Equine Veterinary Journal

Equine Veterinary Journal ISSN 0425-1644
DOI: 10.1111/evj.13144

Agreement among equine veterinarians and between equine veterinarians and inertial sensor system during clinical examination of hindlimb lameness in horses

P. LEELAMANKONG[†*], R. ESTRADA[†‡], K. MÄHLMANN[†], P. RUNGSRI[†§] and C. LISCHER[†]

[†]Department of Veterinary Medicine, Equine Clinic, Freie Universität Berlin, Berlin, Germany
[‡]Large Animal Hospital, School of Veterinary Medicine, National University, Heredia, Costa Rica
[§]Department of Companion Animal and Wildlife Clinic, Faculty of Veterinary Medicine, Chiang Mai University, Chiang Mai, Thailand.

*Correspondence email: p_leelamankong@hotmail.com; Received: 22.10.18; Accepted: 15.06.19

Summary

Background: Hindlimb lameness evaluation is known to be challenging. Experience is essential for the ability of equine veterinarians to detect lameness. Nevertheless, even an experienced veterinarian is still subject to bias. Objective lameness detecting methods have been established to aid veterinarians.

Objectives: 1) To estimate the effect of experience on the interobserver agreement and the agreement between a body-mounted inertial sensor system (BMISS) and veterinarians on detecting hindlimb lameness, and 2) to estimate the agreement between the BMISS and highly experienced veterinarians on change in lameness after diagnostic analgesia.

Study design: Cross-sectional study.

Methods: Twenty-six horses with hindlimb lameness were evaluated in clinical conditions by clinicians and simultaneously measured by the BMISS. Videos of their lameness examination were recorded and shown to 13 veterinarians from three groups of varying experience for evaluation. The interobserver agreement and the agreement between veterinarians and the BMISS were calculated.

Results: Interobserver agreement from all three groups was recorded as 'fair'. The strength of agreement between veterinarians and BMISS was 'fair' for the highly experienced group, 'slight to fair' for the moderately experienced group and 'slight' in the inexperienced group. The BMISS and the highly experienced veterinarians declared a 'strong' agreement in assigning an improvement in lameness after diagnostic analgesia.

Main limitations: Lameness evaluation through video viewing might be more challenging for some evaluators than live situations.

Conclusions: Given the task of evaluating videos of horses trotting in a straight line, the more experienced veterinarians did not show more reliability than those with less experience. Due to 1) the moderate agreement between the BMISS and clinicians (highly experienced and moderately experienced) in the live clinical evaluation in determining hindlimb lameness, and 2) the strong association between the BMISS and highly experienced veterinarians in determining improvement of lameness after anaesthesia, therefore the use of the BMISS as a supporting tool for veterinarians is encouraged.

Keywords: horse; lameness; interobserver; agreement; inertial sensors

Introduction

Examining lame horses while they are in motion is a crucial part in forming a diagnosis of the cause of lameness [1,2]. The ability to accurately recognise movement patterns of lame horses require practice and experience [3–5]. Pelvic movement has been described as the most significant criterion in recognising hindlimb lameness. Several text books use terms like 'hip hike', 'hip drop' or 'gluteal rise' and which phase of trotting they should be aware of in order to detect hindlimb lameness [1,2,6,7]. Since there are no universal rules for identifying lameness, clinicians may use different ways of viewing hindlimb lameness according to the conception of the individual's cognitive powers and perception. Even among experienced veterinarians, agreement during hindlimb lameness evaluation was reported to be only acceptable [8]. This leads to potential inaccuracy during subjective assessments of diagnostic analgesia. A study using verbal scores of change of lameness after diagnostic analgesia resulted in high agreement between veterinarians [9]. Controversially, another study reported that veterinarians can be biased towards a positive result of the diagnostic analgesia when they were aware which one had been performed [10].

Several objective methods to aid lameness evaluation have been developed during the past decades. Kinetic and kinematic methods such as force plate or video-assisted motion analysis techniques are shown to be very precise and accurate [11–13], yet expensive and not suitable for clinical practice. Therefore, alternative methods using inertial sensors attached to different anatomical structures of a horse have been considered [14–16]. A user-friendly inertial sensor-based method, measuring asymmetry of head and pelvic movement, was noted for having a high correlation to the video-based motion analysis system and having sufficient repeatability in lameness evaluations [15,17]. This method has been tested to be able to detect change in lameness after diagnostic analgesia and flexion tests, when correlated with the opinion of experienced veterinarians [18,19].

The objectives of this study were 1) to estimate the interobserver agreement on detecting hindlimb lameness by veterinarians of different experience levels, 2) to estimate the agreement between an inertial sensor system and veterinarians from different experience levels on detecting hindlimb lameness and 3) to estimate the agreement between the body-mounted inertial sensor system, BMISS (Lameness Locator®)[a] and the highly experienced veterinarians on changes in lameness after diagnostic analgesia. We hypothesised that 1) both the interobserver agreement and the agreement between veterinarians and the BMISS from the highly experienced group would be higher compared to the moderate experienced and the inexperienced groups of veterinarians and 2) the agreements for improvement after diagnostic analgesia by highly experienced clinicians would be high.

Materials and methods

Study design

Horses: Between July 2012 and February 2013, 26 horses were presented to the Equine Clinic, Free University of Berlin, for an evaluation of hindlimb lameness. Selection criteria were 1) mild to moderate hindlimb lameness (grade 1–3 from Ross 2011 [1], Supplementary Item 1) when the horse was evaluated trotting in hand in a straight line on a hard surface, 2) improvement in lameness after diagnostic analgesia was determined by clinician or the BMISS and 3) owner permission for the use of the BMISS and video records in the study. Horses ranged from 5- to 18-year-old (mean = 10 years). Breeds included 18 Warmbloods, five Trotters, one American Quarter horse and two others.

Clinical evaluation of lameness and diagnostic analgesia: A trained groom led each horse along a 30-m long concrete surface. To enable the 25 contiguous strides recommended by the inertial sensor system to take place, the horse was trotted up and down the trotting surface twice. Two clinicians took part in the lameness evaluation. The first clinician performed a routine lameness examination and was in charge of decision-making. Both clinicians were asked to grade lameness on scale of 0 to 5 before and after each analgesia, while unaware of each other's opinion and the BMISS data. They were also asked to assess the improvement in lameness after each diagnostic analgesia using score of 1 to 6, with a score of 1 showing no improvement, 2 an improvement of less than 50%, 3 an improvement of more than 50%, 4 an improvement with residual lameness, 5 the lameness abolished and 6 lameness abolished and switched to the contralateral limb. In total, there were seven clinicians who participated in the examination. Four were classified as highly experienced veterinarians with experience in equine orthopaedics ranging from 8 to 25 years and three were classified as moderately experienced veterinarians with experience ranging from 1 to 3 years. For the 26 horses, the role of first clinician (the clinician who performed the decision of lameness examination) was carried out by the highly experienced vets in 16 horses and by the moderately experienced vets in 10 horses. The role of second clinician (who only observed the examination) was carried out by the highly experienced veterinarians in four horses and by moderately experienced vets in 22 horses.

All 26 horses underwent low 6-point block on the first day. In 21 horses, the deep branch of lateral plantar nerve block was consequently performed. In three horses, the tibial and fibular nerves were blocked to determine if further improvement could be observed.

Eighteen horses received lameness evaluations on the second day, whereupon synovial analgesia was performed based on the results of the regional anaesthesia from the first day.

The body-mounted inertial sensor system and data collection: The BMISS consists of three inertial sensors; one uni-axial accelerometer attached to the poll region, a second uni-axial accelerometer affixed to the most dorsal aspect of the pelvis between tuber sacrale and a gyroscope attached to the dorsal aspect of the right forelimb pastern.

The gyroscope detected the right forelimb stance phase and the position of the other three limbs was determined in relation to right forelimb position. The head and pelvis acceleration, measured by poll and pelvic sensors, was converted through sophisticated proprietary algorithms to their relative position throughout the stride cycle, and thereafter calculated into head and pelvic height differences (in millimetres) between right and left halves of stride.

The maximum and minimum head and pelvic height differences were calculated for each stride and reported as mean values for all of the strides in the measuring trial with their standard deviations (in mm). For this study, only the mean maximum pelvic height difference (P_{max}) and the mean minimum pelvic height difference (P_{min}) were the data of interest. In a conceptually symmetric horse, both P_{max} and P_{min} were 0 mm. A positive P_{max} value indicated decreased upward pelvic movement of the right hindlimb after push-off, and a positive P_{min} value indicated decreased downward movement during the right hindlimb stance. A negative value of P_{max} and P_{min} indicated the same instances for the left hindlimb. The threshold for both P_{max} and P_{min} between sound and lame was defined as ± 3 mm [20].

A non-random change at a 95% confidence interval of P_{max} and P_{min} between two measurements has been determined to be 3 mm [15]. Therefore, only changes in P_{max} and P_{min} after diagnostic analgesia compared to a baseline of 3 mm were used to calculate percent improvement in hindlimb lameness as follows;

(P_{max} baseline−P_{max} after analgesia)/(P_{max} baseline−P_{max} threshold) × 100, and (P_{min} baseline−P_{min} after analgesia)/(P_{min} baseline−P_{min} threshold) × 100.

A 100% improvement after local diagnostic analgesia was assigned when both P_{max} and P_{min} after block were under the threshold between sound and lame.

Video collection and evaluation: All straight trotting trials were recorded with a digital HD video camera[b], which was placed on a tripod approximately 2 m behind the horse before the start. At the beginning of each video, the horse appeared in the video frame from the top of the rear down to the hocks. As the horse trotted away, the whole horse could be seen down to the hooves after two to three steps. From then on, zooming in and out was manually optimised to maintain the proportion of the horse in the frame while the horse was trotting away from and towards the video camera. The recordings were transferred to a computer and video test units were made with video-editing software (Windows Live Movie Maker 2012)[c]. Each video test unit consisted of a 'baseline trial' and a 'corresponding after blocking trial' and was about 3 minutes long. Sound was not played. This resulted in a total of 99 video test units. To reduce bias towards choosing just one of the hindlimbs as the lame limb, they were tested together with 101 other video test units of horses with forelimb lameness (data not shown). The 200 video test units were randomly divided into 20 sessions, with each session consisting of 10 video test units.

Thirteen veterinarians from the Equine Clinic of the Freie Universität, Berlin, including those who performed the clinical examination, took part in evaluating these videos. They were categorised into three groups based on years of experience in equine lameness evaluations. There were four highly experienced veterinarians with experience ranging from 8 to 25 years, four moderately experienced veterinarians with 1 to 3 years of experience, and five inexperienced interns with less than 1 year of experience. The video evaluation was carried out at least 2 months after the clinical examination to prevent the clinicians who had seen the horses in live situations from being unduly influenced by their own recent memories. Each individual was only permitted to view one session at a time (about 30 min long) to avoid fatigue. For the evaluation of each video test unit, the baseline trial was played twice and the video was paused to enable the veterinarian to note a lameness score in the same manner as in a clinical situation. Then the after blocking trials (without suggesting which limb was blocked) were also displayed twice. The veterinarians evaluated the lameness score again and assessed the improvement in lameness after diagnostic analgesia using a score of 1–6 (to the limb believed to be blocked). All evaluators were deliberately kept unaware of the prior clinical exam results, inertial sensor system results and the results from the other veterinarians.

Data analysis

A statistical analysis was performed using RStudio version 1.1.419[d]. Subjective lameness scores and BMISS measurements were classified into four different categories based on hindlimb lameness; 1) left hindlimb lameness, 2) right hindlimb lameness, 3) bilateral hindlimb lameness and 4) no hindlimb lameness (Supplementary Item 2). A total of 26 day 1 baseline trials from each horse were used to analyse the interobserver agreement and the agreement between subjective and objective evaluation.

Interobserver agreement and the agreement between subjective and objective evaluation: For interobserver agreement, Fleiss' Kappa (κ) statistics were used to estimate the agreement on hindlimb lameness category between two veterinarians in clinical situations and among each experience group during video evaluation. For the agreement between subjective and objective evaluation, Fleiss' Kappa (κ) statistics were also used to estimate the agreement on hindlimb lameness category between the objective method and each veterinarian. The Landis and Koch benchmark scale was used to estimate the strength of agreement, with $\kappa<0$ representing a poor agreement, $0.0<\kappa<0.20$ a slight agreement, $0.21<\kappa<0.40$ a fair agreement, $0.41<\kappa<0.60$ a moderate agreement, $0.61<\kappa<0.80$ a substantial agreement and $0.81<\kappa<1.00$ an almost perfect agreement.

Effect of lameness severity on interobserver agreement, and the agreement between subjective and objective: The 26-day 1 baseline trials were subsequently divided equally into two severity groups based on the amplitude of Pmax and Pmin (13 baseline trials per severity group). Then, the interobserver agreement and the agreement between subjective veterinarians and objective BMISS were estimated for each severity group.

Agreement on response to diagnostic analgesia between subjective and objective evaluation: A test unit was selected for a response evaluation when all highly experienced veterinarians and the BMISS indicated the same lame limb at baseline as the limb which was currently blocked. The response to the anaesthesia score from the highly experienced group and objective evaluation are classified in five categories; 1 = no improvement, 2 = improvement <50%, 3 = improvement >50%, 4 = lameness eliminated, 5 lameness eliminated and switched to contralateral hindlimb (Supplementary Item 3).

The strength and direction of association between objective and subjective response categories was estimated by a calculation of Kendall's Tau-test (T_b) rank correlation test. Subjective and objective evaluations were considered to be in agreement when the response category matched exactly. Kendall's Tau-test ranges from 1 to −1, with 1 indicating a perfect positive relationship and −1 indicating a perfect negative relationship. The Strength of agreement is described using a benchmark scale of I T_b I<0.1, indicating a 'very weak' relationship, 0.1<I T_b I<0.19 indicating a 'weak' relationship, 0.20<I T_b I<0.29 indicating a 'moderate' relationship, and 0.30<I T_b I indicating a 'strong' relationship. The significance of Tb Statistics for dependence between subjective and objective methods of evaluation was set at $\alpha = 0.05$ [21,22].

Results

Baseline lameness

From the day 1 baseline trials of all 26 horses, the clinicians who performed the examinations gave the lameness score 1 to four horses, 2 to thirteen horses and 3 to nine horses. The absolute value of P_{max} ranged from 0.145 to 20.662 mm (mean = 7.382 mm), while the absolute value of P_{min} ranged from 0.823 to 19.888 mm (mean = 6.529 mm).

Interobserver agreement

For all 26 trials, interobserver agreement on hindlimb lameness evaluation was higher for the live clinical evaluation than for the video evaluation, for all three experience groups. Two clinicians in a clinical situation showed 'moderate' agreement, while three experienced groups of veterinarians viewing videos showed only 'fair' agreement (Table 1). Trials which revealed disagreements among veterinarians within each group were further investigated. Each trial was categorised into one of the three disagreement types which were 1) sound vs. lame, 2) left vs. right hindlimb lameness and 3) unilateral vs. bilateral hindlimb lameness (Supplementary Item 4). The percentage of each disagreement type was calculated from the ratio between the number of trials assigned to each type divided by all trials which showed disagreements from that group (Supplementary Item 5).

Agreement between subjective and objective evaluation for determination of hindlimb lameness

The percentage of agreement between subjective and objective evaluation is presented in Table 2. The overall agreement between clinicians in clinical situations and objective evaluations is higher than for the three experience-based groups of veterinarians evaluating video trials. In each experience-based group of veterinarians evaluating videos, agreement between each individual and objective evaluation varied greatly. In the highly experienced group, the percentage of agreement ranges from 54 to 77%. In the moderately experienced group, the percentage of agreement ranges from 42 to 77%, and in the inexperienced group, the percentage of agreement ranges from 38 to 69%. Overall, agreement within the highly experienced group is higher than in both the moderately experienced and inexperienced groups, while there is no difference observed in the percentage of agreement between moderately experienced and inexperienced group.

The agreement between objective evaluation and the clinician who performed the lameness examination was recorded at 77% ($\kappa = 0.546$, strength = moderate). From 26 baseline trials, the objective method identified hindlimb lameness category as: sound for two trials, left hindlimb lameness for 17 trials, right hindlimb lameness for five trials and bilateral hindlimb lameness for two trials. Of these, the clinicians who performed the lameness examination identified them as sound for none of the trials, left hindlimb lameness for 17 trials, right hindlimb lameness for seven trials and bilateral hindlimb lameness for two trials. Trials which resulted in disagreement between the inertial sensor system and veterinarians in each group were also categorised using the same categorisation as that used for interevaluator disagreement (Supplementary Item 6).

Effect of lameness severity on interobserver agreement, and the agreement between subjective and objective evaluation

Interobserver agreements for more severe trials were higher than for less severe trials of clinician in clinical live evaluation and the highly experienced and moderately experienced veterinarians, but not the inexperienced veterinarians (Table 1). Agreement between objective and subjective evaluation was also higher for more severe trials than for less severe trials in each veterinarian group, except for the two inexperienced individuals (Table 2).

Agreement on the response to diagnostic analgesia between subjective and objective evaluation

There were 43 test units in which all four highly experienced veterinarians selected the same lame limb as the limb which was currently blocked. The

TABLE 1: Interobserver agreement from clinicians in live situation, and the three experience-based groups evaluating videos for all trials, trials with less severe lameness (|P_{max}| = 0.145–8.511 mm, |P_{min}| = 0.823–8.340 mm) and more severe lameness (|P_{max}| = 3.629–20.662 mm, |P_{min}| = 1.395–19.888 mm)

	Live clinical evaluation		Video evaluation					
	Clinicians		High experience		Moderate experience		Inexperience	
	Agreement (%)	Kappa (strength)	Agreement (%)	Kappa (strength)	Agreement (%)	Kappa (strength)	Agreement (%)	Kappa (strength)
All trials (n = 26)	81	0.594 (moderate)	61	0.289 (fair)	63	0.294 (fair)	61	0.241 (fair)
Less severe trials (n = 13)	77	0.513 (moderate)	58	0.087 (slight)	63	0.256 (fair)	78	0.402 (fair-moderate)
More severe trials (n = 13)	77	0.733 (substantial)	86	0.719 (substantial)	77	0.528 (moderate)	48	0.12 (slight)

TABLE 2: The mean agreement of lameness detection between a body-mounted inertial sensor system and clinicians in live situation, and the three experience-based groups evaluating videos for all trials, trials with less severe lameness ($|P_{max}|$ = 0.145–8.511 mm, $|P_{min}|$ = 0.823–8.340 mm), and more severe lameness ($|P_{max}|$ = 3.629–20.662 mm, $|P_{min}|$ = 1.395–19.888 mm)

	Live clinical evaluation		Video evaluation					
	Clinicians		High experience		Moderate experience		Inexperience	
Trial group	Mean agreement (%)	Kappa (strength)	Mean agreement (%)	Kappa (strength)	Mean agreement (%)	Kappa (strength)	Mean agreement (%)	Kappa (strength)
All trials (n = 26)	75	0.546 (moderate)	66	0.385 (fair)	55	0.205 (slight-fair)	56	0.162 (slight)
Less severe (n = 13)	65	0.296 (fair)	60	0.16 (slight)	44	0.001 (slight)	52	−0.014 (poor)
More severe (n = 13)	81	0.722 (substantial)	92	0.846 (substantial)	81	0.603 (moderate-substantial)	63	0.337 (fair)

agreement between the BMISS and the three highly experienced individuals was positive and strong (T_b HE1 = 0.603, T_b HE2 = 0.612, T_b HE3 = 0.385). However, the fourth highly experienced veterinarian had a positive, but weak agreement with the BMISS (T_b HE4 = 0.106).

Discussion

The inter-rater agreement of veterinarians in live clinical evaluations was higher than when evaluating videos, regardless of the level of experience. The agreement between two veterinarians in clinical situations was moderate and the percentage of agreement was 80%, whereas the veterinarians from the three experience-based groups evaluating videos recorded only a fair agreement, with the percentage of agreement recorded at around 60%. Even though the clinicians had the opportunity to see horses in full lameness examinations including lungeing and flexion tests in the live clinical evaluation, they were asked to give scores directly after the straight line trials, and before the lungeing and flexion tests. Consequently, is it not possible that the advantage of a full lameness examination would increase the agreement between the two clinicians. Moreover, Keegan *et al.* [8] showed that the inter-rater agreement among experienced clinicians did not improve after seeing horses in full lameness examinations compared to straight line examinations. However, the fact that they were both aware of owners' complaints about hindlimb lameness would likely cause them to focus more on the hindlimbs, in contrast to video situations in which the veterinarians were unaware of the anamnesis of the horses.

As the inter-rater agreements of the three different experience-based groups evaluating the videos were all at a similar level, such a finding indicates that highly experienced veterinarians are not more reliable in evaluating hindlimb lameness than their more inexperienced counterparts when using this methodology. Similar results were found in a forelimb lameness study in which experienced clinicians and residents or interns evaluated videos of horses trotting on a treadmill [4].

Agreement between the inertial sensor system and the subjective evaluation was also strongest for the live evaluations and is located in the moderate margin. This is to be expected, assuming that, 1) the inertial sensor system provides relevant measurements indicating hindlimb lameness and 2) lameness evaluations in live situations were more reliable than in video evaluations. In the study from Donnell *et al.* [23], clinicians in live clinical evaluation also showed higher agreement with the BMISS and force plate in determining mild forelimb lameness when compared to video evaluations. Considering each experience-based group as a homogenous group, agreement was stronger in the highly experienced group and decreased with lowering experience. This would support the above statement about the reliability of the inertial sensor system. However, when considering the agreement of each individual, some highly experienced individuals had lower agreement with the inertial sensor method than some of their inexperienced counterparts. This could be explained by the fact that certain experienced clinicians complained that the height from which the horses were filmed clearly differed from the vision they normally see in clinical situations. As the video camera was mounted on a tripod which was 110 cm in height, it may have proved difficult for some clinicians to evaluate lameness from this particular angle. While the interns in general have less experience in participating in lameness examinations, they might have lower and more adaptable expectations and be more flexible to the video angle than certain more experienced clinicians.

The disagreement between the two clinicians in the live clinical evaluation in defining hindlimb lameness, according to the four outlined hindlimb lameness categories, occurred in 5 out of 26 trials (19%; Supplementary Item 5), in which they disagreed as to whether the horses were suffering from unilateral or bilateral hindlimb lameness. In four of these five trials, the 'lamer' limb in assumed bilateral hindlimb lame horses matched the lame limb assigned by the other clinicians. Therefore, the disagreement would have been much lower (1 out of 26, 4%) if hindlimb lameness had been categorised differently in this study (no hindlimb lameness, right hind lamer than left hind, and left hind lamer than right hind). A similar trend was found in the disagreement between the inertial sensor system and the two clinicians in the live clinical evaluation. Contrastingly, the veterinarians who evaluated the videos rather disagreed among each other in terms of assigning horses as not having hindlimb lameness, or on whether horses were suffering from unilateral left hindlimb lameness or unilateral right hindlimb lameness. Similar disagreements also arose among the veterinarians assessing both the videos and the inertial sensor system.

Generally, more pronounced lameness appeared to increase both interobserver agreement and the agreement between subjective and objective evaluation. Interobserver agreement in trials with more severe lameness was stronger than in trials with less severe lameness, in the live group, the highly experienced group and the moderately experienced group. These results were consistent with the results from another study in which experienced clinicians showed higher interobserver agreement for trials with a higher mean lameness score [8]. The agreements between the veterinarians and the BIMSS were also higher for trials with more severe lameness for most of the veterinarians, with the exception of two inexperienced individuals.

Care should be taken when conducting lameness evaluations through video viewing. The absence of the trotting sound, the angle from which the videos were taken, and seeing horses moving two dimensionally appear to have had an influence on the agreement of veterinarians in determining hindlimb lameness. The results of a forelimb lameness study by Rungsri *et al*. [24], in which horses were examined and filmed under identical conditions, also revealed lower agreement among veterinarians evaluating videos than those in the live clinical evaluation, both as interobservers and between the objective method. It has been demonstrated that when assessing videos of horses with hindlimb lameness only on the lunge, inter-rater agreement was also low [25]. On the other hand, in another study in which three clinicians evaluated videos with sound of horses trotting in both straight lines and lungeing, the interobserver agreement was found to be higher than the current study [26]. Therefore, observing horses both lungeing and trotting in straight lines, as well as being able to hear the sound of the horse trotting, should help to improve the reliability of the veterinarians evaluating videos in such a study.

The agreement on the response to anaesthesia between the objective method and experienced veterinarians was strong for three highly experienced veterinarians and followed the same direction. The BMISS and three highly experienced veterinarians agreed most strongly on improvement categories 1 and 5 (no improvement, and lameness switched to contralateral limb). However, there is a less clear agreement for improvement categories 2, 3 and 4 (less than 50% improvement, more than 50% improvement and lameness abolished). A similar trend of association was also found in the study performed by Rungsri *et al.* [24]. The reason could be that the task of categorising lameness improvement into either less or more than 50% is a rather subjective one for veterinarians. On the other hand, the inertial sensor system, especially for hindlimbs, has two separate values of interest, Pmax and Pmin, which do not facilitate a direct comparison to one improvement category given by a subjective individual. On the contrary, the agreement was weak for one highly experienced veterinarian (HE4). This veterinarian also had low agreement with the BMISS and the first clinician (K = 0.126 and 0.233) from the clinical situation (data not presented in results section). Therefore, we assumed that he was not familiar with the video evaluation, as discussed above.

This study does have some limitations. Firstly, the horses which participated in this study were all determined by veterinarians to have hindlimb lameness, and thereafter underwent a lameness evaluation which included diagnostic analgesia. In this way, the horses which were only identified as lame by the BMISS were left out of the study. Since the BMISS has been tested in order to be able to identify lower degrees of lameness compared to clinicians [27], it is very likely that the agreement between the BMISS and the clinicians in live clinical evaluation in this study would have been lower if the decision had been made to include those horses. Secondly, the effect of experience on interobserver agreement and on the agreement between veterinarians and the objective method was carried out by video evaluation. Even though veterinarians from each experience-based group encountered the same challenge for the video evaluation, some appeared to have more difficulty than others in familiarising themselves with the complications stated above, which arose from the video evaluation. In order to estimate the weakening of agreement due to video assessment, an intra-observer agreement could have been estimated. However, this was not the objective of the study. Lastly, the small number of horses is another limitation.

In conclusion, given the task of evaluating videos of horses trotting in a straight line, more experienced veterinarians did agree more among each other than their less experienced counterparts. Additionally, even though the agreement between veterinarians and the BMISS was stronger in the highly experienced group in comparison to the moderate and inexperienced groups, the agreement of some highly experienced individuals was lower than those with less experience. Therefore, it could be ascertained that more experienced veterinarians evaluating videos of horses trotting in a straight line were not necessarily more reliable than those with less experience. The results from this study encourage the use of the inertial sensor system as a supporting lameness diagnosing tool because of: 1) the moderate agreement between the inertial sensor system and the clinicians in the live clinical evaluation in determining hindlimb lameness, and 2) and the strong association between the sensor system and experienced veterinarians in determining an improvement in lameness after anaesthesia.

Authors' declaration of interests

No competing interests have been declared.

Ethical animal research

This study was approved by the Ethics Committee of the Freie Universität Berlin.

Owner informed consent

Owners consented for their horses to take part in the study.

Sources of funding

No funding was received for this study.

Acknowledgements

We would like to thank the veterinarians who helped to perform the lameness examination and took part in video evaluation, as well as the owners, grooms and horses that made this study possible.

Authorship

P. Leelamankong was the principal author and contributed to study design, data collection and analysis, and manuscript preparation. R. Estrada contributed to study design and data collection and revising the content. K. Mählmann contributed to data analysis and revising the manuscript. P. Rungsri contributed to study design, data collection, and revising the manuscript. C. Lischer was the senior author and contributed to overall study design, project coordination, data analysis, manuscript preparation and revising the manuscript. All authors gave their final approval of the manuscript.

Data availability statement

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

Manufacturers' addresses

[a]Equinosis LLC, Columbia, Missouri, USA.
[b]Sony Corporation, Minato, Tokyo, Japan.
[c]Microsoft Corporation, Redmond, Washington, USA.
[d]RStudio, Inc., Boston, Massachusetts, USA.

References

1. Ross, M.W. (2011) Movement. In: *Diagnosis and Management of Lameness in the Horse*, 2nd edn., Eds: M.W. Ross and S.J. Dyson, W.B. Saunders, St. Louis, MO. pp 64–80.
2. Baxter, G.M. (2011) Examination of lameness. In: *Adams and Stashak's Lameness in Horses*, 6th edn., Ed: G.M. Baxter, Chichester, Wiley, pp 109–206.
3. Ross, M.W. (2011) Lameness examination: historical perspective. In: *Diagnosis and Management of Lameness in the Horse*, 2nd edn., Eds: M.W. Ross and S.J. Dyson, W.B. Saunders, St. Louis, MO. pp 1–2.
4. Keegan, K.G., Wilson, D.A., Wilson, D.J., Smith, B., Gaughan, E.M., Pleasant, R.S., Lillich, J.D., Kramer, J., Howard, R.D., Bacon-Miller, C., Davis, E.G., May, K.A., Cheramie, H.S., Valentino, W.L. and van Harreveld, P.D. (1998) 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.* **59**, 1370–1377.
5. Parkes, R.S., Weller, R., Groth, A.M., May, S. and Pfau, T. (2009) Evidence of the development of 'domain-restricted' expertise in the recognition of asymmetric motion characteristics of hindlimb lameness in the horse. *Equine Vet. J.* **41**, 112–117.
6. Buchner, H.H., Savelberg, H.H., Schamhardt, H.C. and Barneveld, A. (1996) Head and trunk movement adaptations in horses with experimentally induced fore- or hindlimb lameness. *Equine Vet. J.* **28**, 71–76.
7. Kaneps, A.J. (2004) Diagnosis of lameness. In: *Equine Sports Medicine and Surgery. Basic and Clinical Sciences of the Equine Athlete*. Eds: K.W. Hinchcliff, A.J. Kaneps and R.J. Geor, Saunders, St. Louis, MO. pp 247–259.
8. Keegan, K.G., Dent, E.V., Wilson, D.A., Janicek, J., Kramer, J., Lacarrubba, A., Walsh, D.M., Cassells, M.W., Esther, T.M., Schiltz, P., Frees, K.E., Wilhite, C.L., Clark, J.M., Pollitt, C.C., Shaw, R. and Norris, T. (2010) Repeatability of subjective evaluation of lameness in horses. *Equine Vet. J.* **42**, 92–97.

9. Hewetson, M., Christley, R.M., Hunt, I.D. and Voute, L.C. (2006) Investigations of the reliability of observational gait analysis for the assessment of lameness in horses. *Vet. Rec.* **158**, 852–857.
10. Arkell, M., Archer, R.M., Guitian, F.J. and May, S.A. (2006) Evidence of bias affecting the interpretation of the results of local anaesthetic nerve blocks when assessing lameness in horses. *Vet. Rec.* **159**, 346–349.
11. Keg, P.R., Barneveld, A., Schamhardt, H.C. and van den Belt, A.J. (1994) Clinical and force plate evaluation of the effect of a high plantar nerve block in lameness caused by induced mid-metatarsal tendinitis. *Vet. Q.* **16**, *Suppl.* **2**, S70–75.
12. Weishaupt, M.A., Hogg, H.P., Wiestner, T., Denoth, J., Stussi, E. and Auer, J.A. (2002) Instrumented treadmill for measuring vertical ground reaction forces in horses. *Am. J. Vet. Res.* **63**, 520–527.
13. Peloso, J.G., Stick, J.A., Soutas-Little, R.W., Caron, J.C., DeCamp, C.E. and Leach, D.H. (1993) Computer-assisted three-dimensional gait analysis of amphotericin-induced carpal lameness in horses. *Am. J. Vet. Res.* **54**, 1535–1543.
14. Pfau, T., Robilliard, J.J., Weller, R., Jespers, K., Eliashar, E. and Wilson, A.M. (2007) Assessment of mild hindlimb lameness during over ground locomotion using linear discriminant analysis of inertial sensor data. *Equine Vet. J.* **39**, 407–413.
15. Keegan, K.G., Kramer, J., Yonezawa, Y., Maki, H., Pai, P.F., Dent, E.V., Kellerman, T.E., Wilson, D.A. and Reed, S.K. (2011) Assessment of repeatability of a wireless, inertial sensor-based lameness evaluation system for horses. *Am. J. Vet. Res.* **72**, 1156–1163.
16. Thomsen, M.H., Persson, A.B., Jensen, A.T., Sorensen, H. and Andersen, P.H. (2010) Agreement between accelerometric symmetry scores and clinical lameness scores during experimentally induced transient distension of the metacarpophalangeal joint in horses. *Equine Vet. J.* **42**, *Suppl.* **38**, 510–515.
17. Kramer, J., Keegan, K.G., Kelmer, G. and Wilson, D.A. (2004) Objective determination of pelvic movement during hind limb lameness by use of a signal decomposition method and pelvic height differences. *Am. J. Vet. Res.* **65**, 741–747.
18. Maliye, S., Voute, L., Lund, D. and Marshall, J.F. (2013) An inertial sensor-based system can objectively assess diagnostic anaesthesia of the equine foot. *Equine Vet. J.* **45**, *Suppl.* **45**, 26–30.
19. Marshall, J.F., Lund, D.G. and Voute, L.C. (2012) Use of a wireless, inertial sensor-based system to objectively evaluate flexion tests in the horse. *Equine Vet. J.* **44**, *Suppl.* **43**, 8–11.
20. Keegan, K.G., Wilson, D.A., Kramer, J., Reed, S.K., Yonezawa, Y., Maki, H., Pai, P.F. and Lopes, M.A. (2013) Comparison of a body-mounted inertial sensor system-based method with subjective evaluation for detection of lameness in horses. *Am. J. Vet. Res.* **74**, 17–24.
21. Agresti, A. (2010) *Analysis of Ordinal Cateogorical Data*, 2nd edn., John Wiley & Sons, New Jersey.
22. Sen, P.K. (1968) Estimates of the regression coefficient based on Kendall's Tau. *J. Am. Stat. Ass.* **63**, 1379–1389.
23. Donnell, J.R., Frisbie, D.D., King, M.R., Goodrich, L.R. and Haussler, K.K. (2015) Comparison of subjective lameness evaluation, force platforms and an inertial-sensor system to identify mild lameness in an equine osteoarthritis model. *Vet. J.* **206**, 136–142.
24. Rungsri, P., Staecker, W., Leelamankong, P., Estrada, R., Rettig, M., Klaus, C. and Lischer, C. (2014) Agreement between a body-mounted inertial sensors system and subjective observational analysis when evaluating lameness degree and diagnostic analgesia response in horses with forelimb lameness. *Pferdeheilkunde* **30**, 633–650.
25. Hammarberg, M., Egenvall, A., Pfau, T. and Rhodin, M. (2016) Rater agreement of visual lameness assessment in horses during lungeing. *Equine Vet. J.* **48**, 78–82.
26. Fuller, C.J., Bladon, B.M., Driver, A.J. and Barr, A.R. (2006) The intra- and inter-assessor reliability of measurement of functional outcome by lameness scoring in horses. *Vet. J.* **171**, 281–286.
27. McCracken, M.J., Kramer, J., Keegan, K.G., Lopes, M., Wilson, D.A., Reed, S.K., LaCarrubba, A. and Rasch, M. (2012) Comparison of an inertial sensor system of lameness quantification with subjective lameness evaluation. *Equine Vet. J.* **44**, 652–656.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Supplementary Item 1: Lameness grading scale adapted from Ross 2011 [1].

Supplementary Item 2: Description of four hindlimb lameness categories according to subjective lameness score and the BMISS measurements.

Supplementary Item 3: Description of improvement categories according to subjective score and percentage decrease of Pmax and Pmin.

Supplementary Item 4: Type of disagreement of lameness evaluation between two evaluators.

Supplementary Item 5: Percentage disagreement of lameness evaluation from each group of veterinarians.

Supplementary Item 6: Percentage disagreement of lameness evaluation between inertial sensor system and observers from each group.