues. Second, after the performance of relatively few procedures, the trainee has the anatomical knowledge to perform a thorough and systematic examination of the bronchial tree. Finally, the ability to move the scope in a gentle, efficient and seemingly effortless way requires extensive training.

The assessment of the description of the procedure and the Thompson Bronchitis Index possessed the lowest interrater reliability, perhaps because it is the most subjective part of the assessment tool. Furthermore, these results were almost constant regardless of bronchoscopy experience (fig. 3c), meaning that this part of the tool had very little discriminative ability. This time-consuming part could be omitted from the tool without significant

---

**Fig. 3.** a Probability density curves: technical abilities. b Probability density curves: anatomical knowledge. c Probability density curves: ability to describe the procedure and findings.
changes in reliability or validity. The revised tool would consist of the first 11 items with a maximum score of 22 points.

Our video-based approach allows assessment by one or more raters that do not need to be present at the exact time and place of the procedure performance. However, the most important advantage is the avoidance of all the potential bias of human relations between rater and ratee: subjectivity, false impressions, the three ‘isms’ (ageism, racism, sexism), rumor, grudge and misinterpretation [32]. However, some important aspects are not possible to assess based on video recordings, such as posture, equipment handling, and the ability to select patients or perform and interpret history, clinical examination and imaging results. These abilities must still be assessed using direct observation by a skilled supervisor. Furthermore, we chose to omit the assessment of different bronchial procedures (that is, lavages and biopsies) from this validation study. The reasons were that this inclusion would make the procedures studied much more heterogeneous with great differences in difficulty, and that many novices/junior trainees are not allowed to perform these procedures. We suggest that these abilities are assessed using a virtual reality bronchoscopy simulator where procedures can be performed in a standardized manner without risk to the patients [31].

Simulation-based education in bronchoscopy is growing. A decade ago pulmonary fellows learning bronchoscopy almost entirely depended on one-on-one instruction by faculty [33]. Since then more and more institutions have acquired a virtual reality bronchoscopy simulator [34]. These have been sold at prohibitive prices, which have encouraged groups to make paper mache models [35], use porcine or rubber models [36, 37], or develop their own virtual reality bronchoscopy simulator [38]. Fortunately, new professional competition [39] and small, portable systems [40] are emerging which are bound to lower the prices and thereby increase the prevalence of bronchoscopy simulators. However, before investing time and money on different simulators, it is important to prove their effect, preferably by randomized controlled studies. Our reliable and valid assessment tool could be used in those trials.

Assessing video performance can be administered in an unbiased way and hence be used for high stakes examinations – certification and re-certification. This requires the creation of a defensible passing score. Such a standard setting must be based on the opinions of experts in the field and is a matter for future research. To further validate the tool it would be interesting to correlate the obtained scores with clinical outcome – however, that would require the assessment of a very large number of procedures.

There are several weaknesses in our study. The number of participants is limited. Even though we included almost all trainees from three large hospitals, we would have liked to include more inexperienced bronchoscopists. A better powered study would likely find a statistically significant difference between the technical abilities of junior and senior residents, and perhaps also a difference between the anatomical knowledge of senior residents and consultants. We have not proved the generalizability of the assessment tool, but the video-based approach and the high interrater reliability between raters from different institutions across the country suggest that the tool can be used in other places and by other raters. Future studies are necessary to explore the usability of the tool for a variety of purposes. Can it be used to accelerate learning by providing structured feedback? Can the assessment tool measure the effect of new training methods, that is, simulation-based training? Perhaps the assessment tool can also be used to select candidates for advanced training, that is, endoscopic ultrasound procedures.

**Conclusion**

This new tool for assessing clinical bronchoscopy performance has a high interrater reliability and requires only a single rater. It is necessary to assess three bronchoscopies to achieve a reliable assessment. Measuring anatomical knowledge differentiates between novices and more experienced trainees while measuring technical abilities differentiates between trainees and experts. The first 11 items of the assessment tool precisely differentiate between physicians with varying experience.

**Acknowledgement**

We thank Dr. Martin Iversen for his aid in the construction of the assessment tool.

**Financial Disclosure and Conflicts of Interest**

This work was in part supported by the Research Foundation of the Heart Center, Rigshospitalet. Dr. Konge, Dr. Larsen, Dr. Clemensen, Dr. Arendrup, Prof. von Buchwald and Prof. Ringsted have no conflicts of interest.
References


