Rater Agreement on Gait Assessment during Neurologic Examination of Horses

E. Olsen, B. Dunkel, W.H.J. Barker, E.J.T. Finding, J.D. Perkins, T.H. Witte, L.J. Yates, P.H. Andersen, K. Baiker, and R.J. Piercy

Background: Reproducible and accurate recognition of presence and severity of ataxia in horses with neurologic disease is important when establishing a diagnosis, assessing response to treatment, and making recommendations that might influence rider safety or a decision for euthanasia.

Objectives: To determine the reproducibility and validity of the gait assessment component in the neurologic examination of horses.

Animals: Twenty-five horses referred to the Royal Veterinary College Equine Referral Hospital for neurological assessment (n = 15), purchased (without a history of gait abnormalities) for an unrelated study (n = 5), or donated because of perceived ataxia (n = 5).

Methods: Utilizing a prospective study design; a group of board-certified medicine (n = 2) and surgery (n = 2) clinicians and residents (n = 2) assessed components of the equine neurologic examination (live and video recorded) and assigned individual and overall neurologic gait deficit grades (0–4). Inter-rater agreement and assessment-reassessment reliability were quantified using intraclass correlation coefficients (ICC).

Results: The ICCs of the selected components of the neurologic examination ranged from 0 to 0.69. “Backing up” and “recognition of mistakes over obstacle” were the only components with an ICC > 0.6. Assessment-reassessment agreement was poor to fair. The agreement on gait grading was good overall (ICC = 0.74), but poor for grades ≤ 1 (ICC = 0.08) and fair for ataxia grades ≥ 2 (ICC = 0.43). Clinicians with prior knowledge of a possible gait abnormality were more likely to assign a grade higher than the median grade.

Conclusion and Clinical Importance: Clinicians should be aware of poor agreement even between skilled observers of equine gait abnormalities, especially when the clinical signs are subtle.

Key words: Agreement; Ataxia; Physical examination; Reliability.

A thorough physical examination in combination with the history is the primary source of initial information for any clinician[1,2] and influences decision making for further diagnostic and therapeutic intervention.[3] Examinations should therefore have good reliability (how well can patients be distinguished from each other), high agreement (low measurement error for repeated assessments)[4] and results of the examination should be valid (accurate).[5] Inaccurate or unreliable assessment of an underlying problem might lead to misdiagnosis or inappropriate further testing and treatments. Consequently, inherent limitations of the physical examination because of its subjectivity should be recognized and the examination optimized to reduce variability and bias.

Abbreviations:

CMID	clinically minimal important difference
CVM	cervical vertebral malformation-malarticulation
ERH	Equine Referral Hospital
ICC	intraclass correlation coefficient
PCA	principal component analysis
RVC	The Royal Veterinary College
SARA	scale for the assessment and rating of ataxia

Few studies have examined reproducibility and validity of physical examinations in horses and humans.[5,6] There is expectation bias in gait and lameness assessment of horses[7] despite good agreement between observers where a lameness score (AAEP, 0–5) was >1.5/5 (κ = 0.86), or low agreement (κ = 0.23), for lameness grades ≤1.5/5.[8] Low agreement is especially of concern when additional testing options are limited and when decisions are made relevant to human safety or animal euthanasia. One such example is neurologic examination of horses where

From the Department of Large Animals Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Taastrup, Denmark (Olsen); the Department of Clinical Sciences and Services, The Royal Veterinary College, London, UK (Olsen, Dunkel, Barker, Finding, Perkins, Witte, Yates, Piercy); the Structure & Motion Laboratory, The Royal Veterinary College, London, UK (Olsen, Perkins, Witte); the Department of Clinical Sciences, Swedish University of Agricultural Sciences, Uppsala, Sweden (Andersen); and the School of Veterinary Medicine and Science, University of Nottingham, Nottingham, UK (Baiker). This work has been presented:

(a) Abstract and poster at the conference of the European College of Equine Internal Medicine (ECEIM), Le Touquet, France, 2013.
(b) Abstract and oral presentation at the 2013 American College of Veterinary Internal Medicine Forum, Seattle, WA.
(c) BEVA spring clinical workshop, U.K., April 2013.
(d) Abildgaard symposium, Taastrup, Denmark, March 2013.
(e) PhD defence, Taastrup, Denmark, 19 June 2013.

Corresponding authors: E. Olsen, Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Hojbakkegaard Allé 5, 2630 Taastrup, Denmark; e-mail: eo@sund.ku.dk and R.J. Piercy, Department of Clinical Sciences and Services, The Royal Veterinary College, London, UK; e-mail: rpiercy@rvc.ac.uk.

Submitted September 11, 2013; Revised December 5, 2013; Accepted January 7, 2014.

DOI: 10.1111/jvim.12320

diagnostic imaging and other testing options are limited by physical constraints, by poor inherent sensitivity or specificity[9] or both and where the prognosis for recovery often is poor.[10]

During assessment of gait in neurologic examination of horses, clinicians typically grade the severity of the neurologic deficit according to a modified grading scale (0–5), where grade 0 is assigned to horses without neurologic deficits and grade 5 is assigned to horses that are recumbent.[11,12] Despite general acceptance and application of this system,[12–18] its reproducibility and validity have never been tested; neither has there been any formal assessment of the role for expectation bias in neurologic examination of horses.

The aim of this study was to evaluate the subjective assessment of horses with and without varying neurologic deficits by evaluating the rater and assessment-reassessment agreement of a modified ataxia-grading scale. In addition, we aimed to determine whether expectation bias plays a role in neurologic assessment of horses. We hypothesized that (1) agreement between raters (inter-rater) is good when assessing each part of the neurologic examination of horses with and without suspected neurologic deficits; (2) assessment-reassessment agreement is good when raters assess horses with and without suspected neurologic deficits twice on video; (3) the modified ataxia-grading scale has good agreement applied to horses with moderate to severe ataxia (ataxia grade ≥ 2), but poor agreement for normal horses or horses with low-grade deficits (ataxia grade ≤ 1); (4) raters who know a horse is suspected of having an abnormal gait (unrelated to lameness) assign a higher grade than the median ataxia grade.

Materials and Methods

The project was designed and conducted as a prospective cross-sectional reproducibility study according to the Guidelines for Reporting Reliability and Agreement Studies.[19] The study was approved by the Ethics and Welfare Committee of the Royal Veterinary College and where appropriate, was conducted under specific Home Office License according to the Animal (Scientific Procedures) Act (1986) of the United Kingdom. Please see Data S1 for the full details on methodology.

Raters

Six raters were recruited among clinicians at The Equine Referral Hospital (ERH) at the Royal Veterinary College (RVC), UK: 2 were board-certified internists (Large Animal; ACVIM) and 2 were board-certified surgeons (ECVS or ACVS). Two raters were second-year residents in either large animal internal medicine or equine surgery. Anonymity in scoring was maintained by assigning each rater a random, single digit number unknown to the author responsible for data entry and analysis.

Raters were not given any information about the horse's signalment, history, presentation or other clinical or clinicopathologic findings before the assessment of gait; however if a clinician had knowledge about the examined horse the rater was asked to disclose knowledge of the history or source of the horse and its reason for presentation.

Horses

Examinations were conducted between October 2010 and November 2012 at RVC ERH. Horses were recruited from 3 sources: Group 1 included horses with no known history of gait abnormalities that were purchased for an unrelated study of recurrent laryngeal neuropathy. Group 2 comprised horses referred to the ERH from first opinion practice for evaluation of gait deficits or perceived ataxia. Horses were recruited to Group 3 if a decision for euthanasia had been made in first opinion practice because of perceived moderate to severe ataxia. Horses that were considered on ethical grounds to be too ataxic to travel were excluded from Group 3. The horses were examined in order of presentation to the ERH.

Neurologic Examination

Every horse underwent a full and identical neurologic gait assessment that included walk and trot in a straight line, walking with the head elevated, walking with a blindfold, walking and standing tail pull, walking in small circles in both directions, backing up, lateral placement of distal thoracic limbs, crossing over of distal thoracic limbs, hopping on each thoracic limb, bilateral cervicofacial and panniculus reflexes, walking over an obstacle (10 × 20 cm pole), walking up and down a slope with and without head elevation. Raters completed a questionnaire for each step of the examination and were asked if the result was normal or abnormal as well as further characteristics (Table 2 and Data S2). The horses' gaits were graded according to a 5-point ataxia scale first proposed by Mayhew et al[11] and later modified by Reed[12] (Table 1). In a separate question, the horses were assessed and graded for lameness according to an 11-point scale.[7,20] The grades were assigned after walk and trot in a straight line and rescored with an overall grade, after the complete examination.

Video

All examinations were filmed with a video camera (1080 p, 50 Hz, shutter speed: 1/250 second[a]) with standardized recording and editing (further details in Data S1). A set of 10 videos (median grades: grade 0 [n = 3], grade 1 [n = 2], grade 2 [n = 2], grade 3 [n = 2], and grade 4 [n = 1]) were selected for test–retest reliability and edited. The assessment-reassessment study was based on scoring of these 10 videos on 2 separate occasions, making it

Table 1. The modified ataxia-grading scale used in this study. The text explaining each grade was printed on the questionnaire. No recumbent horses (grade 5) were included in the study.

Grade 0	No gait deficits at the walk
Grade 1	No gait deficits identified at the walk and deficits only identified during further testing
Grade 2	Deficits noted at the walk
Grade 3	Marked deficits noted at the walk
Grade 4	Severe deficits noted at the walk and may fall or nearly fall at normal gaits

a video-to-video comparison to reduce the impact of bias between live and video assessment.

Postmortem Examination and Histopathology

For all euthanized horses, the entire spinal cord was removed and fixed. Sections of the spinal cord were examined segmentally at the level of the dorsal root from C1-T2 with additional sections examined at T9, T16, and L5. A Diplomate of the European College of Veterinary Pathology with a specific interest in neuropathology (author KB) examined the transverse and longitudinal sections that were processed and stained with hematoxylin and eosin and Luxol fast blue. The pathologist was blinded to the signalment, history, clinical examination, and case number of the horse. Histopathologic findings consistent with pathology were recorded and assigned as the final diagnosis for each horse.

Statistics

Statistical analyses were performed using R[21] with the packages "lme4"[22] for mixed models, FactoMineR for PCA[23] and "ggplot2"[24] for graphics. Reliability for each question was calculated according to Equations 1, 2, and 3 (Data S1) by fitting a mixed effects model with rating as the repeated measure outcome variable to determine the random effects of rater, horse, and time of assessment (live, video1, video2).[25,26] The function "lmer" was used for dichotomous, ordinal, and continuous variables. Odds ratios (OR) were calculated and significance set using Fisher's exact test and the package "epicalc."[27] The intraclass correlation coefficient (ICC) was calculated as defined in formulas 1 to 4 (Data S1), all redefined.[4,28,29]

Agreement within assessors was measured as assessment-reassessment (test–retest, which is similar to intrarater assessment), with the assumption of no change in condition between ratings. The variation between assessments was primarily attributed to variation because of time of assessment[5,30] resulting in Equation 3 (Data S1).

The clinically important interpretation of the ICCs used to assess the quality of ratings was based on Cicchetti[31], where an ICC < 0.40 was poor, 0.40–0.59 was fair, 0.60–0.74 was good, and 0.75–1.0 was excellent agreement.

Expectation bias was assessed using a mixed effects model with the assigned grade as outcome and fixed effects of prior knowledge of the horse presenting with a gait deficit (unrelated to lameness) and random effects of horse and rater.

Multiple raters increase the reliability coefficient (Equation 2, Data S1). We also compared reliability between raters for horses divided into 2 subgroups of similar size. Group 1 had a median ataxia grade of ≤ 1 and Group 2 had a median ataxia grade of ≥ 2.

The correlation of separate parts of the gait assessment was investigated using principal component analysis, (PCA) (defined as construct validity in the field of clinimetrics). The PCA was performed on data derived from the questionnaire where the single rater ICC (Equation 1, Data S1) was >0.4, with the number of principal components determined from a scree plot.[32,33] ORs were calculated for the probability of positive answers in each of the questions correlating with either histopathologic changes consistent with pathology or of a horse being assigned a median ataxia grade ≥ 2.

Results

The number of raters assessing each horse ranged from 3 to 6 with a median of 5 assessors. The group of horses comprised 12 mares and 13 geldings, with a mean age of 6.4 years, ranging from 3 to 16 years. Of the 25 horses, 7 had a median ataxia grade of 0; 5 had a median ataxia grade of 1; 6 had a median grade of 2; 6 had a median grade of 3; and 1 horse was graded as 4 (Table 1 and Table S4).

The reliability for the modified ataxia-grading scale was good (ICC = 0.74) after the full neurologic gait assessment (Tables 2, 3). When the horses were split into 2 groups of either low ($\leq 1/5$, n = 12) or high grade ($\geq 2/5$, n = 13), the ICC for live scorings was 0.08 for the low-grade group and 0.43 for the higher grade group. When comparing agreement for the medicine group and the surgery group (Tables S2, S3), there is good and higher agreement for the surgery group on overall lameness scoring (ICC for medicine = 0.30 and ICC for surgery = 0.55) and a similar agreement on the overall ataxia scoring (ICC for both groups = 0.72). Clinicians often disagreed on grade of ataxia with the greatest disagreement over horses with a median grade of 2 (Fig 1).

The PCA revealed 2 principal components explaining 53% of the variation, with 41% of the first and 12% on the second dimension (Table 3). All gait-related assessments with an ICC ≥ 0.4 correlate significantly to the first principal component except "making mistakes over an obstacle," "deficits within individual limbs" and "hopping," all of which correlated with the second principal component (Table 4).

Sixteen horses were euthanized. Histopathologic changes consistent with pathology were detected in 7 ataxic horses: 1/3 horses with median ataxia grade 1, 4/6 horses with median ataxia grade of 2, and 2/3 horses with a median ataxia grade of 3. Histopathologic changes consistent with pathology identified in the ataxic horses ranged from classical Wallerian degeneration associated with compression (n = 6) to neuraxonal dystrophy and a dorsal root neurofibroma (findings are summarized in Table 4). None of the 3 horses euthanized with a median ataxia grade of 0 had histopathologic changes consistent with pathology. Histopathologic evidence of disease in the brain, cerebellum or spinal cord was not identified in 5 horses with an ataxia grade ≥ 1; these included 1/3 with a median ataxia grade of 1, 2/6 with a median ataxia grade of 2, 1/3 with a median ataxia grade of 3, and 1/1 with a median ataxia grade of 4.

Components of the neurologic gait assessment considered abnormal that had an increased OR for spinal cord pathology included "walking with the head elevated," "with a blindfold," "hopping," and "walking down a slope with the head elevated" (summarized in Table 4).

Raters who knew that a horse was presented for evaluation of a neurologic or abnormal gait problem (excluding lameness) were more likely to assign a grade above the median ataxia grade (P = .02, mixed effect model) with an OR of 2.8 (95% CI = 0.8–9.5).

Discussion

Reproducible assessment of gait during a neurologic examination in horses is essential for the diagnostic

Table 2. Reliability results in the form of ICC for the dichotomous and categorical questions for gait assessment during the neurologic examination of horses. See questionnaire (Data S2) for full details on the questions and Tables S1, S2 for comparison of live to video and average rater agreement.

	ICC[a]	
Examination Parts	Live[b]	Video1 : Video2[c]
Walk and trot on a straight line		
Normal/abnormal	**0.40**	0.43
Neuro deficit type	0.25	0.18
Head elevation		
Normal/abnormal	**0.56**	0.29
Makes mistakes	0.37	0.11
Paretic	0.35	0.31
Ataxic	**0.55**	0.17
Hypermetric	0.38	0.05
Hypometric	0.30	0.03
Blindfold		
Normal/abnormal	**0.42**	0.33
Readily identifiable	**0.60**	0.12
Standing tail pull		
Normal/abnormal	0.27	**0.47**
Walking tail pull		
Normal/abnormal	0.37	0.25
Small circles		
Normal/abnormal left	**0.48**	0.34
Normal/abnormal right	**0.44**	0.39
Circumducting	**0.45**	0.30
Turning normally FL	0.12	0.00
Turning normally HL	0.21	0.12
Backing up		
Normal/abnormal	**0.69**	**0.47**
Limb placement		
Normal/abnormal LF	**0.46**	**0.66**
Normal/abnormal RF	**0.46**	**0.60**
Obstacle		
Normal/abnormal	**0.47**	**0.48**
Makes mistakes	**0.68**	**0.67**
Mistakes are neurologic	**0.45**	0.07
Deficits LF	**0.56**	0.09
Deficits RF	**0.50**	0.37
Deficits LH	**0.52**	**0.70**
Deficits RH	0.39	0.27
Panniculus reflex		
Normal/abnormal left	0.37	**0.67**
Normal/abnormal right	0.24	**0.61**
Cervicofacial reflex		
Normal/abnormal left	0.37	**0.43**
Normal/abnormal right	0.33	**0.41**
Hopping		
Normal/abnormal	**0.45**	0.35
Weight shift to pelvic limbs	0.26	0.21
Stumble	**0.55**	0.34
Asymmetric	0.26	0.17
Slope		
Normal/abnormal	**0.42**	0.19
Mistakes	0.32	0.09
Paretic	0.37	0.34
Ataxic	**0.52**	0.17
Hypermetric	0.26	0.15
Slope head elevation		
Normal/abnormal	**0.44**	0.33
Mistakes	0.34	0.19
Paretic	0.36	0.34
Ataxic	**0.46**	0.21
Hypermetric	0.09	0.13

(continued)

Table 2 (Continued)

ICC, intraclass correlation coefficient; LF, left thoracic limb; RF, right thoracic limb; LH, left pelvic limb; RH, left pelvic limb. **Bolded** numbers have an ICC $\geq$ 0.4 (fair agreement).

[a] ICC for a single rater (Equation 1, ICC1[A,1]).

[b] ICC live scoring only, single rater (Equation 1).

[c] ICC test–retest (video only), from 1st video session to 2nd video session (Equation 3).

Table 3. Reliability results as ICC for the questions with answers on an ordinal scale. The ICCs are calculated from the gait assessment in the neurologic examination of horses. See questionnaire (Data S2) for full details on the questions and Tables S1, S2 for comparison of live to video and average rater agreement.

	ICC[a]	
Examination Parts	Live[b]	Video1 : Video2[c]
Walk and trot on a straight line		
Lame or neurologic	**0.21**	0.09
Lame leg	0.49	0.35
Ataxia grade	**0.71**	**0.48**
Lameness grade	0.29	0.39
Across all		
Ataxia score LF	0.39	0.28
Ataxia score LH	**0.69**	**0.54**
Ataxia score RF	0.37	0.28
Ataxia score RH	**0.60**	**0.42**
Paresis score LF	0.07	0.18
Paresis score LH	**0.54**	**0.57**
Paresis score RF	0.00	0.12
Paresis score RH	0.30	0.36
Overall lameness grade	0.26	0.39
Overall ataxia grade	**0.74**	**0.59**

ICC, intraclass correlation coefficient. **Bolded** numbers have an ICC1 $\geq$ 0.4 (fair agreement).

[a] ICC for a single rater (Equation 1, ICC1[A,1]).

[b] ICC live scoring only, single rater (Equation 1).

[c] ICC test–retest (video only), from 1st video session to 2nd video session (Equation 3).

process and for decision making in prepurchase examinations, for considering treatment options, safety for handlers and riders and animal euthanasia. The most common neurologic diseases of horses affect the spinal cord, with resultant changes in gait caused by general proprioceptive deficits and paresis. In this study, the reliability for each part of the gait assessment ranged from poor to good; however, the only individual

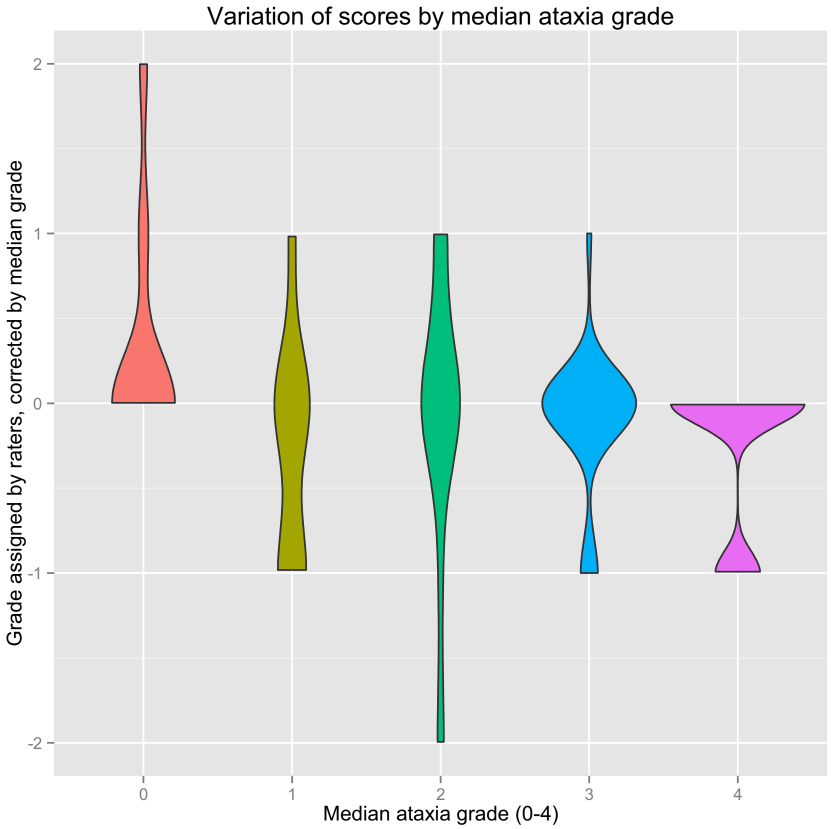

Fig 1. Violin plot of the variation in individual ratings grouped by the median rating for each horse during live scoring only. To align the ratings around 0, each score was subtracted from the median score of the horse. A violin plot is similar to a boxplot, with the addition that the density of data points is illustrated by an increase in width. This figure reveals that most grades have a fluctuation of 1 degree more or less than the median; however, grades 0 and 3 are condensed around the median illustrating better agreement, whereas grade 2 stretches from −2 to +1 grades from the median.

elements with an ICC > 0.6 were "backing up," recognition of a horse "making mistakes over an obstacle," overall ataxia score and pelvic limb ataxia score. The fair to poor agreement between raters and for assessment-reassessment has important implications for the daily clinical assessment of individual horses as well as follow-up examinations. In addition, there are implications for the assessment of neurologic gait deficits in research studies evaluating equine neurologic diseases and their treatments.

A perfect rating scale should be reliable, valid, responsive[5,34,35] and enable identification of the clinically minimal important difference (CMID) between 2 treatment groups or in a change over time.[36,37] Indeed, for decision making for individual patients, some authors recommend that the ICC should be at least 0.90.[19] Based on these recommendations, neither the individual criteria for gait assessment within the equine neurologic examination, nor the modified ataxia-grading scale itself is acceptable for clinical use.

We found a higher overall agreement for live scoring of ataxia compared to previous studies assessing lameness scoring in a live setting using a 6-point scale (AAEP).[8] In normal horses or those with subtle ataxia, agreement on the ataxia grading was worse than found in horses with low-grade lameness.[8] Similarly, there was worse agreement between raters for their assessment of the moderate to high-grade ataxia group, compared to assessment of horses with higher grades of lameness. The relatively lower agreement for the moderate to high-grade ataxia could be explained by the large variation between raters when assessing horses with a median ataxia grade of 2 (Fig 1). We also found a more pronounced disagreement for the assessment of ataxia in horses compared to a new scale applied to assess dogs with spinal cord injury using the Texas Spinal Cord Injury Score (TSCIS) with separate components of gait, proprioceptive positioning and nociception.[38] Two blinded raters had excellent agreement (ranging from 0.72 to 1.00) across all criteria when rating 36 dogs.[38] However, the dogs used in the TSCIS study all had spinal cord dysfunction ranging from mild to severe and the study did not include any unaffected controls. Conceivably, some of the variation between raters in this study might have resulted from the confounding factor of concurrent musculoskeletal disease, as several horses in this study were considered to be lame. Although previous studies of agreement on lameness did not include assessment of ataxia, ideally all horses in this study would have also received a complete lameness investigation and diagnostic analgesia. Unfortunately, this was beyond the scope of our work, but it highlights the

Table 4. OR and PCA results for questions with an ICC > 0.4. OR results for a positive test with evidence of spinal cord pathology. The PCA shows correlation with each question of the 2 dimensions (D1 and D2).

	PCA		OR + Pathology[a]	
Examination Parts	OR	95% CI	Corr D1[b]	Corr D2[b]
Walk and trot on a straight line				
Normal/abnormal	0.0	0.00–52	**0.82**[e]	0.07
Head elevation				
Normal/abnormal	∞[f]	0.65–∞	**0.91**[e]	−0.17
Ataxic	7.0	0.33–417	**0.67**[e]	0.07
Blindfold				
Normal/abnormal	∞	0.26–∞	**0.78**[e]	−0.36
Readily identifiable[c]	0.8	0.05–12	0.20	−0.08
Small circles				
Normal/abnormal left	1.5	0.08–28	**0.83**[e]	0.18
Normal/abnormal right	7.0	0.33–417	**0.94**[e]	0.08
Circumduction	5.0	0.28–294	**0.73**[e]	0.26
Backing up				
Normal/abnormal	3.0	0.15–188	**0.70**[e]	0.05
Limb placement				
Normal/abnormal LF	2.0	0.15–33	0.26	−0.17[d]
Normal/abnormal RF	35.0[g]	1.2–1844	**0.53**[e]	**−0.60**
Obstacle				
Normal/abnormal	1.7	0.12–23	**0.84**[e]	0.03
Makes mistakes	1.5	0.08–29	0.42[e]	**0.61**[d]
Mistakes are neurologic[d]	9.0	0.35–546	**0.79**[e]	−0.18
Deficits LF	0.1	0–3	−0.20	**0.52**[d]
Deficits RF	1.1	1.1–22	−0.09	0.31
Deficits LH	0.3	0.01–7	0.48[e]	**0.62**[d]
Deficits RH	0.9	0.05–15	0.22	0.33
Hopping				
Normal/abnormal	∞[f]	0.91–∞	0.41[e]	−0.49[d]
Stumble	9	0.35–546	0.38	**−0.57**[d]
Slope				
Normal/abnormal	1.5	0.08–29	**0.83**[e]	−0.23
Ataxic	1.5	0.08–29	**0.84**[e]	0.36
Slope with head elevation				
Normal/abnormal	∞	0.26–∞	**0.78**[e]	−0.31
Ataxic	1.5	0.08–29	**0.65**[e]	0.17

CI, confidence interval; OR, odds ratio; PCA, principal component analysis; ∞, the OR is infinitely high. Bolded numbers represent questions with a correlation of 0.5 or higher with that dimension of the PCA.

[a] OR calculation where disease is considered histopathologic evidence of spinal pathology and exposure is a positive test during live assessment.

[b] Correlation of the question with the PCA derived first dimension (D1, 2) and second dimension (D2, 3).

[c] If abnormal, is the deficit readily identifiable? Yes or No answer.

[d] If making mistakes, the mistakes are likely to be a neurologic deficit?

[e] Significant correlation with that dimension.

[f] Significant OR on Fisher's exact test.

importance of establishing better standards for gait assessment in horses with gait deficits caused by neurologic or musculoskeletal disease (or both).

In human medicine, clinicians developed the scale for the assessment and rating of ataxia (SARA) based on the neurologic examination. SARA has 1 underlying construct that explains 80% of the variation with 8 separate criteria, all scored on ordinal scales (gait, stance, sitting, speech disturbance, finger chase, nose-finger test, fast alternating hand movements, and heel-shin slide). SARA has high inter-rater reliability ICC (0.98) and a test–retest reliability ICC of 0.90.[39] The equine modified ataxia scale in this study performed comparably to subscales of the ICARS, but poorly in comparison with the overall excellent agreement and reliability of SARA.[39] However, only neurologists or senior neurology residents assessed the patients in the SARA work and it was conducted on greater numbers of subjects, meaning that a direct comparison with this study might be misleading.

A video-based study using an 11-point lameness scale revealed fair reliability ($\kappa = 0.41$) for agreement between 3 raters scoring lameness based on video (ranging from 0.30 to 0.58 for each pair of raters).[40] Previous studies into agreement on assessment of lameness on the 6-point scale[8,40,41] have a poorer agreement because they are video-based. In a study of objective kinematic assessment of ataxic horses,[15] poor agreement between live and video scoring was attributed to the different conditions for live compared to video assessment; however, our results suggest that it was more likely caused by poor live agreement and low test–retest reliability.

For testing hypotheses, clinical signs should be measurable as either dichotomous (absent or present), categorical (such as absent, mild, moderate or severe) or as ordinal (scaled) variables. Scales with multiple, ordinal divisions are more sensitive to change than dichotomous scales[35,37] and their use has higher reliability because of reduced random error.[32] Furthermore, a scale's validity is better assessed and improved when scores are assigned within multi-item scales.[32] We therefore recommend that a group of experts gather to discuss and refine the observations of horses with neurologic gait deficits[30] based on a series of standardized videos of horses with confirmed spinal cord disease. A multi-item scale with ordinal ratings and with simplified and reliable descriptions of clinical observations of gait could be the aim of such an expert panel and examination components in this study with relatively higher reproducibility could be used as a foundation. The aim would be a scale with improved sensitivity to change and sufficient detection of CMID in order to help discriminate the severity of neurologic gaits for determination of response to treatments and for informed decisions regarding prognosis and horse and rider safety.

We confirm a likely expectation bias in the gait assessment component of the neurologic examination, since raters who were aware that a horse was presented for an abnormal gait (excluding lameness) or a possible neurologic gait deficit, were more likely to assign an ataxia grade higher than the median grade. This finding is comparable to the expectation bias reported in lameness evaluation.[7] Another source of bias is experience and training,[42] where experts in lameness assessment are more consistent in their scoring than residents, interns, and students.[7,43] Excluding residents from our analysis did not improve the ICCs of the live assessment. In addition, the difference between reliability for medicine and surgery was minimal, although surgeons had a higher reliability for assessing lameness compared to the medicine group (Tables S3, S4).

In this study, the median score of all raters was considered as a horse's true grade, and the variation across all raters was examined around this score (Fig 1). This system assumes that all raters have similar ability in identification of neurologic gait deficits, but this might not be true as experience likely varies considerably. Optimally, the raters' scores would have been compared to a reference ("gold") standard for disease severity and presence or absence of disease. We attempted this by examining rater ICCs for horses with and without histopathologic changes consistent with pathology. Spinal cord pathology was commonly associated with a perception of horses' being abnormal either when "walking with the head elevated," "blindfolded," "hopping," or when "walking on a slope", though it was not possible to quantify or compare disease severity by histopathologic examination of the spinal cord. However, the low number of horses in this study reduces the power of these conclusions and the lack of pathologic changes in 5 of 12 euthanized horses with a median ataxia grade ≥ 1 limited our ability to evaluate fully the neurologic examination's validity. Nonetheless, our results emphasize the not uncommon disparity between clinician and pathologist when assessing horses with neurologic gait deficits. Furthermore, histopathologic assessment suffers from similar caveats as clinical assessment with variation or error introduced by experience level of the pathologist or tissue artifacts. For example, incidental background findings (spheroids) are found at all stages and at all levels of the neuroaxis in horses without neurologic disease.[44] In addition, assessment of the importance of neuropathologic changes can best be made when the pathologist is aware of the history and clinical signs,[45] presumably by increasing pretest probability and to maximize the chances of sampling the affected areas. As such, the pathologist's opinion of the significance of histopathologic changes in the context of a history or clinical signs is akin to a clinician's use of history and other clinical information in assessing the significance of perceived deficits detected during subjective (neurologic) examination (ie, expectation bias being a positive discriminatory factor). Furthermore, the assumption that all horses with apparent neurologic gait deficits should have identifiable pathologic changes might well be flawed as dynamic functional deficits might occur with only intermittent spinal compression, as is believed to occur in humans.[46] In addition, recent identification of an important genetic component in controlling gait in horses[47] reveals the extent to which gait alterations can have a functional rather than pathologic basis.

We conclude that clinicians should be aware of poor agreement between skilled and experienced observers of gait abnormalities in horses and poor correlation between pathology and clinical signs. The agreement is worse when signs are mild, and clinicians should be cautious when making decisions about horses on the basis of a subjective assessment of gait during the neurologic examination, especially when signs are subtle. This is particularly important during prepurchase examination and when a decision might lead to euthanasia or retirement or is being made for insurance purposes. Similarly, clinicians should be cautious when drawing conclusions from an apparent change in a horse's degree of neurologic compromise after management changes or treatments, particularly given the anamnesis bias we report. We suggest that the neurologic assessment of horses' gaits could be improved by identification of a set of objective parameters that can quantify severity of ataxia in horses, ideally in a clinical setting.

Footnote

[a] HDC TM-700; Panasonic, Osaka, Japan

Acknowledgments

We thank a number of funding bodies for financial support: Hesteafgiftsfonden, The RCVS Trust, Kustos of 1881, Oticon, University of Copenhagen, Research School KLINIK, Department of Large Animal Sciences, and Danish Agency for Science, Technology and Innovation. We acknowledge Yu-Mei Change for statistical advice, Victoria E. Unt, Luis Lamas, Laura Tulloch, and Anna Liedtke for help with horses during examinations, Sandra D. Starke and Thilo Pfau for help with video recordings, Jenny K. Boye for photography, and Malcolm Morley for helpful discussions. In addition, students and technicians at the ERH are acknowledged for the invaluable help with horses during the study. Finally, we acknowledge the veterinary surgeons who referred cases and horse owners who consented to their horses' inclusion in the study. This manuscript has been approved by the Royal Veterinary College and assigned manuscript number CSS-00604.

Conflict of Interest Declaration: The authors disclose no conflict of interest.

References

1. Sackett DL, Rennie D. The science of the art of the clinical examination. JAMA 1992;267:2650–2652.

2. Yen K, Karpas A, Pinkerton HJ, et al. Interexaminer reliability in physical examination of pediatric patients with abdominal pain. Arch Pediatr Adolesc Med 2005;159:373–376.

3. Reilly BM. Physical examination in the care of medical inpatients: An observational study. Lancet 2003;362:1100–1105.

4. de Vet HC, Terwee CB, Knol DL, et al. When to use agreement versus reliability measures. J Clin Epidemiol 2006;59:1033–1039.

5. McAlister FA, Straus SE, Sackett DL. Why we need large, simple studies of the clinical examination: The problem and a proposed solution. CARE-COAD1 group. Clinical Assessment of the Reliability of the Examination-Chronic Obstructive Airways Disease Group. Lancet 1999;354:1721–1724.

6. Joshua AM, Celermajer DS, Stockler MR. Beauty is in the eye of the examiner: Reaching agreement about physical signs and their value. Intern Med J 2005;35:178–187.

7. Arkell M, Archer RM, Guitian FJ, et al. Evidence of bias affecting the interpretation of the results of local anaesthetic nerve blocks when assessing lameness in horses. Vet Rec 2006;159:346–349.

8. Keegan KG, Dent EV, Wilson DA, et al. Repeatability of subjective evaluation of lameness in horses. Equine Vet J 2010;42:92–97.

9. van Biervliet J, Scrivani PV, Divers TJ, et al. Evaluation of decision criteria for detection of spinal cord compression based on cervical myelography in horses: 38 cases (1981–2001). Equine Vet J 2004;36:14–20.

10. Levine JM, Scrivani PV, Divers TJ, et al. Multicenter case-control study of signalment, diagnostic features, and outcome associated with cervical vertebral malformation-malarticulation in horses. J Am Vet Med Assoc 2010;237:812–822.

11. Mayhew IG, deLahunta A, Whitlock RH, et al. Spinal cord disease in the horse. Cornell Vet 1978;68(Suppl 6):1–207.

12. Reed SM. Neurologic exam. J Equine Vet Sci 2003;23:484–492.

13. Nout YS, Reed SM. Cervical vertebral stenotic myelopathy. Equine Vet Educ 2003;15:212–223.

14. Keegan KG, Arafat S, Skubic M, et al. Detection of spinal ataxia in horses using fuzzy clustering of body position uncertainty. Equine Vet J 2004;36:712–717.

15. Strobach A, Kotschwar A, Mayhew IG, et al. Gait pattern of the ataxic horse compared to sedated and nonsedated horses. Equine Vet J Suppl 2006;36:423–426.

16. Lahunta Ad, Glass EN. Veterinary Neuroanatomy and Clinical Neurology, 3rd ed. St. Louis, MO: W.B. Saunders Company; 2009:552.

17. Ishihara A, Reed SM, Rajala-Schultz PJ, et al. Use of kinetic gait analysis for detection, quantification, and differentiation of hind limb lameness and spinal ataxia in horses. J Am Vet Med Assoc 2009;234:644–651.

18. Hoffman CJ, Clark CK. Prognosis for racing with conservative management of cervical vertebral malformation in Thoroughbreds: 103 cases (2002–2010). J Vet Intern Med 2013;27:317–323.

19. Kottner J, Audige L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol 2011;64:96–106.

20. Wyn-Jones G. The diagnosis of the causes of lameness. In: May SA, ed. Equine Lameness. Oxford UK: Blackwell Scientific Publications; 1988;3:5–6; 6–8.

21. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2013. Available at: http://www.R-project.org. Accessed April 1, 2013.

22. Bates D, Maechler M, Bolker B. lme4: Linear Mixed-Effects Models Using S4 Classes. R Package; 2012. Available at: http://CRAN.R-project.org. Accessed April 1, 2013.

23. Husson F, Josse J, Le S, et al. FactoMineR: Multivariate exploratory data analysis and data mining with R. J Stat Soft 2008;25:1:18.

24. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York, NY: Springer; 2009.

25. Shrout PE, Fleiss JL. Intraclass correlations—Uses in assessing rater reliability. Psychol Bull 1979;86:420–428.

26. Nakagawa S, Schielzeth H. Repeatability for Gaussian and non-Gaussian data: A practical guide for biologists. Biol Rev Camb Philos Soc 2010;85:935–956.

27. Chongsuvivatwong V. epicalc: Epidemiological Calculator. R Package; 2012. Available at: http://CRAN.R-project.org. Accessed 04/01/2013.

28. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 1973;33:613–619.

29. Hallgren KA. Computing inter-rater reliability for observational data: An overview and tutorial. Tutor Quant Methods Psychol 2012;8:23–34.

30. Keszei AP, Novak M, Streiner DL. Introduction to health measurement scales. J Psychosom Res 2010;68:319–323.

31. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 1994;6:284.

32. Hobart JC, Cano SJ, Zajicek JP, et al. Rating scales as outcome measures for clinical trials in neurology: Problems, solutions, and recommendations. Lancet Neurol 2007;6: 1094–1105.

33. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. Qual Life Res 2010;19:539–549.

34. Hayes G, Mathews K, Kruth S, et al. Illness severity scores in veterinary medicine: What can we learn? J Vet Intern Med 2010;24:457–466.

35. Martinez-Martin P. Composite rating scales. J Neurol Sci 2010;289:7–11.

36. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407–415.

37. Saute JA, Donis KC, Serrano-Munuera C, et al. Ataxia rating scales—Psychometric profiles, natural history and their application in clinical trials. Cerebellum 2012;11:488–504.

38. Levine GJ, Levine JM, Budke CM, et al. Description and repeatability of a newly developed spinal cord injury scale for dogs. Prev Vet Med 2009;89:121–127.

39. Schmitz-Hubsch T, du Montcel ST, Baliko L, et al. Scale for the assessment and rating of ataxia: Development of a new clinical scale. Neurology 2006;66:1717–1720.

40. Fuller CJ, Bladon BM, Driver AJ, et al. The intra- and inter-assessor reliability of measurement of functional outcome by lameness scoring in horses. Vet J 2006;171:281–286.

41. Hewetson M, Christley RM, Hunt ID, et al. Investigations of the reliability of observational gait analysis for the assessment of lameness in horses. Vet Rec 2006;158:852–857.

42. Drager LF, Abe JM, Martins MA, et al. Impact of clinical experience on quantification of clinical signs at physical examination. J Intern Med 2003;254:257–263.

43. Keegan KG, Wilson DA, Wilson DJ, et al. Evaluation of mild lameness in horses trotting on a treadmill by clinicians and interns or residents and correlation of their assessments with kinematic gait analysis. Am J Vet Res 1998;59:1370–1377.

44. Summers BA, Cummings JF, De Lahunta A. Veterinary Neuropathology. St. Louia, MO: Mosby; 1995;50–51, 189–198.

45. Jahns H, Callanan JJ, McElroy MC, et al. Age-related and non-age-related changes in 100 surveyed horse brains. Vet Pathol 2006;43:740–750.

46. Zhang L, Zeitoun D, Rangel A, et al. Preoperative evaluation of the cervical spondylotic myelopathy with flexion-extension magnetic resonance imaging: About a prospective study of fifty patients. Spine (Phila Pa 1976) 2011;36:E1134–E1139.

47. Andersson LS, Larhammar M, Memic F, et al. Mutations in DMRT3 affect locomotion in horses and spinal circuit function in mice. Nature 2012;488:642–646.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1. Materials and Methods.

Data S2. Rater Questionnaire.

Table S1. Reliability results in the form of intraclass correlation coefficient for the dichotomous and categorical questions for gait assessment during the equine neurological examination assessed by four to six expert raters. See questionnaire (Data S2) for full details on the questions.

Table S2. Reliability results as intraclass correlation coefficient (ICC) for the questions with answers on an ordinal scale. The ICCs are calculated from the gait assessment in the equine neurological examination assessed by four to six expert raters during live sessions only. See questionnaire (Data S2) for full details of the questions.

Table S3. Overview of the horses and their median ataxia grade, median lameness grade and conclusions from the spinal cord histopathology.

Table S4. Level of clinical significance of the intraclass correlation coefficient ranges for agreement, after Cicchetti.[31]

Table S5. Reliability results as intraclass correlation coefficient for the dichotomous and categorical questions in gait assessment in the equine neurological examination assessed by four to six expert raters during live sessions. See (Data S2) for full details of the questions.

Table S6. Reliability results as intraclass correlation coefficient (ICC) for the questions with answers on an ordinal scale. The ICCs are calculated from the assessment of gait in the equine neurological examination assessed by four to six expert raters. See questionnaire (Data S2) for full details of the questions.