Received: 26 March 2020 | Accepted: 13 August 2020

DOI: 10.1111/evj.13334

GENERAL ARTICLE

Equine Veterinary Journal WILEY

Inter-observer variability of two grading systems for equine glandular gastric disease

Rose Tallon | Michael Hewetson

Royal Veterinary College, Hatfield, Hertfordshire, UK

Correspondence
Rose Tallon, Royal Veterinary College, Hawkshead Lane, North Mymms, Hatfield, Hertfordshire, AL9 7TA, UK.
Email: rosetallon@gmail.com

Abstract

Background: Equine glandular gastric disease (EGGD) is recognised as a separate entity to equine squamous gastric disease (ESGD) and it is recommended that lesions are graded differently. Currently, no validated scoring system exists for EGGD.

Objectives: To determine inter-observer reliability of two previously described grading systems for EGGD and to assess if agreement improved with gastroscopy experience, specialist training or familiarity with the descriptive system.

Study design: Cross-sectional survey.

Methods: A link to an electronic questionnaire containing 20 images of glandular lesions was circulated. Respondents were asked to score lesions using descriptive terminology and a 0-2 verbal rating scale (VRS). Krippendorff's alpha reliability estimate was used to assess inter-rater agreement. A mixed effects model was used to determine which descriptive categories were associated with lesions being described as severe and decision to treat.

Results: Eighty-two veterinarians responded, 49 diplomates and 33 non-diplomates. There was no agreement when all four descriptive variables were combined ($\alpha = 0.19$). Agreement was fair to moderate for severity ($\alpha = 0.52$), distribution ($\alpha = 0.44$), appearance ($\alpha = 0.38$) and shape ($\alpha = 0.32$). Agreement for the VRS was similar to that for severity ($\alpha = 0.53$). Agreement was better among diplomates across all categories. Lesion appearance and shape, but not distribution, were associated with both a decision to treat; and lesions being described as severe ($P =< .05$). A VRS score 2/2 was associated with a lesion being described as severe (OR 75.2, 95% CI 51.12-110.48, $P =< .001$).

Main limitations: Intra-observer variability was not assessed. The number of images is relatively small, and the decision to treat is based on several factors in practice.

Conclusions: Overall, agreement for the descriptive system was poor. Better delineation of descriptive category boundaries and characteristics should be determined. Agreement was similar when comparing the severity category and the VRS. Extrapolation to a VRS based on lesion severity may therefore be possible.

KEYWORDS

horse, gastric ulcer, glandular disease, interobserver variability, grading systems

The abstract is available in Portuguese in the Supporting Information section of the online version of this article.

1 | INTRODUCTION

Ordinal scoring systems are commonly used in both human and veterinary medicine. They facilitate documentation of disease, assessment of response to treatment and standardisation of research. Equine gastric ulcers were historically scored using a 0-4 grading scale recommended by the Equine Gastric Ulcer Council.[1] This system was shown to have better inter-observer agreement than a number/severity scoring system, with high kappa values (>0.8) between three observers for both squamous and glandular lesions.[2] Equine glandular gastric disease (EGGD) is now considered a separate entity to equine squamous gastric disease (ESGD) with regard to risk factors, clinical signs, pathophysiology, treatment and prognosis.[3-5] The incomplete understanding of the pathophysiology of glandular disease makes sub-classification of lesions difficult and means that grading systems for squamous disease may not be accurate for EGGD. The current recommendation is to use descriptive terminology, which classifies lesions based on four categories; severity, distribution, shape and appearance.[6] A recent statement[3] proposed inclusion of the terms nodular and erythematous to allow a more accurate description of lesions. A novel verbal rating scale (VRS), grading lesions from 0 to 2 has also been used in research.[7] Verbal rating scales are commonly used to score pain in people[8-10] and are defined as ordinal scales where words are used to describe the severity of a condition.

Although there is poor correlation between endoscopic findings and histological analysis of lesions,[11-12] gastroscopy remains the best method for antemortem diagnosis of EGGD. The use of an accurate and repeatable grading system is important in both clinical and research settings. Currently, no validated scoring system exists for EGGD, and recent published work has reverted to the original EGUS scale to group data and facilitate statistical analysis.[4,13,14]

The main objectives of this study were (a) to determine inter-observer reliability of descriptive terminology and a verbal rating scale (VRS) for EGGD and (b) to assess if agreement improved with gastroscopy experience, specialist training or familiarity with the descriptive system. It was hypothesised that there would be poor agreement for both scales and that agreement would be better among experienced endoscopists, those with specialist training and those familiar with the descriptive system. To ascertain which factors were associated with respondents considering a lesion to be clinically significant, a secondary objective was to determine which other descriptive variables were associated with lesions being described as severe and which factors influenced the decision to treat.

2 | MATERIALS AND METHODS

An electronic questionnaire containing 20 images of glandular lesions in the antrum and pylorus of the stomach was drafted (Data S1). All questions were close-ended, and the survey was anonymous. A set of introductory questions established the respondent's experience of gastroscopy, specialist status and scoring system currently used. Following this, a series of 20 still images of gastric glandular lesions were displayed sequentially. Respondents were asked to grade each image using the current descriptive terminology[3] (Figure 1) and a verbal rating scale[7] (Figure 2). For each image, respondents were asked whether they would recommend treatment based solely on that image. Although artificial, this was asked as a measure of ascribing clinical significance. Each respondent viewed the images in the same order and could navigate back to previous images.

Electronic invitations containing a link to complete the questionnaire were circulated to both specialists and primary care veterinarians using listservs including the American College of Veterinary Internal Medicine (ACVIM), the European College of Equine Internal Medicine (ECEIM) and a UK-based mailing list of both first opinion and specialist practitioners (Equine Veterinary Group UK). The survey had to be completed on a single attempt and was closed to responses after a 3-month period.

2.1 | Data analysis

A Chi-squared test was used to compare differences between specialist and non-specialist groups with regard to scoring system used and gastroscopy experience. Krippendorff's alpha reliability estimate was used to assess inter-rater agreement for both ordinal and nominal data. This was selected over Fleiss' kappa for its capacity to analyse both ordinal and nominal data and to account for any missing data.[15] Similar to Fleiss' kappa, a coefficient of 0 reflects that any agreement between raters is due to chance alone, while a coefficient of 1 reflects perfect agreement across all raters on all items.[16] Bootstrapping to 1000 was used to generate 95% confidence intervals. Agreement was assessed for the VRS and for each descriptive category individually as well as all four combined. Survey responses were then stratified by scoring system currently used, by experience (>10 gastroscopies per month) and by diplomate status, and the analysis was repeated.

Generalised linear mixed effect models, with image and observer included as random effects, were used to determine which other descriptive parameters, if any, contributed to (a) a respondents' decision to treat; and (b) lesions being described as severe. All responses that considered an image to be normal were excluded. The remaining images were classed as 'severe' or 'non-severe' (mild and moderate), based on how they were graded for severity using the descriptive system. Variables targeted for inclusion in the model were other descriptive categories (size, shape and distribution). Experience, scoring system normally used and diplomate status were also included. Variables where associations with the dependent variable had a $P \leq .2$ were used in the initial model (Table S1). Backward elimination selection methods were used to eliminate any non-significant variables. The intra-class correlation coefficient was determined for each model to assess the contribution of the random effects. The Hosmer-Lemeshow goodness of fit test was used to confirm suitability of the models for the data. A Chi-squared test was used to determine if there was an association between description of severity and the VRS. Significance was set at $P =\leq .05$ and all statistical analyses

FIGURE 1 Descriptive system for equine glandular gastric disease, as described by Rendle et al[3]

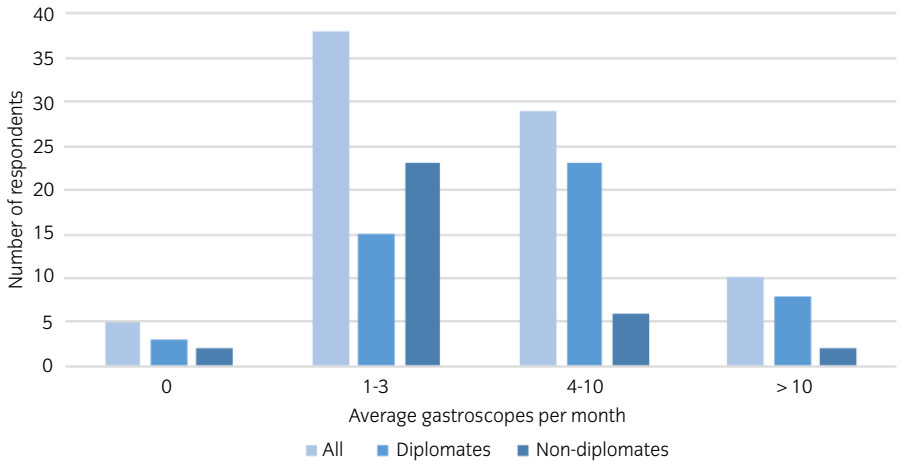
Severity	Mild – Moderate – Severe
Distribution	Focal – Multifocal – Diffuse
Shape	Depressed – Flat – Nodular – Raised
Appearance	Erythematous – Haemorrhagic – Fibrinosuppurative

FIGURE 2 Verbal Rating Scale (VRS) for grading of glandular lesions as described by Sykes et al[7]

Grade	
0	Normal mucosal surface with no evidence of loss of mucosal integrity
1	Mild to moderate lesion(s) with evidence of loss of mucosal integrity
2	Severe lesion(s) with evidence of loss of mucosal integrity

Mild hyperaemia of the glandular mucosa without visible loss of mucosal integrity is not considered to be clinically significant.

FIGURE 3 Experience of respondents, showing average number of gastroscopy examinations performed per month



were carried out using standard software (SPSS v25, IBM, IBM Software).

3 | RESULTS

Eighty-two veterinarians responded. Forty-nine (60%) were diplomates of equine internal medicine (DACVIM/DECEIM) and 33 (40%) were non-specialist veterinarians. Twenty-five respondents (30%) worked in first opinion practice, 24 (29%) in referral practice and 14 (17%) in both fields.

Of the internal medicine diplomates, 28 (57%) used the descriptive system, 16 (33%) used the original EGUS 0-4 scoring system, four (8%) used both and one respondent (2%) used the descriptive system in combination with the VRS. Among non-specialists, 16 (49%) used the descriptive system, 13 (39%) used the EGUS system and three (9%) used both. There was no difference between the groups for scoring system used ($P = .9$). The average number of gastroscopic examinations performed per month is displayed in Figure 3. Specialists were more likely to perform a higher number of gastroscopies per month than non-specialists ($P = .004$).

3.1 | Inter-observer agreement

Inter-observer agreement coefficients for descriptive and VRS grading systems are displayed in Table 1. The same cut-offs as Cohen's kappa can be used when interpreting Krippendorff's alpha coefficient, with $\alpha > 0.8$ considered to reflect strong agreement.[16,17] Agreement was fair to moderate for severity ($\alpha = 0.52$), distribution ($\alpha = 0.44$), appearance ($\alpha = 0.38$) and shape ($\alpha = 0.32$). Agreement for the VRS was similar to that for severity ($\alpha = 0.53$). Agreement was higher among specialists than non-specialists for all descriptive categories and across both scoring systems. Agreement was higher among respondents who currently use the descriptive system in practice, regardless of diplomate status. Overall, the VRS and 'severity' category showed similar inter-observer agreement. Amongst non-diplomates, the VRS had higher agreement than the 'severity'

category. When diplomates, those using the descriptive system in practice and experienced respondents (average >10 gastroscopies per month) were examined separately, the VRS had lower agreement than the 'severity' category.

3.2 | Relationship between VRS and description of severity

Description of severity was associated with VRS score ($P < .001$). A lesion with a VRS score of 2 was more likely to be described as severe than a lesion with a VRS score of 1 (OR 75.2, 95% CI 51.12-110.48, $P < .001$).

3.3 | Factors contributing to lesions being described as severe

Univariable analysis is presented in Table S1 and results of multivariable analysis are presented in Table 2. Appearance ($P = .005$) and shape ($P < .001$), but not distribution ($P = .08$) were associated with lesions being described as severe. Depressed lesions were more likely to be described as severe compared to flat lesions (OR 4.6, 95% CI 2.22-9.55, $P < .001$). Haemorrhagic or fibrinosuppurative lesions were more likely to be described as severe than erythematous lesions (OR 2.9, 95% CI 1.51-5.39, $P = .001$ and OR 2.9, 95% CI 1.51-5.71, $P = .001$, respectively). Diplomates were less likely to describe lesions as severe (OR 0.5, 95% CI 0.28-0.94, $P = .03$). Experience level and scoring system currently used did not contribute to lesions being described as severe. Intraclass correlation coefficient was .6 for image ($P = .01$) and .27 for observer ($P < .001$), indicating that image accounted for the majority of clustering. The Hosmer-Lemeshow test indicated acceptable model fit ($\chi^2 = 11.51$, $P = .17$).

3.4 | Factors contributing to decision to treat

Results of univariable analysis are presented in Table S2 and multivariable analysis is displayed in Table 3. Appearance ($P < .001$) and shape ($P = .03$) were associated with decision to treat. Respondents were more likely to treat depressed lesions compared to flat lesions (OR 3, 95% CI 1.22-7.63, $P = .02$). Distribution was not associated with decision to treat ($P = .13$). Diplomates were less likely to treat lesions (OR 0.5, 95% CI 0.24-0.93, $P = .03$) than non-diplomates. Intraclass correlation coefficient was .4 for image ($P = .02$) and .3 for observer ($P =< .001$). The Hosmer-Lemeshow test indicated acceptable model fit ($\chi^2 = 7.30$, $P = .4$).

4 | DISCUSSION

Poor inter-observer agreement for endoscopy in people has been previously described, with less experienced operators performing worse.[18,19] The Havemeyer scale for grading arytenoid function was found to have fair to moderate agreement between observers, which improved when scales were transposed to dichotomous grades.[20] Ordinal scales have also been shown to have poor inter-rater agreement when assessing lameness in horses.[21,22] Inter-observer agreement for the presence of EGGD has been shown to be weak ($\kappa = 0.42$), with higher agreement ($\kappa = 0.56$) on whether lesions were deemed clinically significant.[23]

TABLE 1 Results of Krippendorff's alpha reliability estimate for the descriptive and verbal rating scale (VRS) scoring systems showing 95% confidence intervals and interpretation of agreement

		All	Diplomates	Non-diplomates	>10 scopes per month	Currently using descriptive system
Descriptive	Overall	0.19 0.1903-0.1965 None	0.21 0.2058-0.2176 Minimal	0.17 01613-0.1778 None	0.21 0.1845-0.2437 Minimal	0.22 0.2119-0.2236 Minimal
	Severity	0.52 0.5169- 0.5286 Weak	0.63 0.6187-0.6350 Moderate	0.41 0.3892-0.4232 Weak	0.49 0.4364-0.5393 Weak	0.61 0.6026-0.6212 Moderate
	Shape	0.32 0.3110- 0.3226 Minimal	0.36 0.3525-0.3718 Minimal	0.26 0.2480-0.2737 Minimal	0.37 0.3217-0.4164 Minimal	0.35 0.3382-0.3567 Minimal
	Appearance	0.38 0.3704-0.3808 Minimal	0.40 0.3921-0.4097 Weak	0.34 0.3280-0.3538 Minimal	0.37 0.3183-0.4067 Minimal	0.41 0.4000-0.4176 Weak
	Distribution	0.44 0.4395- 0.4516 Weak	0.47 0.4645-0.4844 Weak	0.41 0.3983-0.4625 Weak	0.43 0.3867-0.4846 Weak	0.48 0.4726-0.4921 Weak
	VRS (0-2)	0.53 (0.5242-0.5355) Weak	0.55 0.5396-0.5570 Weak	0.51 0.5004-0.5278 Weak	0.45 0.3969-0.5066 Weak	0.59 0.5822-0.5990 Weak

TABLE 2 Multivariable binomial logistic regression to determine which factors were associated with lesions being described as 'severe' in the descriptive scoring system. Image and observer are included as random effects

	Non-severe n = 1116	%	Severe n = 319	%	OR	95% CI	*P* value
Shape							<.001
Flat	629	58.4	127	41.4	Base		
Depressed	88	8.2	81	26.4	4.6	2.22-9.55	<.001
Nodular	124	11.5	27	8.8	2.8	1.27-5.97	.01
Raised	236	21.9	72	23.5	1.9	1.17-3.18	.01
	1077		307				
Appearance							.005
Erythematous	540	49.8	38	12	Base		
Haemorrhagic	230	21.2	159	50.2	2.9	1.51-5.39	.001
Fibrinosuppurative	228	21	94	29.7	2.9	1.51-5.71	.001
Mixed	86	7.8	26	8.2	1.7	0.64-4.43	.3
	1084		317				
Diplomate status							.03
Non-diplomate	435	39	152	47.6	Base		
Diplomate	681	61	167	52.4	0.5	0.28-0.94	.03
	1116		319				

TABLE 3 Multivariable binomial logistic regression to determine which factors were associated with lesions being deemed to warrant treatment. Image and observer are included as random effects

	No Treat n = 239	%	Treat n = 1189	%	OR	95% CI	*P* value
Shape							**.03**
Flat	176	74.6	578	50.4	Base		
Depressed	13	5.5	156	13.6	3.1	1.22-7.63	.02
Nodular	15	6.4	136	11.9	2.3	0.91-5.67	.08
Raised	32	13.6	277	24.1	1.0	0.53-1.97	.9
Appearance							**<.001**
Erythematous	199	90.9	375	31.9	Base		
Haemorrhagic	7	3.2	381	32.4	8.6	3.31-22.54	<.001
Fibrinosuppurative	9	4.1	312	26.5	10.4	4.36-24.91	<.001
Mixed	4	1.7	108	9.2	7.7	2.11-28.03	.002
Diplomate status							**.03**
Non-diplomate	80	33.5	503	42.3	Base		
Diplomate	159	66.5	686	57.7	0.5	0.24-0.93	.03

As expected, overall agreement in this study was poor, particularly when all four descriptive variables were combined. The use of four descriptors generates a large combination of outcomes, particularly for lesions with a mixed appearance (eg erythematous and fibrinosuppurative). This becomes challenging to analyse in a research setting. This was a much larger study than that used to validate the original EGUS scoring system, in which three diplomates were used.[2] The large number of observers increases the likelihood of disagreement. However, even when responses were grouped by specialist status and experience, agreement remained poor. Diplomates had better agreement across all categories, possibly reflecting additional training or greater experience. Interestingly, when experience alone was examined regardless of specialist status, this was not the case but among respondents already using the descriptive system in practice, regardless of specialist status, agreement across all descriptive categories was comparable to the diplomate group. This suggests that additional training to increase familiarity with the system may improve agreement. The VRS had the least agreement amongst the experienced group but performed better than the severity descriptor

amongst non-diplomates. The use of a number may reflect the original EGUS scale which respondents may be more familiar with.

When descriptive variables were examined individually, severity had the highest level of agreement, although this was only moderate. This is an unexpected finding as it is arguably the most subjective parameter, as the endoscopic assessment of severity alone cannot be used to infer clinical signs.[6] Agreement was worst for shape and appearance which may reflect an unfamiliarity with the terms in use and the range of lesions seen in the glandular mucosa. Clear language for both defining and setting boundaries for each category may be necessary to improve agreement. Utilisation of descriptive terminology was less common among non-specialists. This likely represents a difference in training but may be due to unfamiliarity with the descriptive system or to facilitate communication with owners and trainers who may be more familiar with the original scoring system.

Appearance of loss of mucosal integrity (ie haemorrhagic, fibrinosuppurative and depressed lesions) was associated with lesions being described as severe. Depressed lesions may reflect the appearance of a traditional ulcer or erosion, rather than an inflammatory process per se. Flat and erythematous lesions were less likely to be considered severe. Lesions with a mixed appearance were not associated with severity, although this may be due to the low numbers in this category. Distribution was not associated with lesions being described as severe. This is at odds with the original EGUS scoring system which was based around lesion distribution.[1] The VRS does not take distribution into account. Given the association between severity and VRS, it may be possible to extrapolate this to the VRS, or similar, using the severity category of the descriptive system. Grouping lesions for statistical analysis could be done using severity as the primary variable, with a 0-3 scale to incorporate normal, mild, moderate and severe. This would add an additional category, making it easier to document improvement as well as resolution of lesions.

4.1 | Decision to treat

The same factors (shape and appearance) associated with a lesion being considered severe were associated with the decision to treat. This is unsurprising, as a severe lesion would typically infer clinical significance. Diplomates were less likely to treat lesions than non-diplomates. The reason for this is unclear. In practice, the decision to treat may be driven by several factors including clinical signs, owner/trainer demands, financial constraints and individual clinician preference.

4.2 | Limitations

This was an opt-in survey and may not be representative of specialist and non-specialist populations as a whole. It is unknown what degree of familiarity respondents had with the various scoring systems, as data were only collected regarding which system each respondent currently used. The use of a scoring system does not necessary reflect familiarity with that system. It was impossible to exclude bias, with veterinarians being aware that they were participating in a study and knowing that their answers would be viewed. Responses were completely anonymised to minimise the impact of bias, however, an effect on precision and diagnostic accuracy may remain. There is evidence that inter-rater reliability declines under less controlled conditions.[24] Consistency of interpretation was difficult to control. Diagnostic drift is a situation when the assignment of scores may vary slightly in consistency through the scoring process. This may occur in situations such as this, where there are a large number of samples to be examined or when category characteristics/boundaries are poorly defined.[25] Observers may also subconsciously compare and score an image relative to those previously viewed, particularly for more subjective parameters such as severity. Inter-observer variability is arguably less important in a clinical setting, particularly if the same clinician is performing follow-up examinations. Assessment of intra-observer reliability is required to see how these scoring systems perform amongst individual clinicians.

The 0-4 EGUC scoring system was not included in this study as the current recommendation is that it should not be applied to EGGD.[6] Although this scale was previously validated using both squamous and glandular lesions, this was performed at a time when glandular disease was not considered a separate pathological process to squamous disease.

Still images were used in this study and may be less representative of scoring in clinical practice. Good quality video clips may allow better visualisation of depth, colour and assessment of artefacts eg blanching of the mucosa. Further work to compare agreement of scores of still images compared to video clips may help to quantify this difference. In practice, the decision to treat lesions is based on many factors. This study asked respondents this question based solely on the image, without providing any background information regarding history, clinical signs or the presence of ESGD, which is somewhat artificial. The number of images included was based on a previous EGUS validation study[2] but is relatively small and may not have allowed for all possible combinations of lesion types to be included.

Another approach to validate a scoring system is to analyse the relationship between the scores and relevant parameters of disease severity.[25] To the authors' knowledge, there is currently no published work examining the response of specific types of glandular lesions to treatment. The lack of information on biopsies combined with current incomplete understanding of the pathophysiology and clinical signs pertaining specifically to EGGD means that this cannot be undertaken at present but warrants future attention.

5 | CONCLUSION

There was no inter-observer agreement for the descriptive system when all four variables were included. The severity category

showed the best agreement, and this was similar to the VRS. Severity was significantly associated with the VRS, suggesting that it may be possible to extrapolate to the latter. The lack of agreement for appearance, shape and distribution identified in this study questions the need for better definition of these particular parameters. As more is understood about clinical signs pertaining to EGGD and response of specific lesion types to treatment, a more comprehensive scoring system may be developed. In the meantime, additional training to increase familiarity with the descriptive system may improve agreement.

ACKNOWLEDGEMENTS

The authors express their appreciation to Professor David Brodbelt for his advice regarding data analysis.

CONFLICT OF INTEREST

No competing interests have been declared.

AUTHOR CONTRIBUTIONS

R. Tallon gathered, analysed and interpreted the data and drafted the manuscript. M. Hewetson critically revised the manuscript. Both authors were involved in conception, study design and approval of the final version of the manuscript.

ETHICAL ANIMAL RESEARCH

The study was approved by the Social Sciences Research Ethical Review Board at the Royal Veterinary College (URN SR2018-1678).

INFORMED OWNER CONSENT

Images used in this study were obtained from clinical records and used anonymously. Completion of the questionnaire was taken as participant consent.

DATA ACCESSIBILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/evj.13334.

ORCID

Rose Tallon https://orcid.org/0000-0002-9917-8719

REFERENCES

1. The Equine Gastric Ulcer Council. Recommendations for the diagnosis and treatment of equine gastric ulcer syndrome (EGUS): The Equine Gastric Ulcer Council. Equine Veterinary Education. 1999;11(5):262–272. http://doi.org/10.1111/j.2042-3292.1999.tb00961.x
2. Bell RJW, Kingston JK, Mogg TD. A comparison of two scoring systems for endoscopic grading of gastric ulceration in horses. N Z Vet J. 2007;55(1):19–22.
3. Rendle D, Bowen M, Brazil T, Conwell R, Hallowell G, Hepburn R, et al. Recommendations for the management of equine glandular gastric disease. UK Vet Equine. 2018;2(Suppl 1):2–11.
4. Sykes BW, Bowen M, Habershon-Butcher JL, Green M, Hallowell GD. Management factors and clinical implications of glandular and squamous gastric disease in horses. J Vet Intern Med. 2019;33:233–40.
5. Mönki J, Hewetson M, Virtala A-MK. Risk factors for equine gastric glandular disease: a case-control study in a Finnish Referral Hospital population. J Vet Intern Med. 2016;30:1270–5.
6. Sykes BW, Hewetson M, Hepburn RJ, Luthersson N, Tamzali Y. European College of Equine Internal Medicine Consensus Statement-equine gastric ulcer syndrome in adult horses. J Vet Intern Med. 2015;29:1288–99.
7. Sykes BW, Kathawala K, Song Y, Garg S, Page SW, Underwood C, et al. Preliminary investigations into a novel, long-acting, injectable, intramuscular formulation of omeprazole in the horse. Equine Vet J. 2017;49(6):795–801.
8. Brunelli C, Zecca E, Martini C, Campa T, Fagnoni E, Bagnasco M, et al. Comparison of numerical and verbal rating scales to measure pain exacerbations in patients with chronic cancer pain. Health Qual Life Outcomes. 2010;8:42.
9. Haefeli M, Elfering A. Pain assessment. Eur Spine J. 2006;15(Suppl 1):S17–S24.
10. Kim TK. Practical statistics in pain research. Korean J Pain. 2017;30:243–9.
11. Andrews FM, Reinemeyer CR, McCracken MD, Blackford JT, Nadeau JA, Saabye L, et al. Comparison of endoscopic, necropsy and histology scoring of equine gastric ulcers. Equine Vet J. 2002;34(5):475–8.
12. Pietra M, Morini M, Perfetti G, Spadari A, Vigo P, Peli A. Comparison of endoscopy, histology, and cytokine mRNA of the equine gastric mucosa. Vet Res Commun. 2010;34:121–4.
13. Bush J, van den Boom R, Franklin S. Comparison of aloe vera and omeprazole in the treatment of equine gastric ulcer syndrome. Equine Vet J. 2018;50(1):34–40.
14. Varley G, Bowen IM, Habershon-Butcher JL, Nicholls V, Hallowell GD. Misoprostol is superior to combined omeprazole-sucralfate for the treatment of equine gastric glandular disease. Equine Vet J. 2019;51:575–80.
15. Zapf A, Castell S, Morawietz L, Karch A. Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? BMC Med Res Methodol. 2016;16:93.
16. Hayes A, Krippendorff K. Answering the call for a standard reliability measure for coding data. Commun Methods Meas. 2007;1:77–89.
17. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med. 2012;22:276–82.
18. Hyun YS, Han DS, Bae JH, Park HS, Eun CS. Interobserver variability and accuracy of high-definition endoscopic diagnosis for gastric intestinal metaplasia among experienced and inexperienced endoscopists. J Korean Med Sci. 2013;28:744–9.
19. Amano Y, Ishimura N, Furuta K, Okita K, Masaharu M, Azumi T, et al. Interobserver agreement on classifying endoscopic diagnoses of nonerosive esophagitis. Endoscopy. 2006;38:1032–5.
20. McLellan J, Plevin S. Evaluation of videoendoscopic examinations of arytenoid function in the 2-year-old Thoroughbred: can we all agree? Equine Vet J. 2019;51(3):364–9.
21. Leelamankong P, Estrada R, Mählmann K, Rungsri P, Lischer C. Agreement among equine veterinarians and between equine veterinarians and inertial sensor system during clinical examination of hindlimb lameness in horses. Equine Vet J. 2020;52:326–31.

22. Hewetson M, Christley RM, Hunt ID, Voute LC. Investigations of the reliability of observational gait analysis for the assessment of lameness in horses. Vet Rec. 2006;158:852–8.
23. Hewetson M, Venner M, Volquardsen J, Sykes BW, Hallowell GD, Vervuert I, et al. Diagnostic accuracy of blood sucrose as a screening test for equine gastric ulcer syndrome (EGUS) in weanling foals. Acta Vet Scand. 2018;60:24.
24. Topf M. Interrater reliability decline under covert assessment. Nurs Res. 1988;37:47–9.
25. Gibson-Corley KN, Olivier AK, Meyerholz DK. Principles for valid histopathologic scoring in research. Vet Pathol. 2013;50:1007–15.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Tallon R, Hewetson M. Inter-observer variability of two grading systems for equine glandular gastric disease. *Equine Vet J*. 2020;00:1–8. https://doi.org/10.1111/evj.13334