Journal of Equine Veterinary Science 154 (2025) 105704

Contents lists available at ScienceDirect

Journal of Equine Veterinary Science

journal homepage: www.elsevier.com/locate/jevs

Original Research

Agreement between subjective evaluations and a markerless AI-based gait analysis system during lungeing assessment in traditional racehorses

F. Meistro *, M.V. Ralletti, R. Rinnovati, A. Spadari

Department of Veterinary Medical Sciences, University of Bologna, Via Tolara di Sopra 50, 40064 Ozzano dell'Emilia BO, Italy

ARTICLE INFO

Keywords:
Agreement analysis
Equine gait evaluation
Artificial intelligence
Objective gait analysis
Pre-race inspection

ABSTRACT

Background: Subjective lameness evaluation during lungeing is routinely performed in equine practice, but its consistency remains limited, especially in cases of mild or complex asymmetry.
Aims: This study aimed to assess the agreement between subjective gait evaluations and a markerless AI-based gait analysis system (OAI-MS) in traditional racehorses during lungeing. Intra- and inter-observer agreement of subjective evaluations was also investigated.
Methods: 24 traditional racehorses were assessed during routine pre-race inspections (T0) while trotting on a soft surface. Two experienced equine clinicians independently evaluated each horse on both reins using the AAEP 0–5 scale; scores were then converted to a 3-level ordinal scale (0 = sound, 1 = mild, 2 = severe). Simultaneously, gait data were collected using the OAI-MS. A subset of 10 horses was re-evaluated after 10 days (T1) to assess short-term repeatability of the OAI-MS. Video-based reassessment (T2) was used to evaluate intra-observer agreement. Agreement was calculated using weighted Cohen's and Fleiss' kappa. $p < 0.05$.
Results: Inter-observer agreement ranged from $\kappa = -0.20$ to 0.36. Agreement between subjective evaluators and the OAI-MS ranged from slight to moderate ($\kappa = 0.13$–0.47). Intra-observer agreement was fair ($\kappa \approx 0.22$), and OAI-MS repeatability reached $\kappa = 0.43$. Agreement was higher for forelimbs than hindlimbs. Most discrepancies were of low magnitude.
Conclusion: Subjective gait evaluations during lungeing showed limited agreement. The OAI-MS demonstrated moderate repeatability, supporting its usability in the field and its potential role as a complementary tool in clinical decision-making, particularly when asymmetries are mild or disagreement occurs.

1. Introduction

Early recognition of lameness is crucial for enabling effective clinical decisions and preserving the horse's athletic longevity [1–3]. Standard subjective lameness examinations typically include trotting in a straight line and on a circle. The latter is particularly useful, as it may amplify subtle asymmetries not visible on a straight path [3–7]. However, evaluating gait on a circle presents unique biomechanical challenges. It involves physiological asymmetries in limb loading and the combined effects of centripetal and centrifugal forces, which can influence locomotor patterns even in the absence of lameness [8–10]. Despite being the clinical standard [11], subjective lameness evaluation lacks universally accepted criteria, leading to considerable variability between observers, particularly in subtle cases [12,13]. Even among experienced veterinarians, agreement is typically moderate, with reported accuracies ranging from 72 to 76 % for the forelimb and 64–69 % for the hindlimb. Notably, inter-observer agreement improves when lameness exceeds 1.5/5 on the AAEP scale [13,14]. During lungeing, the complexity of circular biomechanics further complicates visual interpretation, making variability in subjective evaluation even more evident and reducing inter-observer agreement [15]. While video recordings could theoretically allow repeated review and more objective comparison, their use has not consistently enhanced inter-observer reliability, especially when evaluating mild gait asymmetries [13,15]. Variability in clinical experience among veterinarians may further contribute to inconsistent evaluation [13,16].

To overcome these limitations, objective gait analysis technologies such as optical motion capture (OMC) systems and inertial measurement units (IMUs) have been developed [16–19]. While highly accurate, these systems require specialized equipment and controlled environments, limiting their application in the field [12,16,20,21].

Recently, a markerless artificial intelligence (AI)-based gait analysis

* Corresponding author.
E-mail address: federica.meistro@unibo.it (F. Meistro).

https://doi.org/10.1016/j.jevs.2025.105704
Received 6 August 2025; Received in revised form 9 September 2025; Accepted 27 September 2025
Available online 28 September 2025

system (OAI-MS) (Sleip AI®) has emerged as a practical alternative for field use [22]. This smartphone-based application uses computer vision to automatically track the horse's head, pelvis, and hooves throughout the stride cycle. It calculates vertical asymmetries with frame-by-frame precision, delivering objective results without the use of sensors or reflective markers applied directly to the horse's body [22,23]. When compared to an OMC, it showed a mean error of 2.2 mm for head and pelvis vertical displacement [22].

A clear definition of output parameters is critical when presenting AI-based gait analysis results, as it directly influences interpretability and clinical applicability. This need for unambiguous parameter specification has been highlighted both in human gait recognition research, [24], and in broader AI domains, where standardized evaluation frameworks have been proposed to ensure reliability, consistency, and interpretability of outcomes [25].

The OAI-MS has demonstrated strong agreement with IMUs and fair to substantial agreement with subjective evaluations on a straight line [18,26,27]. Moreover, a recent comparative study evaluating straight-line trot, including the OAI-MS among other objective systems, reported that individual agreement scores between objective methods were higher and less variable than those obtained by veterinary evaluators. The data also suggested that objective systems were more consistent than subjective examination in detecting subtle hindlimb asymmetries [28].

However, no studies to date have investigated the agreement between subjective evaluations and the OAI-MS during lungeing. Furthermore, intra-observer consistency between live and delayed video-based assessments has not been systematically assessed.

In this study, horses were evaluated during routine pre-race veterinary inspections, which are part of the official pre-competition assessment in a traditional racing event, such as l the Palios [29]. This provided a real-world context to assess the reliability of subjective gait assessment and its agreement with objective gait analysis under field conditions.

Therefore, the aims of this study were to 1) assess the level of agreement between subjective evaluations performed by equine veterinarians and the OAI-MS during lungeing; 2) assess inter-observer agreement between clinicians during lungeing evaluations; 3) evaluate intra-observer agreement between live and delayed video-based assessments performed four months apart.

2. Materials and methods

2.1. Ethical approval

As no experimental procedures were performed, ethical approval was waived in accordance with the University of Bologna's internal guidelines. All procedures were conducted during routine pre-race veterinary inspections. Written informed consent was obtained from all owners for the use of their horses' clinical data in this study.

2.2. Horses

The study population included 24 racehorses (20 Thoroughbreds and 4 Anglo-Arabians), aged 4 to 13 years (mean: 7), all actively competing in the "Palio of Faenza" traditional racing circuit, stabled at the same training facility, where they followed similar training routines.

Before enrolment, each horse underwent a basic clinical examination. It included auscultation, rectal temperature, and general physical assessment, palpation of distal limbs and back, and application of hoof testers on the forelimbs. Horses with evident clinical abnormalities were excluded.

Given the within-subject design of the study, each horse served as its own control, allowing direct comparison between subjective and objective evaluations.

2.3. Lameness assessment protocol (T0–T1-T2)

Two equine veterinarians (Observer 1 and Observer 2), each with > 10 years of experience in equine orthopaedics, independently evaluated all horses during lungeing at T0, T1, and T2. All horses were lunged on a consistent soft sand surface using a 12–15-meter diameter circle for a minimum of 45 seconds in each direction (left and right rein). This setup ensured standardization of environmental conditions and met the technical requirements of the OAI-MS, which needs a sufficient number of consecutive strides under uniform footing to generate reliable data [23].

The initial lameness evaluation (T0) was performed on all 24 horses and involved subjective and objective assessments. During this phase, the two observers independently assigned lameness scores using the AAEP scale (0–5) to each limb (forelimbs: FL, FR; hindlimbs: HL, HR) [14], which provides a standardized ordinal system for evaluating the severity of lameness based on visual gait assessment. Each horse was evaluated separately on the lunge to the right and to the left, and a distinct score was assigned for each direction. Observers recorded their scores immediately after each trial on individual paper forms, specifying the affected limb(s) (forelimb or hindlimb). They were blinded to each other's evaluations throughout the study and did not communicate during or after the assessments to avoid any potential bias. At the same time, a third veterinarian conducted the objective gait analysis using the OAI-MS.

A follow-up evaluation **at T1 (repeat assessment on a subset)** was conducted 10 days later on a randomly selected subset of 10 horses under identical environmental and procedural conditions. Although minor changes in lameness presentation over this short interval cannot be entirely ruled out, the reduced interval was intended to minimize clinical variability. The primary aim of this follow-up was to assess the short-term repeatability and practical applicability of both evaluation methods under consistent conditions. Given the observational design and the logistical constraints of official pre-race inspections, a priori power analysis was not feasible. Instead, sample size was justified ex post based on the precision of agreement estimates, considering confidence interval width for Cohen's κ as the relevant indicator of adequacy.

At T2 (video-based reassessment four months post-T0), both subjective evaluators independently re-scored the original T0 videos to assess intra-observer agreement. They were blinded to their prior scores and to each other's assessments. The four-month interval was intentionally chosen to minimize recall bias and reduce the likelihood that evaluators would remember their initial assessments.

2.4. Objective motion analysis system

Objective gait data were acquired using Sleip AI® (version 1.0.10). Recordings were made with an iPhone® 14 Pro mounted on a tripod approximately 10 m from the lunge circle, following the manufacturer's guidelines [22,23]. The system does not require markers or sensors and automatically identifies anatomical landmarks across the stride cycle.

Although straight-line trot-ups were performed during the clinical evaluation, they were not considered in the present analysis, as they were not relevant to the study's objectives. The operator ensured that the entire horse remained fully visible within the frame for the duration of each recording, as this is a critical requirement for the system to correctly track movement and extract reliable data. Video acquisition was performed in the field, and the files were subsequently uploaded to the OAI-MS platform for analysis. Although the system allows immediate feedback, in this study, all recordings were processed only after completing data collection, in order to maintain consistency across horses and time points.

Once uploaded, each video was analysed by the software, which automatically extracted asymmetry parameters. The OAI-MS system extracted asymmetry parameters related to vertical displacement of the upper body: HDmin and HDmax for the head, and PDmin and PDmax for

the pelvis. HDmin and PDmin quantify differences between the minima vertical positions reached during successive strides and are associated with asymmetries during the impact phase. HDmax and PDmax represent differences between the maxima vertical positions and reflect asymmetries during the push-off phase. These values are computed separately for each stride and averaged across the analysed sequence.

The OAI-MS automatically processes these raw values and calibrates them, transforming the millimetric displacement data into a standardized numerical scale ranging from 0 to 2. This conversion is accompanied by a color-coded visual output that facilitates interpretation by end users. Based on internally defined thresholds, vertical displacement differences are categorized as follows: Green: symmetry (<0.2); Grey: very mild asymmetry (0.2–0.49); Yellow: mild asymmetry (0.5–0.99 mm); Orange: moderate asymmetry (1.0–1.49); Red: severe asymmetry (≥ 1.5)

Each limb is displayed in the interface with a corresponding colour, allowing quick visual identification of asymmetry severity. These thresholds are predefined by the system and applied automatically during the analysis. In line with previous studies [30,31], asymmetries ≥ 0.5 were considered potentially detectable by trained observers. However, in this study, the term "asymmetry" strictly refers to objectively measured vertical displacement differences and does not imply clinical lameness.

When bilateral asymmetries were detected, typically consisting of a push-off asymmetry in one limb and an impact asymmetry in the contralateral limb, the limb associated with the highest amplitude of asymmetry is considered the most affected for the per-trial analysis. Nonetheless, all four limb scores were retained for the limb-specific analysis. Upon processing, the software generated an individual report for each video, summarizing the numerical values and their corresponding visual indicators. Videos that failed to meet the system's internal criteria for data quality, such as recordings with erratic gait, camera misalignment, or an insufficient number of strides, were flagged as unsuitable for analysis. However, in this study, all 24 horses included produced high-quality recordings that were successfully processed and retained for further evaluation.

2.5. Data processing

Following video analysis, the OAI-MS system extracted displacement measurements for HDmin, HDmax, PDmin, and PDmax, which served as the basis for generating the final calibrated asymmetry scores. By default, the system assigns negative values to asymmetries involving the left limbs and positive values to those affecting the right. To enable consistent interpretation regardless of laterality, all values were converted to their absolute magnitude before analysis. The classification of certain values may vary depending on the presence of single versus multiple asymmetry components. For instance, a 0.5 asymmetry may be scored as 1 if detected in isolation, but as 2 when accompanied by an additional asymmetry component (e.g., 0.5 in impact and 0.3 in push-off). This context-dependent classification is consistently applied by the system and was reflected in the ordinal scores used throughout this study.

To enable direct comparison, all scores were converted to a 3-level ordinal scale (0 = sound/no asymmetry; 1 = mild; 2 = severe) following the approach adopted in previous studies [27,28]. Subjective scores recorded using the AAEP 0–5 scale were reclassified as follows: 0 = "sound" (0); 1–2 = "mild" (1); ≥ 3 = "severe" (2), consistent with published methods. Objective asymmetry values provided by the OAI-MS were also reclassified using pre-defined thresholds: asymmetries <0.5 were considered "no asymmetry" (0); values between 0.5–0.99 as "mild" (1); and values ≥ 1.0 as "severe" (2). This harmonized classification was applied uniformly across all agreement analyses to allow direct comparison between observers and the OAI-MS system.

All processed data, including raw asymmetry values, scaled scores, and the evaluations from both observers, were compiled into a single Excel worksheet. Each row represented a unique trial, defined by horse ID, session (T0, T1, or T2), lungeing direction (left or right), and limb (right front: RF, left front: LF, right hind: RH, left hind: LH), resulting in a coherent dataset ready for statistical analysis. This structure allowed for comparison both globally (per trial) and regionally (forelimbs vs hindlimbs), based on the most affected limb per category.

2.6. Statistical analysis

Descriptive statistics were performed at two levels. First, limb-level scores (scale 0–5) were analysed to provide a detailed overview of the distribution of values assigned by each evaluator to each limb (RF, LF, RH, LH), for each time point (T0, T1, and T2). This analysis included the calculation of the mean, standard deviation, median, interquartile range, and full range of scores. Second, a descriptive summary was conducted for the total asymmetry scores (three-category ordinal scale), based on the most asymmetric limb per evaluation, separately for each session. Normality was assessed using the Shapiro–Wilk test, and values were expressed as mean $\pm$ SD or median with range, as appropriate. Frequencies and percentages were calculated for each score category and for the magnitude of disagreement between raters.

The primary aim of the statistical analysis was to assess the level of agreement between the two observers and with the OAI-MS, as well as to assess intra-observer and intra-system consistency over time. Although lameness scores were collected for all four limbs individually (RF, LF, RH, LH), the analysis of agreement was conducted at the trial level, selecting the limb with the highest asymmetry (subjective or objective) for each horse and direction. In addition, a subgroup analysis was performed by anatomical region (forelimbs vs hindlimbs) to explore possible differences in agreement patterns between limb groups.

Agreement analysis was performed at multiple levels. Inter-rater agreement was assessed by comparing the scores assigned by Observer 1 and Observer 2 at T0, T1, and T2. Intra-rater agreement was evaluated by comparing each observer's live scores at T0 with their own video-based reassessments at T2. Inter-method agreement was investigated by comparing the subjective scores of each observer with the objective scores generated by the OAI-MS at T0 and T1. Finally, intra-system agreement was assessed by comparing the OAI-MS results obtained at T0 and T1 for the subset of 10 horses that underwent repeated evaluation (Table 1).

For all pairwise comparisons, agreement was quantified using quadratically weighted Cohen's kappa, which accounts for the ordinal nature of the scoring scales and gives more weight to larger

Table 1
Summary of agreement comparisons conducted in this study, with corresponding statistical methods. Weighted Cohen's kappa was used for all pairwise analyses to account for the ordinal nature of the scores. Fleiss' kappa was applied to assess overall agreement among the two subjective evaluators and the OAI-MS. In addition, the absolute difference in scores was calculated for each pairwise comparison to evaluate the magnitude of disagreement.

Comparison	Type of Agreement	Test Used
Observer 1 vs Observer 2 (T0, T1, T2)	Inter-rater	Weighted Cohen's kappa
Observer 1 vs OAI-MS (T0, T1)	Inter-method	Weighted Cohen's kappa
Observer 2 vs OAI-MS (T0, T1)	Inter-method	Weighted Cohen's kappa
Observer 1 T0 vs T2	Intra-rater	Weighted Cohen's kappa
Observer 2 T0 vs T2	Intra-rater	Weighted Cohen's kappa
OAI-MS vs T1	Intra-system repeatability	Weighted Cohen's kappa
Observer 1 vs Observer 2 vs OAI-MS (T0, T1)	Global agreement (3 raters)	Fleiss' kappa
All pairwise comparisons	Magnitude of disagreement	Absolute score differences

disagreements. Fleiss' kappa was used to assess the overall agreement between the two subjective evaluators and the OAI-MS at T0 and T1, which provides a global measure of agreement across more than two raters. This allowed us to examine the general consistency between all three methods (Table 1).

Kappa coefficients were interpreted according to established benchmarks: <0.20 = poor; 0.21–0.40 = fair; 0.41–0.60 = moderate; 0.61–0.80 = substantial; >0.80 = almost perfect agreement. All kappa values were reported with 95 % confidence intervals.

To further contextualize agreement, absolute score differences were calculated for each comparison, and the frequency of 1-, 2-, or >2-point differences was reported.

While agreement was primarily assessed per horse and direction based on the most asymmetric limb, a secondary analysis was also conducted to explore differences between forelimb and hindlimb evaluations.

All statistical tests were conducted using SPSS Statistics (version 29) and GraphPad Prism (version 10). A significance level of $p < 0.05$ was applied throughout.

3. Results

3.1. Overview of evaluations

All 24 horses were included in the analysis. Each horse was evaluated at the trot on the lunge in both directions (left and right) at three time points: T0 (initial live assessment), T1 (repeat assessment on a subset), and T2 (video-based reassessment).

At T0, each horse was evaluated in both directions on the lunge by two independent observers in real time, and simultaneously assessed using the OAI-MS system, resulting in 96 subjective evaluations and 48 objective assessments. At T1, a subset of 10 horses underwent repeat live evaluation, generating 40 additional subjective evaluations and 20 objective assessments. At T2, the original videos recorded during T0 were re-evaluated in a blinded fashion by both observers, producing an additional 96 subjective assessments; no OAI-MS evaluations were performed at this time point.

In total, Observer 1 and Observer 2 each performed 68 live assessments (T0 + T1) and 48 video-based assessments (T2), for a total of 116 subjective evaluations per observer. The OAI-MS contributed 68 objective assessments across T0 and T1. Altogether, 232 subjective and 68 objective evaluations were included in the study, for a cumulative total of 300 assessments across all methods and sessions (Table 2).

3.2. Distribution of lameness scores

Each evaluation involved four limb-specific observations, producing a total of 1,200 individual scores: 928 from subjective evaluations (464 each by Observer 1 and Observer 2) and 272 from the objective analysis system (OAI-MS) (Table 3).

The results were analysed both by considering the most asymmetric limb per evaluation (total asymmetry) and by examining all individual limb scores.

When considering the most asymmetric limb per trial and using the three-category ordinal scale (0 = sound, 1 = mild/moderate asymmetry, 2 = severe), the OAI-MS system classified 70.2 % of evaluations as sound, 28.7 % as mildly/moderately asymmetric, and 1.1 % as severe. Observer 1 reported 85.7 % as sound, 13.6 % as mild/moderate, and 0.7 % as severe. Observer 2 showed a slightly broader distribution, with 78.7 % sound, 20.2 % mild/moderate, and 1.1 % severe cases.

Table 3
Number of evaluations and corresponding limb-level scores performed during each session. Subjective evaluations were conducted by two observers (Observer 1 and Observer 2) in real time (T0, T1) and via blinded video reassessment (T2). Objective assessments were obtained using the OAI-MS during live sessions only. Each evaluation consisted of four limb-specific observations, resulting in a total of 1,200 individual scores (928 subjective, 272 objective).

Session	Subjective evaluations (n)	Objective evaluations (n)	Subjective limb scores (n)	Objective limb scores (n)	Total limb scores (n)
T0	96	48	384	192	576
T1	40	20	160	80	240
T2	96	0	384	0	384
Total	232	68	928	272	1200

The distribution of scores assigned to individual limbs on the original 0–5 AAEP-based scale reflected similar patterns. The majority of limbs were classified as sound: 85.4 % by Observer 1, 78.7 % by Observer 2, and 69.9 % by the OAI-MS. Very mild asymmetries (score 1) were recorded in 10.3 % of limbs by Observer 1, 13.2 % by Observer 2, and 20.1 % by the OAI-MS. Mild asymmetries (score 2) were less frequent but still showed a consistent trend: 3.3 % for Observer 1, 7.0 % for Observer 2, and 8.5 % for OAI-MS. Moderate scores (≥3) were rare, not exceeding 1.1 % in any group, and severe scores of 4 were only used by the OAI-MS in 0.4 % of cases (Fig. 1).

3.3. Descriptive statistics

At the limb level (0–5 AAEP-based scale), mean scores assigned by the OAI-MS ranged from 0.22 to 0.68, with the highest values observed in the front limbs (0.68 for LF and 0.44 for RF). Observer 1 reported mean scores between 0.09 and 0.36, while Observer 2 ranged from 0.03 to 0.39. Median values were 0 for all limbs and evaluators. Standard deviations were generally higher in the forelimbs, particularly for the OAI-MS (up to 0.84 in LF). Maximum values reached 3 for both observers and 4 for the OAI-MS, but only in the forelimbs (Table 4).

Descriptive statistics for total asymmetry scores (three-category ordinal scale), calculated as the maximum value among the four limbs per evaluation, showed a similar distribution. Mean total asymmetry scores were 0.49 for the OAI-MS, 0.23 for Observer 1, and 0.28 for Observer 2. Median values were 0 for all evaluators, and standard deviations were 0.66 (OAI-MS), 0.51 (Observer 1), and 0.58 (Observer 2). The maximum score of 2 was recorded for each evaluator in a small number of cases.

The full set of descriptive values for each evaluator and limb is reported in Table 5.

The number of evaluations in which at least one limb was scored above 0 was calculated for each subjective evaluator and the OAI-MS

Table 2
Number of assessments performed at each time point, categorized by evaluator and method. At T0 and T1, evaluations were conducted live by two independent observers, with simultaneous objective analysis using the OAI-MS. At T2, the original video recordings from T0 were reassessed in a blinded fashion by both observers; no objective evaluations were performed at this stage. The table reports live and video-based subjective assessments separately, along with objective assessments and cumulative totals.

Session	Observer 1 (live)	Observer 2 (live)	Observer 1 (video)	Observer 2 (video)	OAI-MS (objective)	Subjective total	Objective total	Total assessments
T0	48	48	0	0	48	96	48	144
T1	20	20	0	0	20	40	20	60
T2	0	0	48	48	0	96	0	96
Total	68	68	48	48	68	232	68	**300**

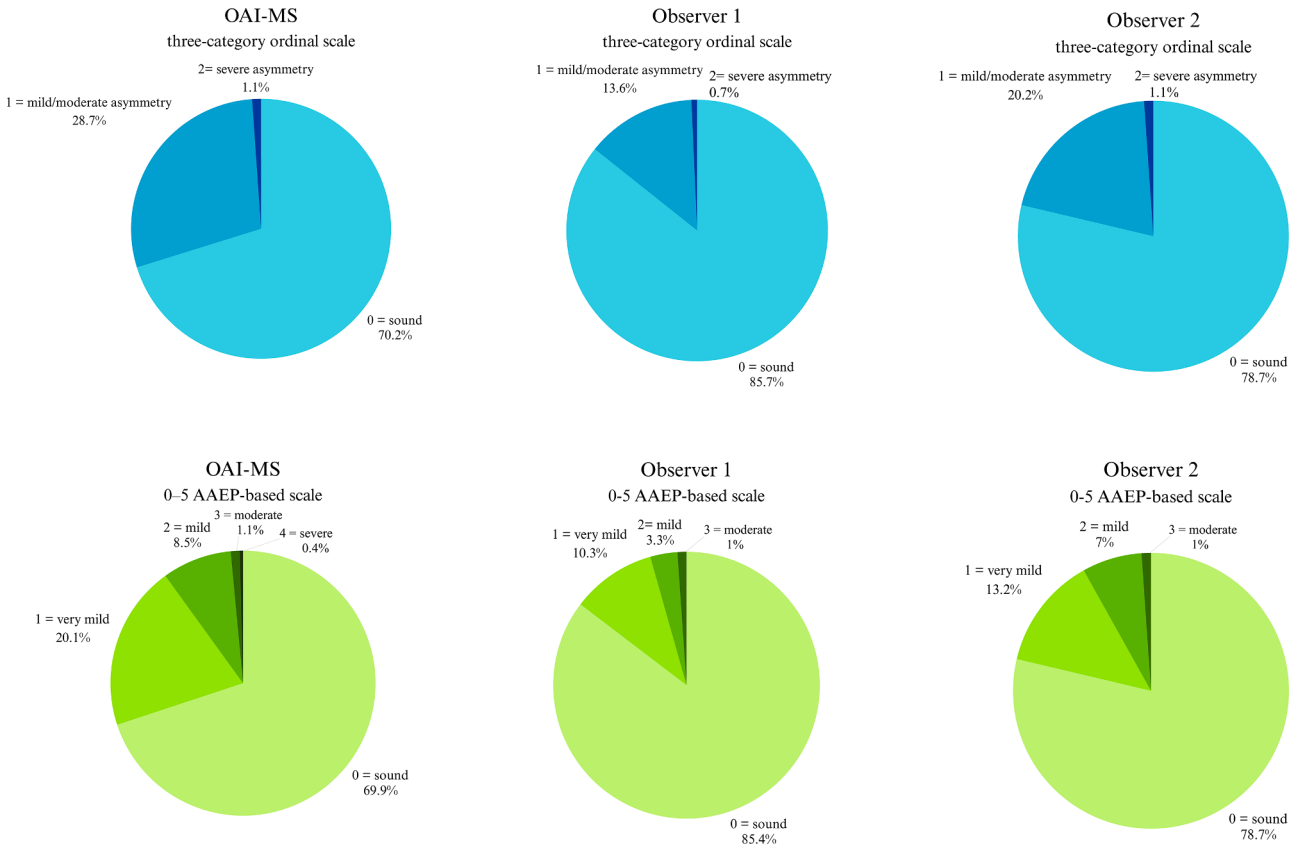

Fig. 1. Distribution of lameness scores assigned by the objective gait analysis system (OAI-MS), Observer 1, and Observer 2. The top row shows the classification based on the most asymmetric limb per evaluation (total asymmetry), using the three-category ordinal scale: 0 = sound, 1 = mild to moderate asymmetry, 2 = severe asymmetry. The bottom row illustrates the distribution of all individual limb scores using the full AAEP-based scale (0–5). Score 5, which corresponds to a non-weight-bearing limb, does not appear as it was not assigned in any case.

Table 4
Descriptive statistics of limb-specific lameness scores assigned by the OAI-MS, Observer 1, and Observer 2 using the 0–5 AAEP-based scale. For each limb (right front: RF, left front: LF, right hind: RH, left hind: LH), the table reports the number of observations (n), mean, median, standard deviation (SD), interquartile range (Q1–Q3), and minimum and maximum values. Data were pooled across all sessions (T0, T1, T2).

Evaluator	Limb	n	mean	median	std	Q1	Q3	min	max
OAI-MS	RF	68	0.44	0	0.8	0	1	0	3
OAI-MS	LF	68	0.68	0	0.84	0	1	0	4
OAI-MS	RH	68	0.22	0	0.48	0	0	0	2
OAI-MS	LH	68	0.29	0	0.57	0	0	0	2
Observer 1	RF	116	0.36	0	0.66	0	1	0	3
Observer 1	LF	116	0.29	0	0.56	0	0,25	0	3
Observer 1	RH	116	0.09	0	0.28	0	0	0	1
Observer 1	LH	116	0.09	0	0.28	0	0	0	1
Observer 2	RF	116	0.39	0	0.66	0	1	0	3
Observer 2	LF	116	0.38	0	0.69	0	1	0	3
Observer 2	RH	116	0.03	0	0.16	0	0	0	1
Observer 2	LH	116	0.23	0	0.57	0	0	0	3

Table 5
Descriptive statistics of total asymmetry scores (three-category ordinal scale), based on the maximum score among the four limbs per evaluation. The table includes the number of observations (n), mean, median, standard deviation (SD), interquartile range (Q1–Q3), and minimum and maximum values for each evaluator. Data reflect total asymmetry scores calculated across all sessions.

Evaluator	n	mean	median	std	q25	q75	min	max
OAI-MS	68	1.26	1	0.82	1	2	0	4
Observer 1	116	0.78	1	0.7	0	1	0	3
Observer 2	116	0.95	1	0.79	0	1	0	3

using the total asymmetry data. The OAI-MS identified asymmetry (score > 0) in 57 out of 68 evaluations, corresponding to 83.8 %. Observer 1 reported asymmetries in 74 out of 116 evaluations (63.8 %), while Observer 2 did so in 81 out of 116 evaluations (69.8 %).

3.4. Agreement analysis

Agreement between observers and with the OAI-MS varied depending on the type of comparison and session. To reflect the ordinal nature of the three-category ordinal scale scoring system used for total asymmetry (0 = sound, 1 = mild to moderate asymmetry, 2 = severe asymmetry), Cohen's kappa coefficients were calculated using quadratic

weighting, which penalises larger score discrepancies more heavily.

3.4.1. Inter-rater agreement

The agreement between Observer 1 and Observer 2 showed marked variability across sessions. In session T0, $\kappa = 0.36$, corresponding to fair agreement, which indicates limited consistency between the two observers. Agreement decreased in session T1, with $\kappa = 0.18$ (poor agreement), and became negative in session T2 ($\kappa = -0.20$), reflecting disagreement worse than chance. Suggesting.

3.4.2. Inter-method agreement

Agreement between subjective evaluations and the OAI-MS also varied. For Observer 1, $\kappa = 0.13$ in T0, indicating poor agreement, while $\kappa = 0.47$ at T1, corresponding to moderate agreement. These values suggest more consistent recognition of asymmetries under repeated conditions. For Observer 2, the agreement with the OAI-MS was $\kappa = 0.15$ (poor agreement) in T0 and $\kappa = 0.30$ (fair agreement) in T1, both falling within the fair range. For this subset (10 horses; 20 repeated objective evaluations), the width of the 95 % confidence interval for κ was approximately ± 0.30, reflecting the limited sample size and indicating that only large repeatability effects could be reliably detected.

3.4.3. Intra-rater agreement

Intra-rater consistency was assessed by comparing live assessments at T0 with video-based reassessments at T2. Observer 1 showed a $\kappa = 0.23$ and Observer 2 $\kappa = 0.22$, both indicating fair agreement across sessions. Evaluators maintained only limited consistency when reassessing the same horses after a four-month interval.

3.4.4. Intra-system agreement

The repeatability of the OAI-MS between T0 and T1 was $\kappa = 0,43$, which is considered moderate agreement. This suggests a moderate level of consistency in the objective system's detection of asymmetries under field conditions, though variability between sessions remained evident.

3.4.5. Global agreement

In session T0, the overall agreement across all three evaluators yielded $\kappa = 0.15$. In session T1, the agreement was slightly higher, with $\kappa = 0.18$. Both values fall within the "slight agreement" category ($\kappa = 0.00–0.20$), highlighting the generally low level of concordance between observers and the objective system when considered together.

3.4.6. Agreement by anatomical region

Agreement between Observer 1 and Observer 2 was $\kappa = 0.00$ for forelimbs and $\kappa = 0.08$ for hindlimbs, both corresponding to slight agreement. For Observer 1 and the OAI-MS, agreement was fair in both regions, with $\kappa = 0.27$ for forelimbs and $\kappa = 0.21$ for hindlimbs. Observer 2 showed fair agreement with the OAI-MS in the forelimbs ($\kappa = 0.32$) and slight agreement in the hindlimbs ($\kappa = 0.17$). All results are reported in Table 6.

Table 6
Cohen's kappa values (quadratically weighted) were calculated separately for forelimbs (RF, LF) and hindlimbs (RH, LH), based on total asymmetry scores on a three-category ordinal scale (0 = sound, 1 = mild to moderate asymmetry, 2 = severe asymmetry). Each comparison is presented for Observer 1 vs Observer 2, and each observer vs the OAI-MS.

Comparison	Cohen kappa
Forelimbs – Observer 1 vs Observer 2	0.0035
Forelimbs – Observer 1 vs OAI-MS	0.2733
Forelimbs – Observer 2 vs OAI-MS	0.3210
Hindlimbs – Observer 1 vs Observer 2	0.0798
Hindlimbs – Observer 1 vs OAI-MS	0.2109
Hindlimbs – Observer 2 vs OAI-MS	−0.0312

3.4.7. Magnitude of discordance

The absolute differences between total asymmetry scores (three-category ordinal scale) were calculated for all pairwise comparisons. Exact agreement ($\Delta = 0$), as well as differences of one ($\Delta = 1$) or two points ($\Delta = 2$), were reported.

For inter-rater comparisons ($n = 116$), Observer 1 and Observer 2 agreed exactly in 64 evaluations (55.2 %), differed by one point in 51 cases (44.0 %), and by two points in 1 case (0.9 %).

In the inter-method comparisons with OAI-MS ($n = 68$ for each observer), Observer 1 matched the system exactly in 38 cases (55.9 %), with a one-point difference in 29 cases (42.6 %) and a two-point difference in 1 case (1.5 %). Observer 2 showed exact agreement with the system in 46 cases (67.6 %), and a one-point difference in 22 cases (32.4 %).

Intra-rater comparisons between T0 and T2 ($n = 48$) showed exact agreement in 24 cases (50.0 %) for Observer 1 and in 25 cases (52.1 %) for Observer 2. One-point differences were observed in 23 and 22 cases, respectively, with two-point differences occurring once in both comparisons (2.1 %).

The OAI-MS showed intra-system agreement ($n = 20$) between T0 and T1 with 11 exact matches (55.0 %) and 9 one-point differences (45.0 %) (Table 7).

4. Discussion

This study assessed the level of agreement between subjective evaluations performed by two experienced equine veterinarians and an OAI-MS in detecting gait asymmetries during lungeing in a population of traditional racehorses. The findings highlight the variability of visual lameness assessment when horses are trotted in a circle [15], and describe how a new, portable objective system [23] performs in this setting.

Overall agreement, expressed through Cohen's kappa coefficients with quadratic weighting, was low in most comparisons: inter-rater agreement ranged from $\kappa = 0.36$ at T0 to $\kappa = 0.18$ at T1 and $\kappa = -0.20$ at T2 (poor to fair); intra-rater agreement was fair, with $\kappa = 0.22–0.23$; and inter-method agreement was poor to moderate, ranging from $\kappa = 0.13$ to 0.47The poor intra-rater agreement observed at 4 months (T0 vs T2) should be interpreted with caution, as video-based reassessment introduces specific limitations, including reduced depth perception, restricted viewing angles, and the absence of real-time interaction with the horse. These factors likely contributed to the variability observed, and video evaluations cannot be fully equated with live clinical examinations.

In contrast, the OAI-MS showed moderate intra-system agreement (κ

Table 7
Magnitude of disagreement between evaluators and across sessions, calculated as the absolute difference (Δ) in total asymmetry scores (three-category ordinal scale): (0 = sound, 1 = mild to moderate asymmetry, 2 = severe asymmetry). For each pairwise comparison, the table reports the number and percentage of evaluations showing exact agreement ($\Delta = 0$), one-point differences ($\Delta = 1$), and two-point differences ($\Delta = 2$).

Comparison	n	Exact agreement ($\Delta = 0$)	$\Delta = 1$	$\Delta = 2$
Observer 1 vs Observer 2	116	64 (55.2 %)	51 (44.0 %)	1 (0.9 %)
Observer 1 vs OAI-MS	68	38 (55.9 %)	29 (42.6 %)	1 (1.5 %)
Observer 2 vs OAI-MS	68	46 (67.6 %)	22 (32.4 %)	0 (0.0 %)
Observer 1 T0 vs T2	48	24 (50.0 %)	23 (47.9 %)	1 (2.1 %)
Observer 2 T0 vs T2	48	25 (52.1 %)	22 (45.8 %)	1 (2.1 %)
OAI-MS T0 vs T1	20	14 (70.0 %)	6 (30.0 %)	0 (0.0 %)

= 0.43 between T0 and T1), which is consistent with previous studies reporting modest to moderate reliability of markerless AI systems under field conditions [18,26–28]. However, the time interval between assessments may have allowed for subtle changes in the horses' locomotor patterns, potentially contributing to variability in asymmetry measurements. Therefore, the observed level of agreement should be interpreted with caution, as it may reflect both technical repeatability and natural fluctuations in gait asymmetry over time.

While overall agreement values were low, most discordances were of small magnitude. Over 95 % of all mismatches involved a one-point difference on the adapted three-category ordinal scale (. This suggests that subjective evaluators often recognised similar asymmetry patterns but applied different scoring thresholds, reflecting the inherent variability and interpretative character of subjective lameness evaluation. In line with this, the κ value of 0.00 reported for forelimb agreement should not be interpreted as a complete absence of concordance, but rather because of the predominance of "sound" scores and the resulting low variability, which reduces the stability of κ estimates and makes them highly sensitive to small discrepancies. A one-point difference on the three-category scale is unlikely to alter the clinical outcome when the same evaluator consistently assesses the horse, as the decision to investigate a lameness would generally remain unchanged. However, such differences may still contribute to under-recognition of subtle or bilateral asymmetries, which could be clinically relevant in some contexts. In several previous studies, the highest discordance between clinicians and objective assessments has been reported in mild asymmetry categories, which reflects the known difficulty of detecting subtle gait abnormalities visually [13,15,27,32,33].

The biomechanics of lungeing must be considered when interpreting these results. In fact, it poses systematic biomechanical adaptations, such as body lean and asymmetric loading of the inner limbs, which can either resemble true lameness or mask subtle pathological signs [6–8,10,34]. Even clinically sound horses show consistent vertical asymmetries on the circle, particularly in the hindlimbs, and these adaptations may impact both visual and objective evaluation [9,35]. In our study, agreement was consistently higher for forelimbs than for hindlimbs, across all evaluator comparisons. Inter-method agreement between the OAI-MS and observers was fair for the forelimbs (κ = 0.27–0.32) and slight for the hindlimbs (κ = 0.17–0.21). This pattern aligns with previous studies reporting that forelimb lameness is more readily identified by clinicians due to the clearer correlation between asymmetry and head movement [28,33,36]. In contrast, interpreting pelvic motion requires attention to subtler displacement patterns, such as variations in tuber coxae rotation or vertical pelvic excursion, which are more prone to inter- and intra-observer variability [7,37,38].

As a result, hindlimb lameness, especially when mild or bilaterally distributed, is frequently under-recognised in subjective evaluations and may also be underestimated by objective systems [28,39].

The magnitude of disagreement was systematically explored. Observer 1 and Observer 2 showed exact agreement in 55 % of evaluations, with most of the remaining differences within a single-point range. When comparing with the OAI-MS, exact matches ranged from 56 % to 68 %. These findings suggest that scoring variability often reflects uncertainty at the border between "mild" and "moderate". This is particularly true when horses are evaluated on the circle, where dynamic and asymmetric movement complicate scoring consistency. These scoring differences are influenced by several factors, including the observer's clinical judgement, the viewing angle during evaluation [40], and when using automated systems, the way multiple asymmetry signals are integrated into a single classification output [18,26].

The variability and modest agreement levels observed in our study reflect challenges commonly reported in other AI-assisted diagnostic fields, where data heterogeneity, limited external validation, and difficulties in model interpretability undermine consistency and clinical adoption. In this context, the call for standardized benchmarking frameworks in AI is particularly relevant, as it highlights the importance of reliability, transparency, and reproducibility [25]. In equine medicine, this highlights the importance of aligning subjective evaluations with AI-based assessments to build confidence and support the clinical use of AI in gait analysis.

Previous research by the same group involving Palio horses assessed with the OAI-MS system during pre-race clinical evaluations, where over 80 % of horses considered fit to race exhibited detectable asymmetries, particularly during lungeing [41]. These findings highlight the possibility that some movement asymmetries observed during lungeing may reflect individual variation rather than overt pathology and support the use of objective tools as a complementary aid in clinical interpretation. The Palio horse population offers a particularly relevant model for this type of analysis, given their distinctive biomechanics, the challenges posed by uneven and non-standardised track surfaces, and the high physical demands associated with traditional racing [41,42].

A key strength of this study lies in its methodological consistency. The evaluations were standardized in surface, direction, duration, and video acquisition. Blinded video reassessment minimized recall and observer bias by preventing evaluators from being influenced by prior assessments or expectations [43]. Using a simplified three-category ordinal scale improved the clarity of interpretation and allowed for a more robust statistical evaluation by reducing noise from minor variations. The agreement analysis incorporated descriptive statistics at the limb level, comparison by anatomical region, and a global assessment using Fleiss' kappa. The overall Fleiss' κ values ranged from 0.15 to 0.18, indicating slight agreement among the three evaluators, findings that are in line with previous studies on visual lameness assessment [18,27,28].

This study has some limitations that should be acknowledged. Although the evaluations were blinded with respect to scores and system outputs, horse identity could not be concealed during live assessments, which may have influenced observer perception. The use of an aggregated ordinal scale, while helpful for analytical analysis and inter-method comparison, may have limited the ability to identify more subtle or complex asymmetry patterns. In addition, the assessments were carried out during routine pre-race inspections, which are primarily aimed at identifying obvious lameness. They did not involve a comprehensive lameness examination with flexion tests, nerve blocks, or diagnostic imaging. Although the duration and structure of the evaluations were equivalent to a standard visual assessment, the setting may have potentially influenced the evaluators' scoring approach. Finally, the T1 subset (10 horses; 40 subjective and 20 objective evaluations) provided only an exploratory estimate of short-term repeatability. With just 20 objective repeats, the 95 % CI half-width for κ was ~0.30, meaning that only large repeatability effects could be reliably detected. This limitation reflects a common challenge in AI-based diagnostic studies, where small datasets reduce generalizability [44].

The findings of this study support the growing role of objective gait analysis systems in the clinical evaluation of lameness, particularly when visual assessment is highly variable, such as in lungeing or in settings where multiple observers are involved. While not intended to replace clinical judgment, tools like the OAI-MS can help standardize evaluations and reduce subjectivity. This is particularly valuable not only for preventive purposes, such as identifying subtle but potentially relevant asymmetries in high-performance horses, but also in clinical contexts where disagreement or uncertainty arises during visual evaluation. It offers an objective reference that can assist in decision-making and safeguard all parties involved. In this sense, objective systems may contribute to greater transparency and confidence in veterinary evaluations, particularly in regulatory or competitive contexts.

Further studies should aim to better define the clinical significance of asymmetries detected by objective systems, distinguishing between normal individual variation and early indicators of pathologies. Evidence from human gait recognition has shown that relying on clearly defined and reproducible parameters improves stability and interpretability [24]. In line with this, our findings highlight the need for

standardized and consistently applied metrics in AI-based gait analysis to enhance comparability across studies and support clinical implementation. At the same time, assessing how these tools can be practically integrated into routine veterinary workflows will be essential to increase the transparency, reproducibility, and reliability of lameness assessment.

5. Conclusions

This study assessed the level of agreement between subjective evaluations and an OAI-MS in horses trotting during lungeing. Overall agreement was poor, with the lowest consistency observed in hindlimb assessments and in cases of mild asymmetry. Inter-observer variability was higher than intra-observer consistency, highlighting the persistent subjectivity of visual gait assessment and the absence of standardized scoring thresholds. The OAI-MS produced consistent results across repeated sessions, reinforcing its potential role as a practical tool to improve consistency and reduce subjective variability, particularly under field conditions. Rather than replacing clinical judgment, objective systems may provide additional support in cases of uncertainty or disagreement, resulting in more transparent and reliable assessment.

Defining precise, clinically significant thresholds and developing practical approaches for their use in routine lameness evaluations will be essential to more consistent, objective, and evidence-based decision-making.

Ethics in publishing statement

All authors confirm that the manuscript adheres to the highest standards of ethical publishing. No experimental procedures were conducted for the purposes of this study. All data were obtained during routine clinical veterinary examinations as part of standard pre-race inspections. Ethical approval was waived in accordance with the internal guidelines of the University of Bologna. Written informed consent was obtained from all horse owners prior to data collection and inclusion in the study. The authors affirm that the work is original, has not been published elsewhere, and is not under consideration for publication by any other journal.

CRediT authorship contribution statement

F. Meistro: Writing – review & editing, Writing – original draft, Formal analysis, Data curation, Conceptualization. **M.V. Ralletti:** Writing – review & editing, Writing – original draft, Visualization, Investigation. **R. Rinnovati:** Writing – review & editing, Validation, Supervision, Methodology, Investigation. **A. Spadari:** Writing – review & editing, Supervision, Project administration, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Baxter GM, Stashak TS. History, visual exam, and conformation. editor. In: Baxter GM, editor. Adams and stashak's lameness in horses. HobokenNJ: Wiley-Blackwell; 2020. p. 67–92.

[2] Oke SL, McIlwraith CW. Review of the economic impact of osteoarthritis and oral joint-health supplements in horses. In: AAEP Proceedings; 2010. p. 12–6.

[3] Ross MW. The lameness examination. In: Ross MW, Dyson SJ, editors. Diagnosis and management of lameness in the horse. 2nd ed. St. Louis: Saunders; 2011. p. 1–79.

[4] Hobbs SJ, Licka T, Polman R. The difference in kinematics of horses walking, trotting and cantering on a flat and banked 10 m circle. Equine Vet J 2011;43(6):686–94. https://doi.org/10.1111/j.2042-3306.2010.00334.x.

[5] Clayton HM, Sha DH. Head and body centre of mass movement in horses trotting on a circular path. Equine Vet J 2006;38(5):462–7. https://doi.org/10.1111/j.2042-3306.2006.tb05588.x.

[6] Robartes H, Fairhurst H, Pfau T. Head and pelvic movement symmetry in horses during circular motion and in rising trot. Vet J 2013;198(1):e33–8. https://doi.org/10.1016/j.tvjl.2013.09.033.

[7] Rhodin M, Pfau T, Roepstorff L, Egenvall A. Effect of lungeing on head and pelvic movement asymmetry in horses with induced lameness. Vet J 2013;198(1):e39–43. https://doi.org/10.1016/j.tvjl.2013.09.031.

[8] Denoix JM. A look at lameness through the eyes of functional anatomy (and biomechanics). In: Proceedings of the 67th Annual Convention of the AAEP; 2021. p. 106–33.

[9] Starke SD, Willems E, May SA, Pfau T. Vertical head and trunk movement adaptations of sound horses trotting in a circle on a hard surface. Vet J 2012;193(1):73–80. https://doi.org/10.1016/j.tvjl.2011.10.019.

[10] Macaire C, Hanne-Poujade S, De Azevedo E, Denoix JM, Coudry V, Jacquet S, et al. Asymmetry thresholds reflecting the visual assessment of forelimb lameness on circles on a hard surface. Animals (Basel) 2023;13(21):3327. https://doi.org/10.3390/ani13213319.

[11] Keegan KG. Objective measures of lameness evaluation. In: Proceedings of the American Association of Equine Practitioners; 2012. p. 127–31.

[12] Donnell JR, Frisbie DD, King MR, Goodrich LR, Haussler KK. Comparison of subjective lameness evaluation, force platforms and an inertial-sensor system to identify mild lameness in an equine osteoarthritis model. Vet J 2015;206(2):136–42. https://doi.org/10.1016/j.tvjl.2015.08.004.

[13] Keegan KG, Dent EV, Wilson DA, Janicek J, Kramer J, Lacarrubba A, et al. Repeatability of subjective evaluation of lameness in horses. Equine Vet J 2010;42(2):92–7. https://doi.org/10.2746/042516409X479568.

[14] American Association of Equine Practitioners (AAEP). Guide to veterinary services for horse shows. Lexington (KY): AAEP; 1999.

[15] Hammarberg M, Egenvall A, Pfau T, Rhodin M. Rater agreement of visual lameness assessment in horses during lungeing. Equine Vet J 2016;48(1):78–82. https://doi.org/10.1111/evj.12385.

[16] Keegan KG, Kramer J, Yonezawa Y, Maki H, Frank Pai P, Dent EV, et al. Assessment of repeatability of a wireless, inertial sensor-based lameness evaluation system for horses. Am J Vet Res 2011;72(9):1156–63. https://doi.org/10.2460/ajvr.72.9.1156.

[17] McCracken MJ, Kramer J, Keegan KG, Lopes M, Wilson DA, Reed SK, et al. Comparison of an inertial sensor system of lameness quantification with subjective lameness evaluation. Equine Vet J 2012;44(6):652–6. https://doi.org/10.1111/j.2042-3306.2012.00571.x.

[18] Calle-González N, Lo Feudo CM, Ferrucci F, Requena F, Stucchi L, Muñoz A. Objective assessment of equine locomotor symmetry using an inertial sensor system and artificial intelligence: a comparative study. Animals (Basel) 2024;14(6):1121. https://doi.org/10.3390/ani14060921.

[19] Marshall JF, Lund DG, Voute LC. Use of a wireless, inertial sensor-based system to objectively evaluate flexion tests in the horse. Equine Vet J 2012;44(1):8–11. https://doi.org/10.1111/j.2042-3306.2012.00611.x.

[20] McCracken MJ, Kramer J, Keegan KG, Lopes M, Wilson DA, Reed SK, et al. Comparison of an inertial sensor system of lameness quantification with subjective lameness evaluation. Equine Vet J 2012;44(6):652–6. https://doi.org/10.1111/j.2042-3306.2012.00571.x.

[21] Pfau T, Landsbergen K, Davis BL, Kenny O, Kernot N, Rochard N, et al. Comparing inertial measurement units to markerless video analysis for movement symmetry in quarter horses. Sensors (Basel) 2023;23(20):8414. https://doi.org/10.3390/s23208414.

[22] Lawin FJ, Byström A, Roepstorff C, Rhodin M, Almlöf M, Silva M, et al. Is markerless more or less? Comparing a smartphone computer vision method for equine lameness assessment to multi-camera motion capture. Animals (Basel) 2023;13(3):383. https://doi.org/10.3390/ani13030390.

[23] Sleip AI AB. Sleip – scientifically validated equine gait analysis app – tutorial and guides. Uppsala, Sweden: Sleip AI AB; 2023. Available from, https://www.sleip.com.

[24] Awad KM, Tulaib LF, Saleh HM. Gait recognition by computing fixed body parameters. Babylonian J Network 2024;2024:191–7. https://doi.org/10.58496/bjn/2024/019.

[25] Sallam M, Khalil R, Sallam M. Benchmarking generative AI: a call for establishing a comprehensive framework and a generative AIQ test. Mesopotamian J Artif Intell Healthcare 2024;2024:69–75. https://doi.org/10.58496/mjaih/2024/010.

[26] Kallerud AS, Marques-Smith P, Bendiksen HK, Fjordbakk CT. Objective movement asymmetry in horses is comparable between markerless technology and sensor-based systems. Equine Vet J 2024;57(1):115–25. https://doi.org/10.1111/evj.14089.

[27] de Chiara M, Montano C, De Matteis A, Guidi L, Buono F, Auletta L, et al. Agreement between subjective gait assessment and markerless video gait-analysis in endurance horses. Equine Vet J 2025:1–8. https://doi.org/10.1111/evj.14516.

[28] McPeek JL, Menarim BC, Sponseller B, Adams A, McClendon M, Page AE. Agreement between multiple objective and subjective equine lameness evaluators. J Equine Vet Sci 2025;148:105451. https://doi.org/10.1016/j.jevs.2025.105451.

[29] Gabriele SBT, Valazza A, Parrilli A, Fini M, Pagliara E, Putame G, et al. Traditional equestrian events in Italy: results of the official veterinary controls carried out during the period 2017–2023. Large Anim Rev 2024;30(6):281–90.

[30] Dyson S. Recognition of lameness: man versus machine. Vet J 2014;201(3):245–8. https://doi.org/10.1016/j.tvjl.2014.05.018.

[31] Dyson S. Can lameness be graded reliably? Equine Vet J 2011;43(4):379–82. https://doi.org/10.1111/j.2042-3306.2011.00391.x.

[32] Keegan KG, Wilson DA, Wilson DJ, Smith B, Gaughan EM, Scott R, 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 1997;58(4):414–8.
[33] 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 (3):326–31. https://doi.org/10.1111/evj.13144.
[34] Rhodin M, Roepstorff L, French A, Keegan KG, Pfau T, Egenvall A. Head and pelvic movement asymmetry during lungeing in horses with symmetrical movement on the straight. Equine Vet J 2016;48(3):315–20. https://doi.org/10.1111/evj.12446.
[35] Pfau T, Jennings C, Mitchell H, Olsen E, Walker A, Egenvall A, et al. Lungeing on hard and soft surfaces: movement symmetry of trotting horses considered sound by their owners. Equine Vet J 2016;48(1):83–9. https://doi.org/10.1111/evj.12374.
[36] Dyson SJ. The clinician's eye view of hindlimb lameness in the horse: technology and cognitive evaluation. Equine Vet J 2009;41(2):99–100. https://doi.org/10.2746/042516409X399963.
[37] Greve L, Pfau T, Dyson S. Thoracolumbar movement in sound horses trotting in straight lines in hand and on the lunge and the relationship with hind limb symmetry or asymmetry. Vet J 2017;220:95–104. https://doi.org/10.1016/j.tvjl.2017.01.003.
[38] Rhodin M, Persson-Sjodin E, Egenvall A, Serra Bragança FM, Pfau T, Roepstorff L, et al. Vertical movement symmetry of the withers in horses with induced forelimb and hindlimb lameness at trot. Equine Vet J 2018;50(6):818–24. https://doi.org/10.1111/evj.12844.
[39] van Weeren PR, Pfau T, Rhodin M, Roepstorff L, Serra Bragança F, Weishaupt MA. Do we have to redefine lameness in the era of quantitative gait analysis? Equine Vet J 2017;49(5):567–9. https://doi.org/10.1111/evj.12715.
[40] Greve L, Dyson S. What can we learn from visual and objective assessment of non-lame and lame horses in straight lines, on the lunge and ridden? Equine Vet Educ 2020;32(9):479–91. https://doi.org/10.1111/eve.13016.
[41] Meistro F, Ralletti MV, Rinnovati R, Spadari A. Objective evaluation of gait asymmetries in traditional racehorses during pre-race inspection: application of a markerless AI system in straight-line and lungeing conditions. Animals (Basel) 2025;15(12):1797. https://doi.org/10.3390/ani15121797.
[42] Pagliara E, Pasinato A, Valazza A, Riccio B, Cantatore F, Terzini M, et al. Multibody computer model of the entire equine forelimb simulates forces causing catastrophic fractures of the carpus during a traditional race. Animals (Basel) 2022;12(6):607. https://doi.org/10.3390/ani12060737.
[43] Hróbjartsson A, Thomsen ASS, Emanuelsson F, Tendal B, Hilden J, Boutron I, et al. Observer bias in randomized clinical trials with measurement scale outcomes: a systematic review of trials with both blinded and nonblinded assessors. CMAJ 2013;185(4):E201–11. https://doi.org/10.1503/cmaj.120744.
[44] Hamad RH. A systematic review of artificial intelligence's function in the diagnosis of lung cancer (2018–2024). Mesopotam J Artif Intell Healthc 2025;2025:12–25. https://doi.org/10.58496/mjaih/2025/002.