Validity and reliability of grade scoring in the diagnosis of exercise-induced laryngeal obstruction

Walsted, Emil Schwarz; Hull, James H; Hvedstrup, Jeppe; Maat, Robert Christiaan; Backer, Vibeke

Published in:
ERJ Open Research

DOI:
10.1183/23120541.00070-2017

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

ORIGINAL ARTICLE
EILO

Validity and reliability of grade scoring in the diagnosis of exercise-induced laryngeal obstruction

Emil Schwarz Walsted [1,2], James H. Hull[2], Jeppe Hvedstrup [1], Robert Christiaan Maat[3] and Vibeke Backer[1]

Affiliations: [1]Respiratory Research Unit, Dept of Respiratory Medicine, Bispebjerg University Hospital, Copenhagen, Denmark. [2]Dept of Respiratory Medicine, Royal Brompton Hospital, London, UK. [3]Dept of Otorhinolaryngology, Röpcke-Zweers Hospital, Hardenberg, The Netherlands.

Correspondence: Emil Schwarz Walsted, Respiratory Research Unit, Dept of Respiratory Medicine, Bispebjerg Hospital, Bispebjerg Bakke 23, DK-2400 Copenhagen NV, Denmark. E-mail: emilwalsted@dadlnet.dk

ABSTRACT The current gold-standard method for diagnosing exercise-induced laryngeal obstruction (EILO) is continuous laryngoscopy during exercise (CLE), with severity classified by a visual grade scoring system. We evaluated the precision of this approach, by evaluating test–retest reliability of CLE and both inter- and intra-rater variability.

In this prospective case–control study, subjects completed four consecutive treadmill CLE tests under identical conditions. Laryngoscopic video recordings were anonymised and graded by three expert raters. 2 months following initial scoring, videos were re-randomised and rating repeated to assess intra-rater agreement.

20 subjects (16 cases and four controls) completed four CLE tests. The time to exhaustion increased by 30 s (95% CI 0.02–57.8, $p<0.05$) in the second CLE compared with the first test, but remained identical in the subsequent tests. Only one-third of subjects retained their initial diagnosis in the subsequent three tests. Inter-rater agreement on grade scores (weighted Cohen's $\varkappa$) was 0.16–0.45, while intra-rater agreement ranged from 0.30 to 0.67.

The CLE test is key in the diagnostic assessment of patients with EILO. However, the widely adopted visual grade scoring system does not appear to be a robust means for reliably classifying severity of EILO.

@ERSpublications
Validity and reliability of grade scoring in exercise-induced laryngeal obstruction (EILO) http://ow.ly/cDWV30cClFX

Cite this article as: Walsted ES, Hull JH, Hvedstrup J, *et al*. Validity and reliability of grade scoring in the diagnosis of exercise-induced laryngeal obstruction. *ERJ Open Res* 2017; 3: 00070-2017 [https://doi.org/10.1183/23120541.00070-2017].

Received: June 08 2017 | Accepted after revision: June 12 2017

Support statement: This study did not receive any funding. Equipment was kindly provided by Bispebjerg Hospital (Copenhagen, Denmark).

Conflict of interest: None declared.

Introduction

Exercise-induced laryngeal obstruction (EILO) describes a clinical state in which the laryngeal inlet narrows during intense exercise, to precipitate respiratory symptoms such as wheeze and breathlessness [1]. It is now recognised to be a prevalent cause of exertional dyspnoea in young individuals [2, 3] and is the key differential diagnosis for exercise-induced bronchoconstriction (EIB) [4, 5].

In order to confirm or refute a diagnosis of EILO it is necessary to directly visualise laryngeal movement during intense exercise [6]. This is typically achieved by placing and then securing a flexible laryngoscope in the nasopharynx; permitting an uninterrupted assessment of laryngeal function during exercise and thus characterising any propensity to closure. This gold-standard diagnostic approach, termed the continuous laryngoscopy during exercise (CLE) test, was first described >10 years ago [7] and has now been used as a diagnostic test in many thousands of patients, employing a variety of exercise modalities [8–10]. It is possible to diagnose EILO using laryngoscopy immediately post-exercise in patients who are is still symptomatic; however, this method is likely to be less sensitive. Thus, the CLE test allows diagnostic assessment of EILO throughout an exercise bout, but also provides important detail regarding the anatomical nature of any laryngeal closure (*i.e.* occurring at either the glottic or supraglottic level or both) and therefore informs treatment decisions (*i.e.* determines whether surgical intervention is appropriate) [11–14].

A number of techniques have been employed to "score" or characterise the degree of laryngeal closure during a CLE test. In the clinical setting, the most widely accepted and utilised classification system to determine severity is the visual grade score [11]. This scoring system grades laryngeal closure at both the glottic and supraglottic level, with a score between 0 (complete patency) and 3 (almost complete closure) (figure 1). Typically, a score of $\geqslant 2$ at either the glottic or supraglottic level is taken as "diagnostic" of abnormal or exaggerated laryngeal closure [2, 3, 6]. Indeed, this cut-off has now been employed in a significant number of studies and informs treatment algorithms [3, 11]. Yet, despite the widespread adoption of this approach and some published evidence reporting inter-rater variation in grade scoring [5, 11, 15, 16], there are no robust data establishing the validity, precision and reliability of this approach to scoring EILO. Moreover, little is currently known about the test–retest reliability of CLE and whether a "learning effect" exists when multiple CLE tests are performed in the same individual.

We therefore undertook this study with the aim of evaluating the test–retest precision and validity of the CLE test, in a cohort of individuals diagnosed with EILO. In addition, we performed a comprehensive evaluation of inter- and intra-rater scoring validity of both the grade EILO scoring system and time to onset of laryngeal closure, using a video assessment with three experienced clinicians evaluating CLE tests. We hypothesised that there would be moderate to good agreement between these clinicians and the grade scoring system would thus be a robust and reliable scoring system in the diagnostic assessment of EILO.

Methods

Study design and subjects

In this prospective single-blinded case–control study, we recruited 23 subjects (17 with EILO and six healthy controls) from the outpatient asthma clinic at Bispebjerg University Hospital (Copenhagen, Denmark). Patients were recruited from a cohort initially referred for investigation of unexplained exertional dyspnoea, were free from respiratory disease and were nonsmokers. One patient and one control were competitive athletes; subjects were otherwise physically active but not highly trained.

All subjects underwent a diagnostic work-up that included spirometry, bronchial provocation testing and CLE testing. All subjects were then requested to re-attend over a period of 3 weeks for three further CLE tests under conditions identical to the initial CLE. Subjects were instructed to maintain their usual lifestyle, diet and exercise level between visits, and did not undergo any therapy for EILO (*i.e.* speech and language therapy or supraglottoplasty) for the duration of the study. The study was approved by the regional ethical committee for the Capital region of Denmark (ID H-3-2010-142).

CLE testing

The CLE test was performed as described previously [5]. In brief, subjects performed an incremental exercise test on a treadmill with a nasendoscope (Olympus, Tokyo, Japan) *in situ*. The exercise protocol started at a self-determined pace with no incline. After a 2-min lead-in the incline was increased by 3% every 90 s while the participant exercised to a self-determined level of exhaustion while a laryngoscopic video was continuously recorded. Time to exhaustion was taken as termination in the exercise bout. Subjects performed all four tests using the same equipment in the same ambient conditions, at the same pace and incline protocol and utilising the same naris for endoscopy placement.

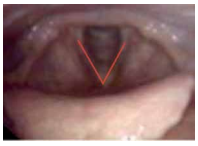
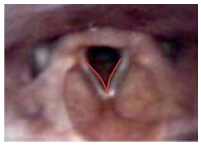
	Glottic Grading of parameters A and C:	Supraglottic Grading of parameters B and D:
Evaluation of the laryngoscopy video recording:#	Expected maximal abduction of the vocal cords (normal) 0 	Expected maximal abduction of the aryepiglottic folds with no visible medial rotation (tops of cuneiform tubercles pointed vertical or slightly lateral) 0 
Glottic / Supraglottic A | B C | D	Narrowing or adduction anteriorly of rima glottidis without visible motion of the arytenoid cartilage synchronised to inhalation 1 	Visible medial rotation of the cranial edge of the aryepiglottic folds and tops of the cuneiform tubercles (synchronous to inhalation) 1 
Sum score: E=A+B+C+D Clustered sum score: I: E = 0,1,2 II: E = 3,4 III: E ⩾5	Inhalation synchronised adduction of vocal cords but no contact between cords 2 	Further medial rotation of the cuneiform tubercles with exposure of the mucosa on the lateral side of the tubercles (synchronous to inhalation) 2 
	Total closure of the glottic space synchronous to inhalation 3 	Medial rotation until near horizontal position of the cuneiform tubercles and tops of the cuneiform tubercles moves towards the midline (synchronous to inhalation) 3 
Moderate effort scores	A 0 1 2 3	B 0 1 2 3
Maximal effort scores	C 0 1 2 3	D 0 1 2 3

FIGURE 1 The grading system illustrated by photographic images of the larynx showing the different grades of laryngeal obstruction at the glottic and supraglottic levels. #: the scores at each level (glottic (A and C) and supraglottic (B and D)) were assessed at moderate (A, B) (when the subject started to run) and at maximal (C, D) (just before the subject stopped running on the treadmill) effort; all four levels (A–D) were noted together with a sum score (E) for each test/subject. Reproduced from [11] with permission from the publisher.

Test–retest reliability and inter-rater agreement

The laryngoscopic video recordings from CLE were anonymised and assigned a computer-generated random identifier and order. Subsequently, three raters, all blinded to the other physiological characteristics and all with clinical experience in the diagnosis and treatment of EILO (all had assessed >300 CLE tests), independently graded the anonymised videos using the visual grade scale described by Maat *et al.* [11]. Specifically, each video was given a supraglottic and a glottic obstruction grade score (none, mild, moderate

or severe) and the time to onset of EILO (grade ⩾2 onset). A CLE test was considered positive (*i.e.* showing clinically significant EILO) if either the glottic or the supraglottic score was ⩾2.

Intra-rater agreement

The clinical raters were then asked to rescore all CLE videos (re-randomised) at an interval >2 months following the initial scoring session. Raters were required to rescore as described earlier, to evaluate internal consistency. The raters did not have access to the original set of videos from the first scoring session. Additionally, the videos were re-encoded, changing the randomisation screen overlay, and thereby the video file size.

Statistical analysis

Inter-rater and intra-rater agreement as well as test-to-test agreement in glottic and supraglottic grade scores were described using Cohen's weighted $\varkappa$ [17] and the level of agreement was classified as described by Landis and Koch [18]. Similarly, agreement on test duration and time to onset was described using intraclass correlation coefficients (ICC) (2,1) [19] and the classification used by Cicchetti [20]. To determine changes and differences in test duration from test to test and between the groups, we used a linear mixed model with random intercept and random slope. Statistical analysis was performed using SAS 9.4 (SAS Institute, Cary, NC, USA) and SPSS 24.0 (SPSS, Chicago, IL, USA).

Results

Out of the 23 included subjects (19 female and four male), 20 (87%) completed all four CLE tests, while three subjects underwent only one test (*i.e.* declined or were unable to attend subsequent tests). The tests were performed a median (interquartile range) 5 (4) days apart. The majority of subjects were female (83%), and all had normal resting spirometric indices (table 1). At initial testing, 17 subjects had a diagnosis of moderate (n=14) to severe (n=3) EILO.

Test–retest reliability

In comparison with the initial CLE (*i.e.* test 1), only one-third of subjects retained their initial diagnosis over the course of the subsequent three tests (figure 2), *i.e.* the diagnostic severity of EILO (crossing the grade severity threshold of 2), was re-classified on subsequent testing in two-thirds of patients and half of controls.

The time to exhaustion was ~30 s longer (95% CI 0.02–57.8, $p<0.05$) in the second CLE compared with the first test, but there was no significant difference in duration of test between the subsequent tests (*i.e.* tests 2–4) (figure 3). Even when any change in EILO severity score from test 1 to 2 was excluded, it remained that only half of subjects retained their initial diagnosis in all three tests (figure 2).

In 54% of cases, subjects classified as negative/normal (*i.e.* grade <2) by a rater at the initial CLE were subsequently classified as positive by the same rater in any one of the remaining three tests.

Overall, for both glottic and supraglottic grade scores, test–retest agreement in tests 1–4 was fair to moderate; Cohen's weighted $\varkappa$ ranged from 0.31 to 0.58 (table 2). The ICC describing test–retest agreement on time to onset of EILO was good (0.61) for tests 1 and 2 and excellent (0.79–0.86) for tests 2, 3 and 4 (table 2).

TABLE 1 Subject characteristics

Sex n (%)	
Female	19 (83)
Male	4 (17)
Age years median (range)	23 (15–45)
Body mass index kg·m^{-2}	21.1 (4.6)
FEV_1 L	3.50 (1.2)
FEV_1 % pred	98.7 (25)
FVC L	4.30 (1.2)
FVC % pred	103.3 (26.1)

Data are presented as median (interquartile range), unless otherwise stated. FEV_1: forced expiratory volume in 1 s; FVC: forced vital capacity.

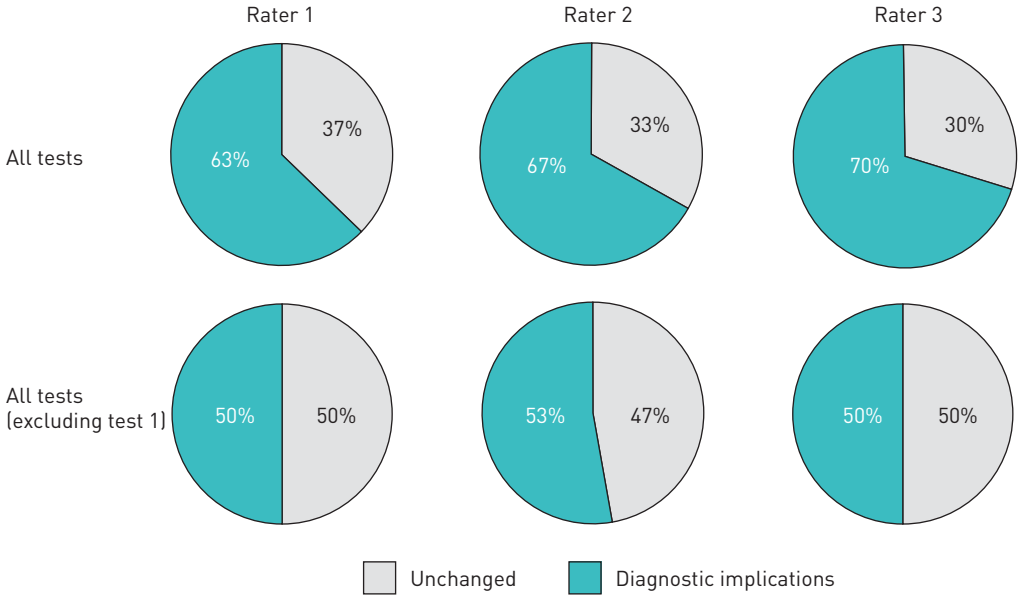


FIGURE 2 Diagnostic implications: one or more changes in exercise-induced laryngeal obstruction diagnosis (*i.e.* having or not having clinically significant laryngeal obstruction at either the glottic or the supraglottic level) over the course of the four tests. In a secondary analysis, the first test was excluded from analysis to depict implications due to intra-rater and test–retest variation only (*i.e.* excluding any learning effect; see also figure 3).

Inter-rater and intra-rater agreement

Overall, the agreement between the three raters ranged from only slight to moderate ($\varkappa$ 0.16–0.45) for glottic grade scores and from fair to moderate ($\varkappa$ 0.30–0.42) for supraglottic grade scores, while the agreement on time to onset of EILO was fair (ICC 0.54–0.56) (table 3).

Intra-rater agreement was marginally better, ranging from fair to substantial ($\varkappa$ 0.30–0.67) for glottic grade scores, while agreement on supraglottic grade scores was moderate ($\varkappa$ 0.41–0.56) and agreement on time to onset was fair (ICC 0.48–0.58) (table 4).

Discussion

Direct laryngoscopy, performed continuously during exercise, with the visual grade scoring system, is currently the gold standard and most widely utilised diagnostic methodology in the assessment of EILO [21, 22]. In this study we provide the first evaluation of the validity of this approach and in fact found poor levels of agreement between expert raters using this approach, to semi-objectively quantify severity of EILO. Moreover, the differences observed were not trivial, with approximately two-thirds of subjects having a potentially clinically relevant change in their diagnosis, over the course of four sequential CLE tests.

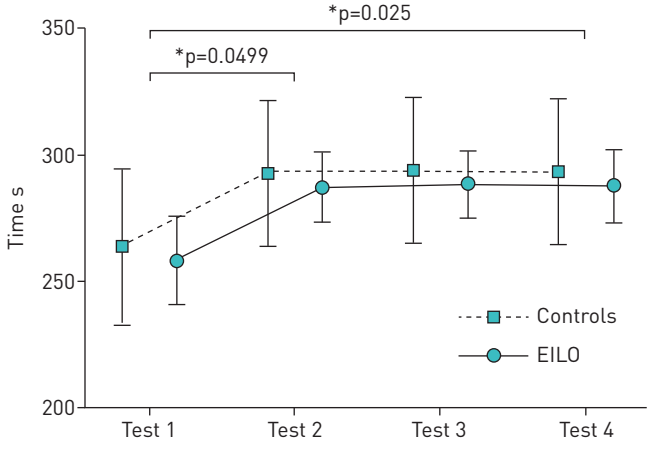


FIGURE 3 Learning effect on test duration. Data are presented as predicted means±SE from a linear mixed model of group and time as predictors of test duration. EILO: exercise-induced laryngeal obstruction. *: p<0.05.

TABLE 2 Test–retest reliability of exercise-induced laryngeal obstruction grade scores and time to onset

	Tests 1 and 2		Tests 2 and 3		Tests 3 and 4	
	Cohen's κ (95% CI)	ICC (95% CI)	Cohen's κ (95% CI)	ICC (95% CI)	Cohen's κ (95% CI)	ICC (95% CI)
Glottic scores	0.49 (0.32–0.67)		0.52 (0.35–0.70)		0.40 (0.23–0.57)	
Supraglottic scores	0.31 (0.07–0.55)		0.45 (0.25–0.65)		0.58 (0.41–0.76)	
Time to onset		0.61 (0.37–0.77)		0.86 (0.73–0.92)		0.79 (0.61–0.89)

ICC: intraclass correlation coefficient (2,1), single measures.

EILO is an increasingly well-recognised cause of exertional dyspnoea in the young and has been estimated to affect between 5% and 10% of the adolescent population [2, 3]. Prompt and accurate diagnosis is essential to facilitate delivery of targeted treatment, which can include surgical intervention in select and well-characterised cases [13, 14, 23].

The widely adopted scoring system for classification of EILO, namely the visual grade score [11], is based on movements of identifiable anatomical structures seen on laryngoscopic video recorded during CLE testing, and was the first systematic approach to provide clinicians with a "vocabulary" to describe the phenomena observed in CLE tests. Although originally intended as a clinical tool for use as a part of a complete assessment, detection of a grade 2 (figure 1) or greater obstruction pattern is now often taken as "diagnostic" of clinical EILO guiding clinical decisions and the score is used as a primary outcome measure in clinical research. This was not the intention of the original scoring system and indeed, the findings from this study challenge the validity of this approach and indicate that that the appearance of laryngeal closure, dictated by this relatively crude and binary threshold approach, appears to be neither robust nor dependable. Specifically, when three experienced raters were asked to independently score CLE video appearance, there was only fair to moderate agreement between scores. A similar finding was apparent when raters were asked to rescore the same videos at a 2-month interval, thus supporting the notion that the differences observed in scores relates to difficulties in providing a consistent appraisal of the video images (*i.e.* high subjective bias), rather than disagreement between three experts, trained and working in different centres and specialisations. In the present study, two of the expert raters were specialists in respiratory medicine and one in otorhinolaryngology, possibly augmenting disagreement between raters compared with previous work describing moderate to good inter-rater agreement between otorhinolaryngologists.

A number of prior studies have reported inter-rater variability of scoring systems from CLE [5, 11, 15, 16]. Depending on scoring methodology (*i.e.* blinding and randomisation), these studies report moderate to perfect agreement on visual grade scores. Most studies use Cohen's weighted κ to describe agreement (as the grade score is ordinal), yet, some studies [15, 16] report percent-agreement; a poor measure of agreement which should be interpreted with care [24]. This acknowledged, in contrast to previous work, the videos analysed in the present study were not preselected for scoring based on any video quality criteria, thus explaining the relatively lower level of inter- and intra-rater agreement.

Importantly, although there was only moderate overall agreement between raters, there appeared to be better agreement for scores for those with supraglottic-pattern EILO, *i.e.* laryngeal closure occurring at the level of the supraglottic/arytenoid structures. Indeed, while two cases were judged from 1 (mild) to 3 (severe) between raters for glottic-level closure EILO, the same was not true of supraglottic closure, with only one case of alteration in severity score. This is important in the context of impact on therapeutic

TABLE 3 Inter-rater agreement on exercise-induced laryngeal obstruction grade scores and time to onset

	Raters 1 and 2		Raters 2 and 3		Raters 3 and 1	
	Cohen's κ (95% CI)	ICC (95% CI)	Cohen's κ (95% CI)	ICC (95% CI)	Cohen's κ (95% CI)	ICC (95% CI)
Glottic scores	0.16 (0.03–0.29)		0.45 (0.30–0.59)		0.16 (0.08–0.25)	
Supraglottic scores	0.42 (0.27–0.57)		0.40 (0.20–0.60)		0.30 (0.15–0.44)	
Time to onset		0.54 (0.29–0.72)		0.56 (0.00–0.80)		0.55 (0.36–0.70)

ICC: intraclass correlation coefficient (2,1), single measures.

TABLE 4 Intra-rater agreement on exercise-induced laryngeal obstruction grade scores and time to onset

	Glottic scores $\varkappa$ (95% CI)	Supraglottic scores $\varkappa$ (95% CI)	Time to onset ICC (95% CI)
Rater 1	0.30 (0.10–0.49)	0.41 (0.27–0.55)	0.45 (0.23–0.62)
Rater 2	0.48 (0.31–0.64)	0.56 (0.40–0.71)	0.85 (0.74–0.92)
Rater 3	0.67 (0.56–0.79)	0.43 (0.23–0.63)	0.43 (0.22–0.60)
All raters	0.58 (0.49–0.66)	0.48 (0.39–0.57)	0.51 (0.39–0.61)

ICC: intraclass correlation coefficient (2,1), single measures.

strategy. Specifically, glottic closure abnormalities are typically treated with targeted breathing control work, regardless of severity, whereas, for more severe supraglottic abnormalities, supraglottaplasty may be a therapeutic option [13, 14, 23]. However, the procedure is still experimental; the total number of surgically treated EILO patients worldwide to date is small (<100 cases published) and the long-term prognosis for the patients is yet unknown. Thus, for most patients, appropriate treatment options will be speech and language therapy, respiratory physiotherapy and optimisation of asthma medications, if relevant.

Differentiating natural variability in the test (*i.e.* true subject variability) *versus* inter- and intra-rater effect is difficult in the absence of a truly objective standard; test–retest reliability will be affected by both intra-rater variation and true variation in the test results. When comparing intra-rater agreement on glottic and supraglottic scores (table 4) with test–retest agreement (table 2), it is evident that the test-retest variation is probably caused by intra-rater variation, although no definitive conclusion can be made. Another caveat when assessing test–retest variability is that, since intra-rater agreement is not perfect, the probability of changing scores over a given number of tests will rise with the number of tests performed.

The current study reveals the presence and potential for a slight "learning effect" between CLE tests. Specifically, time to exhaustion in the test increased by ~30 s (11%) from visit 1 to 2 (figure 3). This is a phenomenon observed in most exercise tests, and for instance has dictated practice in field-based exercise testing (*e.g.* need for a practice 6-min walk test in chronic obstructive lung disease [25]). It is a potentially important consideration when planning intervention and assessment studies in EILO studies, *i.e.* planning a test or familiarisation run may be necessary. Indeed, Olin *et al.* [16] recently described how the visual grade score is effort dependent and most evident at high intensity.

Our findings of suboptimal test–retest variability and the presence of a learning effect in CLE testing are consistent with diagnostic testing employed in the diagnosis of EIB; with poor test–retest reliability dictating an argument that diagnostic EIB tests may be need to be repeated in order to truly establish or refute a borderline diagnosis [26, 27]. Having said this, mean group score for both glottic and supraglottic EILO appeared to change little over the repeated tests, *i.e.* although a familiarisation test may be required to allow the subject to exercise longer and thus to avoid underdetection of EILO (*i.e.* not exercising long enough to a sufficiently high intensity), this finding appeared to be of little relevance compared with the other larger disparities within the rating system. Moreover, in the current study we do not have data available for either ventilation or heart rate, and thus, providing that a subject has exercised to a high intensity (*i.e.* >90% peak predicted heart rate), this may mitigate this confounding issue.

Implications of findings

Any clinical diagnostic decision process is clearly based on a synthesis between clinical features and diagnostic testing. The same is true in EILO and diagnosis is not based simply on a "blinded" video assessment, *i.e.* it is important to determine if the subject developed stridor and their typical exertional symptoms during the test. It is thus an oversimplification to assume that the diagnosis would have changed in cases assessed in this scenario. However, as CLE testing keeps gaining ground in the investigation of exertional dyspnoea, it is important that clinicians and researchers pay attention to potential learning effects and inter-/intra-rater variability, particularly when evaluating effects of treatment interventions, as they could bias the assessment of treatment outcomes.

Likewise, it is apparent that some individuals may appear to develop significant degrees of laryngeal closure (*i.e.* grade $\geqslant$2) and yet remain asymptomatic during exercise, and *vice versa*, some may develop only a minor encroachment, yet develop troublesome symptoms that lead to exercise cessation. The majority of subjects were symptomatic (17 out of 20), yet the reproducibility of grade scores in individuals

with no symptoms despite grade scores of >0 is unknown. Therefore, there is an urgent need to understand the physiological correlates of glottic closure in exercise and, for instance, how this relates to an individual's dyspnoea, but also to respiratory loading. Indeed, it may be more important to record symptoms and measure other physiological correlates (*e.g.* flow, pressure and neural drive) than to simply classify and characterise EILO based on the appearance of the larynx.

Other systems/approaches have been utilised in measuring and classifying the severity of laryngeal closure. One system utilised a computer-generated scoring system based on mapping the glottic structures with a software system, EILOMEA [28]. Although reporting sensitivity and specificity of 1.00 for supraglottic EILO, a recent study by Norlander *et al.* [22] found the method to be comparable to grade scoring, but impractical for clinical application. The software has yet to be validated in repeated measures and relies on a single snapshot still image for the measurements to then be applied. This fails to recognise the highly variable and evolving nature of EILO and a mapping algorithm is needed to capture the dynamic status; in addition, considering the added subjectivity involved in marker placement and selecting a still frame representative of a whole video, it does not qualify as an objective measure.

Less invasive means of detecting EILO would be preferable to CLE testing; unfortunately, neither spirometry [29], bronchial provocation testing [6, 30] or impulse oscillometry [31, 32] have yet proven useful as diagnostic tools in the clinical investigation of EILO; airway perturbation might prove a valuable diagnostic method [33, 34], but needs to be validated in prospective, large-scale clinical studies. Common to all surrogate tests aiming to detect EILO, no truly objective reference exists. Although a fully objective scoring method would be preferable, visualising the larynx when the patient is symptomatic is crucial and CLE remains the current gold standard for diagnosing EILO. Endoscopy technology is evolving rapidly and portable laryngoscopy kits now allow for more specific field-testing [8], which may in time prove even more sensitive than laboratory-based CLE.

Methodological considerations

In the present study, we employed a blinded and randomised setup including multiple tests and a standardised exercise protocol to investigate the precision and test–retest reliability of the CLE test. We acknowledge that the tests have been performed in a selected population; particularly, the majority of subjects had EILO of moderate severity. While this addresses the clinical problem of diagnosing EILO well, intra- and inter-agreement might be higher in a population with a higher proportion of severe/less ambiguous cases. Additionally, ventilatory data or other clinical information could have assisted the raters in making a more precise scoring although it would have masked the variability of the grade scoring system itself.

In the present study, we applied the protocol used for CLE testing in our outpatient clinic, including for re-evaluation of EILO. No established standard exercise protocol exists for CLE testing and any learning effect will depend on the specific protocol employed; indeed, individual adjustments to a standard exercise protocol might be necessary to reproduce a patient's exercise-related respiratory symptoms in a laboratory setting.

Conclusion

In conclusion, the CLE test plays a key and vital role in the diagnostic assessment of patients with EILO. However, the data from this work suggest that the use of the grade scoring system, first established by Maat *et al.* [11], and now widely used, is not a robust means for reliably classifying the severity of EILO. Alternative, fully objective methods are urgently needed to ensure that the severity of EILO is characterised consistently both in clinical practice and in studies reliant on robust outcome measurements.

References

1 Hall A, Thomas M, Sandhu G, *et al.* Exercise-induced laryngeal obstruction: a common and overlooked cause of exertional breathlessness. *Br J Gen Pract* 2016; 66: e683–e685.
2 Johansson H, Norlander K, Berglund L, *et al.* Prevalence of exercise-induced bronchoconstriction and exercise-induced laryngeal obstruction in a general adolescent population. *Thorax* 2015; 70: 57–63.
3 Christensen PM, Thomsen SF, Rasmussen N, *et al.* Exercise-induced laryngeal obstructions: prevalence and symptoms in the general public. *Eur Arch Otorhinolaryngol* 2011; 268: 1313–1319.
4 Hull JH, Ansley L, Robson-Ansley P, *et al.* Managing respiratory problems in athletes. *Clin Med* 2012; 12: 351–356.
5 Nielsen EW, Hull JH, Backer V. High prevalence of exercise-induced laryngeal obstruction in athletes. *Med Sci Sports Exerc* 2013; 45: 2030–2035.
6 Walsted ES, Hull JH, Sverrild A, *et al.* Bronchial provocation testing does not detect exercise-induced laryngeal obstruction. *J Asthma* 2017; 54: 77–83.
7 Heimdal J-H, Roksund OD, Halvorsen T, *et al.* Continuous laryngoscopy exercise test: a method for visualizing laryngeal dysfunction during exercise. *Laryngoscope* 2006; 116: 52–57.

8 Walsted ES, Swanton LL, van van Someren K, *et al.* Laryngoscopy during swimming: a novel diagnostic technique to characterize swimming-induced laryngeal obstruction. *Laryngoscope* 2017; in press DOI: 10.1002/lary.26532.
9 Panchasara B, Nelson C, Niven R, *et al.* Lesson of the month: rowing-induced laryngeal obstruction: a novel cause of exertional dyspnoea: characterised by direct laryngoscopy. *Thorax* 2015; 70: 95–97.
10 Tervonen H, Niskanen MM, Sovijärvi AR, *et al.* Fiberoptic videolaryngoscopy during bicycle ergometry: a diagnostic tool for exercise-induced vocal cord dysfunction. *Laryngoscope* 2009; 119: 1776–1780.
11 Maat RC, Røksund OD, Halvorsen T, *et al.* Audiovisual assessment of exercise-induced laryngeal obstruction: reliability and validity of observations. *Eur Arch Otorhinolaryngol* 2009; 266: 1929–1936.
12 Maat RC, Røksund OD, Olofsson J, *et al.* Surgical treatment of exercise-induced laryngeal dysfunction. *Eur Arch Otorhinolaryngol* 2007; 264: 401–407.
13 Mehlum CS, Walsted ES, Godballe C, *et al.* Supraglottoplasty as treatment of exercise induced laryngeal obstruction (EILO). *Eur Arch Otorhinolaryngol* 2016; 273: 945–951.
14 Norlander K, Johansson H, Jansson C, *et al.* Surgical treatment is effective in severe cases of exercise-induced laryngeal obstruction: a follow-up study. *Acta Otolaryngol* 2015; 135: 1152–1159.
15 Olin JT, Clary MS, Connors D, *et al.* Glottic configuration in patients with exercise-induced stridor: a new paradigm. *Laryngoscope* 2014; 124: 2568–2573.
16 Olin JT, Clary MS, Fan EM, *et al.* Continuous laryngoscopy quantitates laryngeal behaviour in exercise and recovery. *Eur Respir J* 2016; 48: 1192–1200.
17 Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968; 70: 213–220.
18 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–174.
19 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979; 86: 420–428.
20 Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess* 1994; 6: 284–290.
21 Røksund OD, Heimdal J-H, Clemm H, *et al.* Exercise inducible laryngeal obstruction: diagnostics and management. *Paediatr Respir Rev* 2017; 21: 86–94.
22 Norlander K, Christensen PM, Maat RC, *et al.* Comparison between two assessment methods for exercise-induced laryngeal obstructions. *Eur Arch Otorhinolaryngol* 2016; 273: 425–430.
23 Maat RC, Hilland M, Røksund OD, *et al.* Exercise-induced laryngeal obstruction: natural history and effect of surgical treatment. *Eur Arch Otorhinolaryngol* 2011; 268: 1485–1492.
24 Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. *Commun Methods Meas* 2007; 1: 77–89.
25 ATS Committee on Proficiency Standards for Clinical Pulmonary Function Laboratories. ATS statement: guidelines for the six-minute walk test. *Am J Respir Crit Care Med* 2002; 166: 111–117.
26 Price OJ, Ansley L, Hull JH. Diagnosing exercise-induced bronchoconstriction with eucapnic voluntary hyperpnea: is one test enough? *J Allergy Clin Immunol Pract* 2015; 3: 243–249.
27 Anderson SD, Kippelen P. Assessment of EIB: what you need to know to optimize test results. *Immunol Allergy Clin North Am* 2013; 33: 363–380.
28 Christensen P, Thomsen SF, Rasmussen N, *et al.* Exercise-induced laryngeal obstructions objectively assessed using EILOMEA. *Eur Arch Otorhinolaryngol* 2010; 267: 401–407.
29 Christensen PM, Maltbæk N, Jørgensen IM, *et al.* Can flow-volume loops be used to diagnose exercise induced laryngeal obstructions? A comparison study examining the accuracy and inter-rater agreement of flow volume loops as a diagnostic tool. *Prim Care Respir J* 2013; 22: 306–311.
30 Christensen PM, Rasmussen N. Eucapnic voluntary hyperventilation in diagnosing exercise-induced laryngeal obstructions. *Eur Arch Otorhinolaryngol* 2013; 270: 3107–3113.
31 Price OJ, Ansley L, Bikov A, *et al.* The role of impulse oscillometry in detecting airway dysfunction in athletes. *J Asthma* 2016; 53: 62–68.
32 Bikov A, Pride NB, Goldman MD, *et al.* Glottal aperture and buccal airflow leaks critically affect forced oscillometry measurements. *Chest* 2015; 148: 731–738.
33 Gallena SJK, Tian W, Johnson AT, *et al.* Validity of a new respiratory resistance measurement device to detect glottal area change. *J Voice* 2013; 27: 299–304.
34 Gallena SK, Solomon NP, Johnson AT, *et al.* Test-retest reliability of respiratory resistance measured with the airflow perturbation device. *J Speech Lang Hear Res* 2014; 57: 1323–1329.