Equine Veterinary Journal

Equine Veterinary Journal ISSN 0425-1644
DOI: 10.1111/evj.13030

Evaluation of videoendoscopic examinations of arytenoid function in the 2-year-old Thoroughbred: Can we all agree?

J. McLELLAN and S. PLEVIN*

Florida Equine Veterinary Associates, Ocala, Florida, USA.

*Correspondence email: mpvets@hotmail.com; Received: 19.03.18; Accepted: 24.09.18

Summary

Background: Upper respiratory tract (URT) endoscopic examination is a routine part of prepurchase examinations. Discrepancies have been documented in the assessment of arytenoid function grades (AFG) between veterinarians.
Objectives: To document intra- and interobserver agreement for a population of multi-experience level veterinarians for assessment of AFG of 2-year-old Thoroughbreds.
Study design: Observational cohort study.
Methods: One-hundred and fourteen URT videoendoscopic examinations were evaluated by 10 veterinarians. Veterinarians were categorised based on experience, into five groups, each group with two veterinarians. Arytenoid function was graded using the Havemeyer ordinal scale and then reclassified by the authors dichotomously into 'meets conditions of sale' (MCS) or 'does not meet conditions of sale'(DNMCS). Interobserver agreement of arytenoid function was assessed across all 10 veterinarians using Fleiss' kappa and between veterinarians of similar experience levels in the five subgroups using Cohen's unweighted (k) and Cohen's linear weighted kappa (Ck). Intraobserver agreement was similarly calculated for each reviewer using 22 repeated video clips.
Results: Overall interobserver agreement using ordinal scales was fair (k = 0.27, 95% CI 0.22–0.31) to moderate (mean weighted Ck = 0.57, 95% CI 0.46–0.69) depending on statistical methodology used. Using the dichotomous classification, interobserver agreement was good (k = 0.7, 95% CI 0.63–0.77).
Overall intraobserver agreement using ordinal scales was fair (mean k = 0.26, 95% CI 0.14–0.38) to good (mean Ck = 0.61, 95% CI 0.50–0.71); and for the dichotomous reclassification it was good (k = 0.73, 95% CI 0.59–0.87). Experience level differences were identified.
Main limitations: The low number of veterinarians in each experience subgroup.
Conclusions: Subjectivity exists in arytenoid function grading, despite the existence of a well-defined scale. Agreement variation exists depending on the grading scales and statistical methods used for analysis. Future studies pre- and post veterinarian training are indicated to determine if agreement can be improved.

Keywords: horse; upper respiratory tract; endoscopy; Thoroughbred

Introduction

Endoscopic assessment of laryngeal function, specifically arytenoid abduction, is a routine part of prepurchase examinations at Thoroughbred public auctions throughout the world [1]. The results of such examinations provide a useful prognostic indicator of future racing performance [2,3]. Despite the emphasis placed on the results of these examinations, to the authors' knowledge, studies documenting the agreement between multiple veterinarians of varying experience for assessment of arytenoid function in 2-year-old Thoroughbreds are lacking. Previous studies assessing intra- and interobserver agreement for arytenoid function have varied in the breed and age of horses studied and have used a small number of predominantly experienced observers [4,5]. Interpretation of arytenoid function grade (AFG) assignment has been reportedly affected by these variables [4–9], resulting in discrepancies in agreement between veterinarians. Additionally, previous studies have used different arytenoid function grading scales and statistical methods for analysis, leading to confusion and inconsistencies when interpreting data, with inter- and intraobserver agreement varying from fair to excellent [4–7]. Anecdotal reports of disagreement between veterinarians at public auction regarding AFG assignment also exist. As prognostic recommendations and horse purchase prices are often affected by AFG, it is imperative that veterinarians both collectively and individually demonstrate a high level of agreement and repeatability when assigning such grades to 2-year-old Thoroughbreds.

The primary intent of the current study, therefore, was to determine, for the first time, the extent of intra- and interobserver agreement for AFG assignment in 2-year-old Thoroughbreds. A large number of veterinarians with different experience levels participated in the study. All participants viewed the same presale video endoscopic examinations. Discrepancies in grading could not, therefore, be attributed to extraneous factors such as equipment, endoscopic technique, holder, level of excitement, stress or fatigue of the horse. Based on results from previous studies and anecdotal reports of grading discrepancies at public auctions, it was hypothesised that there would be poor to moderate interobserver agreement for AFG. We hypothesised agreement would increase when using a more lenient dichotomous reclassification of 'meets conditions of sale' or 'does not meet conditions of sale' vs. the widely accepted seven-point ordinal Havemeyer scale [10,11]. 'Sales' veterinarians (veterinarians predominately working at public auctions) were predicted to have the best inter- and intraobserver agreement considering the frequency with which they assess arytenoid function.

Materials and methods

This was an observational cohort study. An a priori power analysis was conducted to determine an appropriate sample size with a power of 80% and alpha of 0.05. For assessment of agreement using the seven-point ordinal scale, assuming potential asymmetric proportions of grade in each category by each observer, with an expected minimum value for the Cohen's kappa coefficient of 0.5, the minimum anticipated number of video clips was 16. For assessment of agreement using the two-point dichotomous scale, it was anticipated that agreement would be higher. The anticipated value of kappa, therefore, was increased to 0.8, resulting in a required minimum of 20 video clips. To ensure sufficient video clips to permit post hoc exclusion, 22 examinations were used for analysis of intraobserver agreement.

The endoscopic examinations were collected from fifty-nine 2-year-old Thoroughbred horses located at a single public auction. These examinations were selected from a larger cross section of 'sale horses' to

ensure sufficient variation in AFG findings. Videoendoscopic examinations were compiled into a continuous 52 min and 40 s video sequence with 114 video clips documented by case number 1–114.

All exams were performed by the first author (J.M.), in accordance with the American Association of Equine Practitioners (AAEP) protocol for upper respiratory tract videoendoscopic examination at public auctions [12]. Horses were restrained using a lip chain. A 6.9-mm flexible endoscope[a] was passed through the right ventral meatus. Horses' nasal passages were occluded to allow achievement of maximal abduction of the arytenoid cartilages and the horse was induced to swallow by touching the back of the pharynx with the end of the endoscope. All examinations were performed as part of a routine presales examination and as such were not produced specifically for this study. Videos were, therefore, edited to achieve the aims of this investigation.

One video clip, between 30 and 45 s in length, of each of the 59 horses enrolled in the study was generated; these clips were used to assess interobserver agreement. All 59 examinations were performed in the mornings. As part of a future investigation, 11 of the horses also received follow-up afternoon examinations, which were also randomly added to the case series to increase case numbers.

Twenty-two randomly chosen video clips were exactly repeated and added to the series, to allow assessment of intraobserver repeatability. These duplicate videos and 11 afternoon examinations were, due to their random order in the case series, considered as independent observations for assessment of interobserver agreement as previously described [6]. Therefore, a total of 92 'full length' video clips were used to assess interobserver agreement. As part of an intended future investigation, 22 video clips were shortened to <15 s and randomly inserted into the case series. Results of interpretation of these abbreviated clips were excluded from this study. An Excel[b] spreadsheet was used to randomly generate case numbers for use in the final video series of 114 clips: the random order was intended to make it difficult for study participants to recollect specific cases.

Only video clips where anatomy was clearly visible and where video quality was sufficient to grade arytenoid function were used in this study. Video clips were removed from the study if such quality criteria were not met. Post hoc review of participants' comments related to each video clip were reviewed by primary and secondary authors (J.M. and S.P.).

Twelve equine veterinarians were invited to participate in the study. Inclusion criteria dictated that participants were equine-specific practitioners that fit into one of the five experience level groupings described below. Ten equine practitioners responded to the invitation and were recruited to analyse the endoscopic examinations. The observers consisted of: two interns; two junior associates (<5 years' experience each); two board certified equine surgeons; two 'sale specific' veterinarians; and two senior veterinarians (>15 years' experience each). These five experience groups, therefore, consisted of two veterinarians per group.

The video compilations were sent to the participants via Dropbox[c] with each endoscopic examination represented by a case number. No other identifying features existed for each clip. The veterinarians (observers) were provided with a grading form and an explanation of the Havemeyer arytenoid grading scale [10,11] to be used (Fig 1). Participants were asked to grade arytenoid function for each of the 114 cases using this scale. Additionally, participants were asked to assess pharyngitis and epiglottis grade as well as airway size, for use in future studies and there was an option to include comments for every case. No training was provided prior to the commencement of this study. Participants evaluated the video clips independently of each other and had the ability to replay video clips, to freeze frame them and to play them in slow motion. Participants were able to defer grading a case if they did not think the quality of videoendoscopic examination was diagnostic. Veterinarians were blinded to all information and were unaware that video clips had been randomly repeated throughout the sequence.

Grading system of laryngeal function in the standing un-sedated horse (Havemeyer workshop proceedings, 2003)

Grade	Description	Sub-grade	
I	All arytenoid cartilage movements are synchronous and symmetrical and full arytenoid cartilage abduction can be achieved and maintained		
II	Arytenoid cartilage movements are asynchronous and/or larynx asymmetric at times but full arytenoid cartilage abduction can be achieved and maintained	.1	Transient asynchrony, flutter or delayed movements are seen
		.2	There is asymmetry of the rima glottidis much of the time due to reduced mobility of the affected arytenoid and vocal fold but there are occasions, typically after swallowing or nasal occlusion when full symmetrical abduction is achieved and maintained
III	Arytenoid cartilage movements are asynchronous and/or asymmetric Full arytenoid cartilage abduction cannot be achieved and maintained	.1	There is asymmetry of the rima glottidis much of the time due to reduced mobility of the arytenoid and vocal fold but there are occasions, typically after swallowing or nasal occlusion when full symmetrical abduction is achieved but not maintained
		.2	Obvious arytenoid abductor deficit and arytenoid asymmetry. Full abduction is never achieved
		.3	Marked but not total arytenoid abductor deficit and asymmetry with little arytenoid movement. Full abduction is never achieved
IV	Complete immobility of the arytenoid cartilage and vocal fold		

Fig 1: The Havemeyer arytenoid function grading scale used by observers in this study.

For the purposes of statistical analysis, assigned Havemeyer grades were renumbered 1–7 based on the ordinal nature of the Havemeyer scale. Additionally, to allow assessment of dichotomous grading scales, authors reclassified ordinal grades dichotomously into 'meets conditions of sale' (MCS) or 'does not meet conditions of sale' (DNMCS). Adhering to the criteria set forth by the conditions of sale at public auction, any grade equal to or greater than grade III.1 on the Havemeyer scale was considered to not meet conditions of sale (DNMCS) [13].

Data analysis

All data analysis was conducted using an online software package[d].

Interobserver agreement was assessed for ordinal grades using Cohen's unweighted kappa (K) between observers within each of the five experience groups and using Fleiss' unweighted kappa across all 10 veterinarians, to report the level of perfect agreement between observers.

Interobserver agreement for ordinal arytenoid grading scales was also assessed between veterinarians within each of the five experience groups using Cohen's linear weighted kappa (Ck), with equal weighting between each discordant grade. As previously described [6] a mean linear weighted kappa value was calculated for interobserver agreement for ordinal grades across all 10 veterinarians: considered the mean level of agreement in the general population of clinicians. Finally, interobserver agreement was assessed (unweighted kappa) using the dichotomous reclassification of MCS or DNMCS, between observers of each experience group and using Fleiss' unweighted kappa, across all 10 observers. Percentages of agreement were also reported for both ordinal and dichotomous scales, and 95% confidence intervals (CIs) were recorded for each statistical test.

Using 22 exact repeat video pairs, intraobserver agreement was similarly calculated for each individual veterinarian using ordinal grades (Cohen's unweighted and linear weighted kappa) and the dichotomous reclassification (Cohen's unweighted kappa). As previously described [6], mean intraobserver agreement was calculated for each experience level and the population.

Descriptive agreement levels of kappa were as previously described: 0 = no agreement; 0.01 to 0.20 = poor agreement; 0.21–0.40 = fair agreement; 0.41–0.60 = moderate agreement; 0.61–0.80 = good agreement; >0.81 = excellent [14].

As previously reported [4], to determine which ordinal grades exhibited the most discordance in grading between observers, interobserver disagreement was assessed between each possible combination of veterinarians for each of the 92 video clips and reported as a mean percentage agreement for each AFG. To assess percentage agreement for each grade in any pair of veterinarians, percentage agreement for each AFG was calculated as the mean of the percentage agreement when veterinarian B graded the same as veterinarian A, and when veterinarian A graded the same as veterinarian B, as previously described [4].

Results

On post hoc review of observers' comments, one video clip was deemed nondiagnostic due to mucous accumulation on the lens and was thus excluded from further analysis. As this clip was a duplicate for use in the intraobserver study the number of paired full-length videos for analysis of intraobserver reliability was reduced to 21 and full-length baseline evaluations reduced to 90.

For all 10 veterinarians, ordinal agreement assessed by unweighted kappa was only fair (k = 0.27, 95% CI 0.22–0.31), with perfect agreement present for 37% of all ordinal grade assessments. Analysis of linear weighted kappa improved the mean interobserver agreement to moderate (Ck = 0.57, 95% CI 0.46–0.69) and agreement was good (k = 0.7, 95% CI 0.63–0.77) when grades were dichotomously reclassified. Interobserver variation existed between veterinarians with different experience with interns consistently demonstrating less agreement. Experience level variation also existed for intraobserver agreement, with members of the intern and surgeon groups demonstrating less intraobserver agreement than the other groups for both ordinal and dichotomous grades, as shown

TABLE 1: Interobserver and intraobserver % of agreement, unweighted kappa and linear weighted kappa values for both ordinal (Havemeyer) scale and dichotomous grading scales of arytenoid function

		Ordinal grades		Dichotomous grades	
	Agreement (%) (95% CI)	Unweighted Kappa (95% CI)	Linear weighted Kappa (95% CI)	Agreement (%) (95% CI)	Unweighted Kappa (95% CI)
Interobserver agreement (90 video clips)					
By experience level					
Intern	0.22 (0.14–0.32)	0.06 (0–0.16)	0.36 (0.25–0.48)	0.77 (0.66–0.85)	0.5 (0.32–0.68)
Inexperienced	0.51 (0.40–0.62)	0.34 (0.20–0.47)	0.67 (0.56–0.77)	0.92 (0.84–0.97)	0.79 (0.64–0.94)
Experienced	0.42 (0.32–0.53)	0.31 (0.18–0.43)	0.64 (0.54–0.73)	0.91 (0.83–0.96)	0.75 (0.62–0.89)
Sales vet	0.51 (0.40–0.62)	0.34 (0.21–0.47)	0.65 (0.54–0.76)	0.90 (0.81–0.95)	0.73 (0.57–0.89)
Surgeon	0.34 (0.25–0.45)	0.14 (0.01–0.27)	0.55 (0.45–0.66)	0.81 (0.71–0.88)	0.6 (0.46–0.74)
For all observers	0.37 (0.29–0.52)*	0.27 (0.22–0.31)*	0.57 (0.46–0.69)	0.85 (0.77–0.92)*	0.7 (0.63–0.77)*
Intraobserver repeatability (21 repeated video clips)					
Observer					
Intern A	0.36 (0.18–0.59)	0.22 (0–0.46)	0.38 (0.12–0.63)	0.72 (0.49–0.88)	0.45 (0.07–0.80)
Intern B	0.28 (0.12–0.52)	0.14 (0–0.38)	0.62 (0.42–0.81)	0.81 (0.57–0.94)	0.6 (0.31–0.89)
Mean intern	0.32	0.18	0.5	0.77	0.53
Inexperienced A	0.43 (0.23–0.66)	0.24 (0–0.51)	0.55 (0.28–0.83)	0.86 (0.63–0.96)	0.63 (0.32–0.94)
Inexperienced B	0.48 (0.26–0.70)	0.35 (0.09–0.61)	0.75 (0.62–0.57)	0.90 (0.68–0.98)	0.8 (0.59–1)
Mean inexperienced	0.46	0.30	0.65	0.88	0.715
Experienced A	0.5 (0.29–0.71)	0.39 (0.15–0.63)	0.73 (0.58–0.88)	0.95 (0.75–1)	0.9 (0.74–1)
Experienced B	0.43 (0.23–0.66)	0.29 (0.03–0.55)	0.65 (0.49–0.82)	1 (0.81–1)	1 (1–1)
Mean experienced	0.47	0.34	0.69	0.98	0.95
Sales A	0.38 (0.19–0.61)	0.26 (0.03–0.50)	0.74 (0.61–0.86)	0.90 (0.68–0.98)	0.81 (0.56–1)
Sales B	0.48 (0.26–0.70)	0.28 (0.01–0.56)	0.63 (0.42–0.84)	0.95 (0.74–0.99)	0.88 (0.68–1)
Mean sales	0.43	0.27	0.69	0.93	0.85
Surgeon A	0.38 (0.19–0.61)	0.24 (0.02–0.45)	0.65 (0.48–0.82)	0.90 (0.68–0.98)	0.81 (0.56–1)
Surgeon B	0.19 (0.06–0.43)	0.04 (0–0.24)	0.52 (0.34–0.70)	0.76 (0.52–0.91)	0.52 (0.22–0.91)
Mean surgeon	0.29	0.14	0.59	0.83	0.67
For all observers	0.41 (0.3–0.52)	0.26 (0.14–0.38)	0.61 (0.5–0.71)	0.81 (0.73–0.89)	0.73 (0.59–0.87)

*Values calculated by Fleiss' kappa calculations.

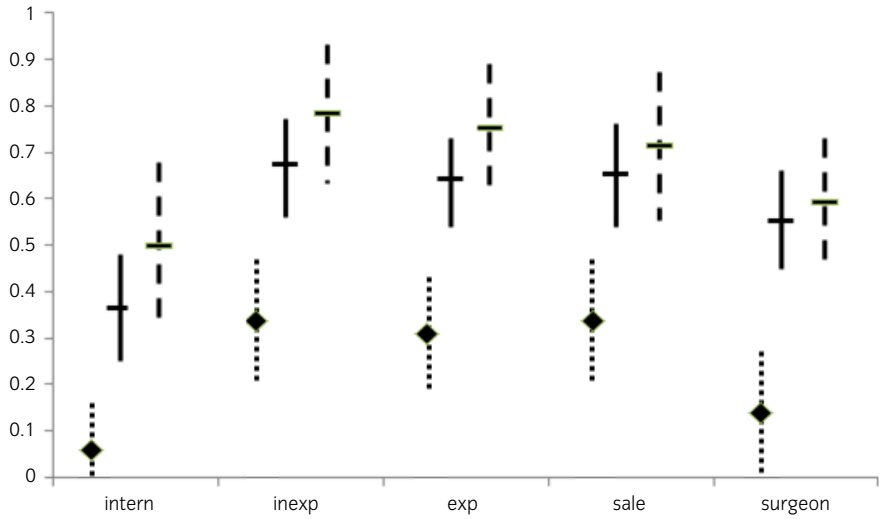

Fig 2: Y axis represents relative kappa values (mean and 95% CI) for interobserver agreement for each of the five experience levels. Dotted line with diamond marker: Cohen's unweighted kappa for ordinal grades; solid line with horizontal marker: Cohen's linear weighted kappa for ordinal grades; dashed line with horizontal marker: Cohen's unweighted kappa for dichotomous grades. Interns consistently had lower interobserver agreement than other experience levels. Agreement increased across all experience levels when Cohen's linear weighted kappa was used for ordinal scales and agreement was greatest when dichotomous scales were used.

in Table 1 and Figure 2. For ordinal grades, mean intraobserver agreement ranged from fair (k = 0.26, 95% CI 0.14–0.38) to good (Ck = 0.61, 95% CI 0.50–0.71), and was best when ordinal scales were transposed to dichotomous grades (k = 0.73, 95% CI 0.59–0.87).

The discordance between individual ordinal grades for each possible combination of inter observer pairs is shown in Figure 3.

Discussion

For the first time, investigation of AFG agreement specific to 2-year-old Thoroughbreds, using veterinarians with different experience has been performed. Levels of agreement were influenced not only by the grading method used (ordinal vs. dichotomous) but by the statistical method chosen for analysis (weighted vs. unweighted). Unweighted kappa requires exact grade agreement between observers for agreement to be considered to exist. As such, this is probably the most realistic test for assessing agreement in a clinical setting, specifically a Thoroughbred sales auction environment where disputes often arise when 'exact' grade agreement does not exist between veterinarians. However, it may unfairly discriminate against subtle interobserver variations which may not be clinically important. It considers, for example, a grade 1 and II.1 interobserver pairing as equally discordant as a grade 1 and IV pairing. As previously reported [6], using this very strict assessment of interobserver agreement, the current investigation identified that if exact agreement of AFG between veterinarians using the ordinal Havemeyer scale was required, only 'fair' agreement could be expected.

Interobserver agreement for the same ordinal grades, using linear weighted kappa analysis was found to be higher, however, with moderate mean agreement documented across the population of observers. This is within the range of agreement previously reported using this statistical method [4,6]. This higher agreement level was anticipated, due to the more forgiving nature of linear weighted kappa, which gives credit for similar grade assignments and does not penalise for the inability to exactly match grades. One flaw of a linear weighted model, however, is that the distance between each grade is considered equal by default, and this may not be optimal for the critical assessment of arytenoid function. For example, the authors believe that more weight should be placed on the difference between grade II.2 and III.1 agreement: the arbitrary line between meeting conditions of sale at public auction or not, than on other grade differences, and this is not reflected in a linear weighted kappa analysis where all weights are equal. Assignment of specific weighting can be controversial [14,15], however, and because the authors did not want to influence the results, an arbitrary weighting model for Kappa was not introduced.

As hypothesised, when ordinal grades were reclassified dichotomously into MCS or DNMCS, the highest agreement across the observer population was achieved, with good interobserver agreement being documented.

No directly comparable investigations to this study exist. Previous investigations have varied on horse type, observer number, observer experience, grading scales used and methods used for statistical analysis [4,5]. Most previous studies have documented findings using the more lenient linear weighted kappa method of analysis [14,15]. One report, using two experienced veterinarians to assess the laryngeal function of draught horses, determined linear weighted interobserver agreement to be good [4], which is higher than the moderate level found in this current study. This discrepancy is likely due to the larger number of veterinarians in this study and their varying experience levels. When studies have reported unweighted kappa data, the levels of agreement are comparable to those in the current investigation. One study in mature Thoroughbreds documented only fair interobserver agreement for ordinal scales across a population of four experienced clinicians. The same study reported substantial agreement when using dichotomous grades of 'asymmetry present or absent' [6]. These observations are similar to those documented in this study.

Similar to the interobserver agreement, the average ability of veterinarians, in this study, to perfectly agree with themselves (unweighted kappa) when using the ordinal seven-point Havemeyer scale was only fair. To the authors' knowledge, such investigation of exact intraobserver agreement has not been previously documented. Mean intraobserver agreement of ordinal grades improved to good across the 10 observers when linear weighted kappa was used, although mean intraobserver agreement for interns and surgeons was only moderate when viewing exact repeat video clips. 'Good' mean intraobserver agreement is slightly less than has been reported previously [4,6]. This is likely due, not only, to the larger number of observers and their varied experience levels but to the presumed higher prevalence of subjective grades included in this study. Similar to the results for interobserver agreement, the intraobserver agreement was highest when ordinal scales were reclassified dichotomously, with good mean agreement being documented across the veterinarian population. Although dichotomous grades in this current study and in previous investigations [4] have elicited the highest intra- and interobserver agreement, this binary scale

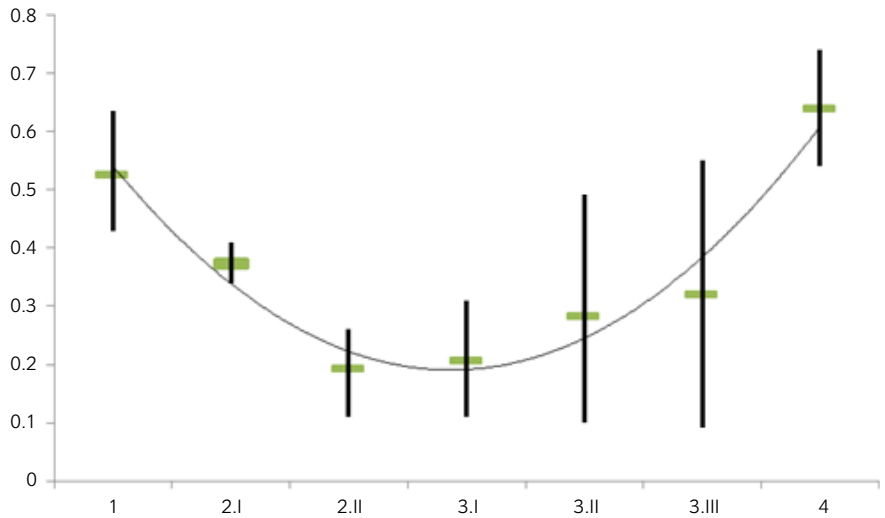

Fig 3: Agreement per observer for ordinal grade of each of 90 videoscopes vs. all other observers. Horizontal bars represent the mean for all 10 observers (95% CI represented by the vertical line transecting the mean marker). Y axis is '% agreement' and x axis represents ordinal laryngeal grades. Polynomial best fit curve is added to highlight the trend observed.

may not currently be acceptable in a clinical setting. In clinical practice and specifically at Thoroughbred public auctions, great significance is placed on exact grade assignments, even when the arytenoid function 'meets conditions of sale'. The ability of a horse to sell, or not, and the sale price can be affected by this individual grade.

The current investigation exposes a lack of strong 'exact' agreement using the Havemeyer scale, despite all observers assessing the same video clips, in which all extraneous and physiological factors are inherently standardised and in which the only remaining variable, therefore, is the veterinarian. Historically, disorders that feature lower levels of agreement suggest either poorly defined grading scales and/or large subjectivity of grading [6]. As the Havemeyer grading scale has been well described [4,10,11], subjectivity in grading is surmised to be responsible for the lack of exact agreement between veterinarians. In addition to the effect on sale prices, the inability of veterinarians to reliably grade arytenoid function has consequences for the ability to accurately monitor disease, communicate successfully with peers and accurately interpret research. In human medicine, training is introduced to improve agreement when large discrepancies between observers exist [15]. It has been suggested that more guidance and formal instruction in grading arytenoid function be instituted for veterinarians [6]. Results from this study would support this proposal. Future studies to determine if additional training could improve agreement or if subjectivity is inherent to arytenoid function grading using ordinal scales are indicated.

We hypothesised that the experience level of the veterinarian would affect agreement. Not surprisingly, interns as a group demonstrated the lowest ability to agree for both intra- and interobserver agreement, using all scales and statistical methods of analysis. Interns likely lack the experience and knowledge to accurately interpret their observations. Surprisingly, however, surgeons in this study demonstrated only moderate ability to agree with themselves when viewing the exact same video clip. It is possible that in-house surgeons and interns, surveyed in this current study, are not exposed to the volume of routine URT examinations that the other three categories of veterinarians are exposed to. Whilst interns demonstrated the lowest level of agreement, interestingly, no one single experience level excelled above the others in their overall agreement ability, using the ordinal scale. This may suggest a level of arytenoid function grading subjectivity that cannot be overcome by experience alone. In this investigation, the only measured parameter where an experience group demonstrated notably higher agreement was for intraobserver agreement using the dichotomous reclassification. 'Experienced' and 'sale' veterinarians demonstrated excellent agreement here, suggesting that experience may increase veterinarians' ability to agree with themselves on whether a horse's arytenoid function 'meets conditions of sale'.

It should be noted that for many assessments in this study, surgeon B's and intern A's interpretation of arytenoid grades were notably different from most of the other veterinarians, including their experience level partners. This may have negatively affected interpretation of interobserver and mean intraobserver agreement for these experience levels, as can be evidenced by the associated large confidence intervals (95% CI). This discrepancy in interpretation for surgeon B and intern A, however, highlights a 'real-world' challenge related to accuracy and reliability of arytenoid assessment when performed by any one individual. It also delineates the risk of using only two observers per category. Future studies using larger numbers of observers in each experience group may offset this potential sampling error. The effect of one individual's grading should also be considered when reviewing published literature. Several previous studies have submitted conclusions regarding AFGs and associations with future performance based on grades assigned by either a single observer [3] or from a small number of observers whose inter- and intraobserver agreement levels were not first formally determined [2]. Whilst we do not dispute the results of these studies, the current investigation highlights the presence and importance of variability in inter- and intraobserver agreement associated with AFG assignment.

Levels of agreement for AFG assignment have also previously been shown to vary depending on the specific ordinal grade being analysed. Historically, most AFG disagreements have occurred in assessment of mildly asynchronous grades, rather than in assessment of the ability to maintain full abduction. One study in which three reviewers evaluated 108 videos demonstrated that the majority (79%) of disagreement was due to fluctuations between synchronous full arytenoid abduction and asynchronous full abduction [5]. Results from the current study demonstrate poorest ordinal grade agreement involved Havemeyer grades II.2 and III.1: grades that signify the difference between 'meets conditions of sale' and 'does not meet conditions of sale'. It should be noted that this does not mean that some veterinarians graded II.2 and some graded III.1 for the same video clip. Because dichotomous interobserver agreement remained 'good', it is likely that veterinarians graded better than II.2 and worse than III.1 for some of these subjective cases but were still able to agree on whether the horse did or did not meet conditions of sale. This highlights the difficulty faced in grading subtle variations in equine arytenoid function, on either side of the meeting conditions of sale grade and underscores the subjectivity involved in grading the equine larynx [5,6]. As expected, laryngeal grades 1 and IV, representing perfect arytenoid function and complete paralysis, respectfully, produced the most agreement in this study, with the interim grades producing an inverted bell curve of agreement between observers (Fig 3). Differences in the grade at which most discrepancy occurs between this and other studies may be

due to many factors, such as the greater number of observers used and the fact that in previous investigations, most observers were considered 'experienced'. Additionally, as the aim of this study was to challenge veterinarians' grading abilities, the video compilation consisted primarily of cases with more subjective arytenoid function grades. Grades II.2 and III.1 were, therefore, likely over-represented and this too may have contributed to AFG disagreement differences between this and other studies.

One limitation of this study design was that the classifications of the 10 veterinarians into groups was based on their experience and qualifications, and cross over may have been present to some degree. For example, one veterinarian in the sales group and one in the senior associates' group were also board-certified surgeons. Both individuals practice as ambulatory veterinarians and have not been dedicated surgeons for over 10 years and, as such, it was deemed acceptable to place them in the subgroup that most accurately reflected their current practice.

In conclusion, AFG agreement levels were found to vary depending on the grading scales chosen and the exacting nature of the statistical agreement criteria imposed.

The ability of veterinarians to exactly agree with each other and themselves on AFG assignment, as would most often be expected to take place in clinical practice, was low. Agreement levels increased for all veterinarians when the ordinal Havemeyer scale was reclassified dichotomously.

It is recommended, therefore, that future studies be directed towards investigating the effect of veterinary AFG training on agreement levels. If ultimately improvement of exact grade agreement cannot be achieved, industry expectations regarding such agreement may need to be modified.

Authors' declaration of interests

No competing interests have been declared.

Ethical animal research

Research ethics committee oversight not required by this journal: retrospective analysis of clinical data. Explicit owner informed consent for inclusion of animals in this study was not stated.

Source of funding

None.

Acknowledgements

The authors would like to thank the veterinarians that generously gave their time to participate in this study.

Authorship

J. McLellan and S. Plevin contributed equally to the study design. S. Plevin was responsible for manuscript preparation. J. McLellan performed all data analysis and interpretation. S. Plevin approved final manuscript.

Manufacturers' addresses

[a]RF Lab Systems, Nagano, Japan.
[b]Microsoft, Redmond, Washington, USA.
[c]Dropbox, San Francisco, California, USA.
[d]www.vassarstats.net

References

1. Embertson, R.M. (1998) Evaluation of the young horse upper airway: what is normal, and what is acceptable? *Proc. Am. Assoc. Equine Practnrs*. **44**, 34-38.
2. Stick, J.A., Peloso, J.G., Morehead, J.P., Lloyd, J., Eberhart, S., Padungton, P. and Derkson, F.J. (2001) Endoscopic assessment of airway function as a predictor of racing performance in Thoroughbred yearlings: 427 cases (1997-2000). *J. Am. Vet. Med. Ass.* **219**, 962-967.
3. Garrett, K.S., Pierce, S.W., Embertson, R.M. and Stromberg, A.J. (2010) Endoscopic evaluation of arytenoid function and epiglottic structure in Thoroughbred yearlings and association with racing performance at two to four years of age: 2,954 cases (1998-2001). *J. Am. Vet. Med. Ass.* **236**, 669-673.
4. Perkins, J.D., Salz, R.O., Schumacher, J., Livesey, L., Piercy, R.J. and Barakzai, S.Z. (2009) Variability of resting endoscopic grading for assessment of recurrent laryngeal neuropathy in horses. *Equine Vet. J.* **41**, 342-346.
5. Hackett, R.P., Ducharme, N.G., Fubini, S.L. and Erb, H.N. (1991) The reliability of endoscopic assessment of arytenoid cartilage movement in horses. Part 1: subjective and objective laryngeal evaluation. *Vet. Surg.* **20**, 174-179.
6. McGivney, C.L., Sweeney, J., David, F., O'Leary, J.M., Hill, E.W. and Katz, L.M. (2016) Intra- and interobserver reliability estimates for identification and grading of upper respiratory tract abnormalities recorded in horses at rest and during overground endoscopy. *Equine Vet. J.* **49**, 433-437.
7. Anderson, B.H., Kannegieter, N.J. and Goulden, B.E. (1997) Endoscopic observations on laryngeal symmetry and movements in young racing horses. *N. Z. Vet. J.* **45**, 188-192.
8. Ducharme, N.G., Hackett, R.P., Fubini, S.L. and Erb, H.N. (1991) The reliability of endoscopic examination in assessment of arytenoid cartilage movement in horses. Part II. Influence of side of examination, reexamination, and sedation. *Vet. Surg.* **20**, 180-184.
9. Archer, R.M., Lindsay, W.A. and Duncan, I.D. (1991) A comparison of technique to enhance the evaluation of equine laryngeal function. *Equine Vet. J.* **23**, 104-107.
10. Dixon, P., Robinson, E. and Wade, J.F. (2003) Proceedings of a workshop on Equine Recurrent Laryngeal Neuropathy. *Havemeyer Found. Monogr. Ser.* **11**, 93-97.
11. Robinson, N.E. (2004) Consensus statements on equine recurrent laryngeal neuropathy: conclusions of the Havemeyer Workshop. *Equine Vet. Educ.* **16**, 333-336.
12. Available at: https://aaep.org/guidelines/aaep-ethical-and-professional-guidelines/aaep-position-statements/sale-issues.
13. Available at: https://www.keeneland.com/sales/public-auction-sales-code-conduct.
14. Brennan, P. and Silmen, A. (1992) Statistical methods for assessing variability in clinical measures. *Br. Med. J.* **304**, 1491-1494.
15. Hallgren, K.A. (2012) Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor. Quant. Methods Psychol.* **8**, 23-34.