Establishing Pass/Fail Criteria for Bronchoscopy Performance

Lars Konge    Paul Clementsen    Klaus Richter Larsen    Henrik Arendrup
Christian Buchwald    Charlotte Ringsted

Centre for Clinical Education, University of Copenhagen and the Capital Region of Denmark, Copenhagen, Denmark

Key Words
Assessment tools · Bronchoscopy · Education · Procedural training · Standard setting

Abstract
Background: Several tools have been created to assess competence in bronchoscopy. However, educational guidelines still use an arbitrary number of performed procedures to decide when basic competency is acquired. Objectives: The purpose of this study was to define pass/fail scores for two bronchoscopy assessment tools, and investigate how these scores relate to physicians’ experience regarding the number of bronchoscopy procedures performed. Methods: We studied two assessment tools and used two standard setting methods to create cut scores: the contrasting-groups method and the extended Angoff method. In the first we compared bronchoscopy performance scores of 14 novices with the scores of 14 experienced consultants to find the score that best discriminated between the two groups. In the second we asked an expert group of 7 experienced bronchoscopists to judge how a borderline trainee would perform on each item of the test. Results: Using the contrasting-groups method we found a standard that would fail all novices and pass all consultants. A clear pass related to 75 procedures. The consequences of using the extended Angoff method were also acceptable: all trainees who had performed less than 50 bronchoscopies failed the test and all consultants passed. A clear pass related to 80 procedures.

Conclusions: Our proposed pass/fail scores for these two methods seem appropriate in terms of consequences. Prior experience with the performance of 75 and 80 bronchoscopies, respectively, seemed to ensure basic competency. In the future objective assessment tools could become an important aid in the certification of physicians performing bronchoscopies.
tant characteristics: they are set by appropriate numbers and types of judges, they utilize an appropriate method, and they produce reasonable outcomes.

The many standard-setting methods developed over the last 40 years can be divided into two broad categories: test-centered methods and examinee-centered methods. In test-centered methods, experts review the items in the test and decide on the level of performance required to meet a performance standard. The experts do not rate actual examinee performances when setting the standard. In the examinee-centered methods, performances of real examinees are evaluated relative to the performance standard [10]. There is no single correct or best standard setting method; the choice of method depends on the type of assessment data, feasibility, resources available, and the preferences of decision makers [11].

There have been no attempts to create a test score that would represent a pass/fail standard for competency in bronchoscopy, and guidelines rely on an arbitrary number of performed procedures. According to the American College of Chest Physicians, trainees should perform at least 100 flexible bronchoscopies in a supervised setting to establish basic competency, and at least 25 procedures per year to maintain competency [12]. The British Thoracic Society acknowledges that the optimal number of procedures performed under supervision will vary depending on the competency of the trainee. Nevertheless, they suggest a minimum of 50 procedures under direct supervision and a further 50 under indirect supervision [13]. The Accreditation Council for Graduate Medical Education in the United States has chosen to follow these recommendations and has increased the number of flexible bronchoscopies procedures from 50 to 100 in their new Program Requirements for Fellowship Education in Pulmonary Disease and Critical Care Medicine, effective July 1, 2011 [14]. These numbers are not evidence based and do not take into consideration that individual trainees learn at different paces. Furthermore, there is no reward for those trainees that spend hours on deliberate practice using bronchoscopy simulators, even though studies have shown that bronchoscopy simulation training enhanced skill acquisition among learners [15–17] and helped reduce repair costs [18, 19].

Our group has developed and validated two assessment tools for assessing competence in bronchoscopy [20, 21]. The first tool is designed to be administered on a virtual-reality bronchoscopy simulator and the second to be used in a clinical setting. Both tools are reliable and allow blinded assessment by one or several raters. They are significantly able to discriminate between trainees and experts. The purpose of this study was to establish pass/fail scores for the two assessment tools and explore how these standards relate to bronchoscopy experience, i.e. whether a certain number of performed procedures guarantee basic competence in bronchoscopy.

### Materials and Methods

**Simulation-Based Assessment of Bronchoscopy – Contrasting-Groups Method**

This assessment instrument was designed for performance on a virtual-reality bronchoscopy simulator (Endoscopy Accu-Touch® System with bronchoscopy modules; CAE Healthcare, Montreal, Que., Canada). Using the instrument we previously assessed a total of 45 persons comprising 14 medical students with no bronchoscopy experience, 17 trainees who had performed between 10 and 120 bronchoscopies, and 14 members of clinical staff who had performed more than 500 bronchoscopies each. The assessment procedure included having the participants complete six standardized, simulated procedures: first, 2 diagnostic bronchoscopies where the participants had to find tumors and state the localizations; next, 2 standardized bronchoalveolar lavages using 150 ml of saline in a segment of the middle lobe, and finally two procedures including the use of biopsy tools (needle, brush, and forceps). Performance was scored on a 28-point checklist [20]. Each person received an aggregate score, the bronchoscopy quality score, by dividing the total score by the amount of time spent.

For standard setting we chose an examinee-based method, the contrasting-groups method [22], because we could meaningfully create two contrasting groups based on an external criterion: one group of noncompetent performers (=medical students) and one group of competent performers (=members of staff). The standard was established by graphing the two score distributions and finding the score that best discriminated between the two groups. The pass score was set at the intersection of the distributions, as we considered false-positives (students passing) and false-negatives (members of staff failing) to be of equal weight. After determining the pass score we investigated the consequences of the standard for the 45 participants and explored how the standard related to participants’ prior experience in terms of the number of bronchoscopies performed.

**Assessment of Clinical Bronchoscopy Based on Video Recordings – Angoff Method**

This assessment instrument was constructed for assessment of clinically performed bronchoscopies. The instrument consisted of 16 items covering technical abilities (administration of local anesthesia, passage through vocal cords, scope collisions, time in red-out, and centering of the scope), anatomical knowledge (correct identification of the right upper lobe, middle lobe, right lower lobe, apical part of left upper lobe, lingula, and left lower lobe), description of findings (using the Thompson bronchitis index [23]), and description of the procedure. The first 15 items were rated on 3-point scales and the last item (concerning the description of the procedure) was rated on a 5-point scale. Using video recordings of bronchoscopy performances the test was administered to 19 physicians. The individual assessment scores were an
average of 3 bronchoscopies performed. All 57 bronchoscopies were rated in a blinded fashion by 2 independent raters.

We had no assessment data from a sizeable group of noncompetent or borderline-competent bronchoscopist, so we used a test-centered standard setting method: the extended Angoff method [24, 25]. With this method, content experts make judgments about every item, so it is fairly easy to defend the resulting pass score [11]. We formed a standard-setting panel of 7 highly experienced bronchoscopists working and teaching in 7 different university hospitals or university affiliated hospitals from all regions of Denmark. We described two performance categories to the experts: (1) needs more supervised practice and (2) able to perform bronchoscopies without supervision. A borderline examinee was defined as one who had an exactly 50/50 probability of ending up in either category, i.e. one who on some days might just pass the assessment, but on other days might fail. Each expert then estimated how a typical borderline examinee would perform on each item, and how much time he or she would spend on the procedure. These judgments were systematically combined (totaled and averaged) to determine the pass scores for technical ability, anatomical knowledge, and ability to describe procedure and findings. We used a noncompensatory approach where each component had to be passed separately because we wanted a trainee to reach a minimum level of competence in all three categories. To help calibrate the experts we provided actual performance data. This was done because experts tend to set unrealistically high pass scores – they almost always expect too much of novices [26].

Statistical Analysis
All data were computed using PASW Statistics ver. 18.

Results

Simulation-Based Assessment of Bronchoscopy – Contrasting-Groups Method
The 14 medical students scored 0.27 points/min (mean; SD = 0.065) and the 14 members of staff scored 0.65 points/min (mean; SD = 0.117). By graphing these score distributions we found the pass score to equal 0.42 points/min (fig. 1). There were no false-positives or false-negatives in our two groups as the highest-scoring medical student scored 0.40 points/min and the lowest-scoring member of staff scored 0.49 points/min. One trainee who
Assessment of Clinical Bronchoscopy Based on Video Recordings – Angoff Method

Table 1 shows the demographics of the expert group including their bronchoscopy and teaching experience. Table 2 shows the expert group’s opinion of the performance of a borderline trainee. The level of agreement concerning the scores was high, ICC = 0.89, but the experts had quite different opinions about how much time a borderline trainee would spend on a bronchoscopy: mean 15.8 min, range 8–30 min, and SD 9.7 min. As mentioned above, the pass scores for the different categories (technical abilities, anatomical knowledge, and ability to describe the procedure and findings) were found by averaging the opinions of the experts. Figure 3a–c shows the consequences of applying the pass scores to a group of 19 physicians with varying bronchoscopy experience. The cut score regarding technical abilities was very low and only 2 trainees failed this part of the test (fig. 3a). All inexperienced trainees failed the test regarding anatomical knowledge (fig. 3b), while all participants demonstrated sufficient ability to describe the procedure and findings. Overall, 1 trainee who had performed 75 bronchoscopies failed the test, but all physicians who had performed 80 bronchoscopies or more passed. None of the physicians failed because they spent too much time on average on the 3 bronchoscopies.

Discussion

This study shows that it is possible to set defensible standards for bronchoscopy skills that have meaningful consequences when applied to bronchoscopists with a varying degree of experience. The expert group agreed
quite well on the characteristics of a borderline trainee. They all found that a trainee on the verge of being able to perform bronchoscopies independently still has serious deficits concerning his or her technical abilities. This agrees well with the finding in our previous validation study of the assessment tool where no significant difference in the performance of junior and senior residents was found regarding these skills. The resulting pass score was quite low. The pass score concerning anatomical knowledge was much higher, indicating that the experts considered a borderline trainee relatively well educated in this area. All the experts underestimated the ability of a borderline trainee to describe the procedure and findings, resulting in a low pass score that all our previously tested physicians would meet. However, even novices were quite adept regarding the ability to correctly describe the procedure and findings. It is possible that this aspect can be learned by studying and watching others perform. The ability to describe the procedure and findings has a very low discriminative ability, which means that several consultants would fail if the cut score was higher (fig. 3c).

Two of the experts estimated that a borderline trainee spends 30 min on a diagnostic bronchoscopy and this resulted in a mean score of almost 16 min (the acceptable amount of time spent on a bronchoscopy). Not even nov-

Fig. 3. Relationship between the proposed cut score (Angoff method) and prior bronchoscopy performance. a Technical abilities. b Anatomical knowledge. c Ability to describe procedure and findings.
ices spent that long on average and nobody failed the time criterion. However, for practical and financial reasons, and for the sake of patient safety and comfort, a physician should be required to perform an uncomplicated, diagnostic bronchoscopy relatively fast. Although rapid task completion is a recognized feature of expert performance, a good bronchoscopy is both fast and accurate [15, 27]. An acceptable time would probably be around 10 min as proposed by 5 of the 7 raters. One way to create better consensus among the experts in a standard setting procedure is to include an iterative procedure in which outlier ratings (i.e. time = 30 min) are discussed and can be revised. We chose not to do this as few judges change their ratings once the initial mental effort has been expended [28]. However, this example of unrealistic expectations of some experts indicates the need for standard setting procedures.

Our proposed cut scores did not allow any complete novices to pass the test, did not fail any bronchoscopy experts, and passed an appropriate number of the most experienced trainees. Performance of 75 and 80 bronchoscopies, respectively, was related to performance above the pass level. This indicates that recommendations by various jurisdictions regarding experience with 100 bronchoscopies seems meaningful and safe regarding basic competence. The similarity in the results obtained by two independent assessment instruments and standard setting procedures supports the validity of our findings. Moreover, the similarity of results from the simulation assessment data and the clinical performance data indicates that simulation performance may be a good indicator of clinical performance. However, further studies are needed to confirm this.

Our study has several limitations. A test based on performance on a simulator is never totally realistic, and the clinical assessment tool does not measure more advanced procedures such as performing biopsies. Moreover, both assessment tools do not include other important traits of a good bronchoscopist, i.e. the ability to select patients or interpret history, clinical exam, and imaging results. Combining these issues with the technical bronchoscopy performance is an area in need of further research and development.

Another limitation for the evaluation of both assessment instruments is the limited number of tested physicians. If the tests were used on a wider scale, there would be a risk of false positives and negatives. In a recent study medical students were given written material about the bronchoscopy procedure and a poster of the anatomy of the bronchial tree, which they were strongly encouraged to study before a test. One of the 10 students was particularly skilled, and managed to pass the test as a complete novice. Ten other students performed between 7 and 19 bronchoscopies on a different virtual-reality simulator just before taking the test – 8 of these passed even though they had never performed an actual bronchoscopy! This indicates that even though bronchoscopy simulators are very realistic [29], it would be unwise to base the important decision about when a physician has reached basic bronchoscopy competency solely on a simulator test.

In the future more physicians should be assessed using the tool, especially more supposedly borderline trainees who have performed between 10 and 100 bronchoscopies. It is also necessary to use the tests with other raters across the world to ensure generalizability, as both tests were validated using a limited number of raters from Denmark.

Until these issues are further investigated it seems reasonable to require that fellows perform a minimum of 100 procedures. Our study indicates that the performance of 80 bronchoscopies ensures basic competency, but because it is very important to protect the public by identifying incompetent physicians, one could argue for a higher number that would minimize pass errors.

Conclusions

Our proposed pass/fail scores seem appropriate in terms of consequences. The performance of 75 and 80 bronchoscopies, respectively, correlated well with clear pass criteria. In the future objective assessment tools could become an important aid in the certification of physicians performing bronchoscopies.

Acknowledgments

We thank Dr. Henrik Kirstein Jensen, Dr. Torben Riis Rasmussen, Dr. Asbjørn Hoegholm, Dr. Niels Christian Hansen, and Dr. Martin Iversen for their kind participation in the expert group.

Financial Disclosure and Conflicts of Interest

Dr. Konge, Dr. Clementsen, Dr. Larsen, Dr. Arendrup, Prof. Buchwald, and Prof. Ringsted have no conflicts of interest to declare or financial disclosure to make.
References

14. ACGME Program Requirements for Fellowship Education in Pulmonary Disease and Critical Care Medicine (Internal Medicine). IM – Pulmonary Critical Care, 2011.