European Journal of Radiology 183 (2025) 111911

Contents lists available at ScienceDirect

European Journal of Radiology

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

Research article

Validation of AI-driven measurements for hip morphology assessment

Louis Lassalle [a,b,c,*], Nor-eddine Regnard [a,b,c], Marion Durteste [c], Jeanne Ventre [c], Vincent Marty [c], Lauryane Clovis [c], Zekun Zhang [c], Nicolas Nitche [c], Alexis Ducarouge [c], Alexia Tran [d], Jean-Denis Laredo [c,e,f,g], Ali Guermazi [h]

[a] Réseau Imagerie Sud Francilien, Lieusaint, France
[b] Ramsay Santé, Clinique du Mousseau, Evry, France
[c] Gleamer, Paris, France
[d] Hôpital Fondation Adolphe de Rothschild, Paris, France
[e] Service de Radiologie, Institut Mutualiste Montsouris, Paris, France
[f] Laboratoire (B3OA) de Biomécanique et Biomatériaux ostéo-articulaires, Faculté de Médecine Paris-Cité, Paris, France
[g] Professeur Émerite d'Imagerie Médicale, Université Paris-Cité, Paris, France
[h] Department of Radiology, VA Boston Healthcare System, Boston University School of Medicine, Boston, MA, USA

ARTICLE INFO

Keywords:
Pelvic and hip measurements
Artificial intelligence
Radiography
Anteroposterior view
False profile view

ABSTRACT

Rationale and Objectives: Accurate assessment of hip morphology is crucial for the diagnosis and management of hip pathologies. Traditional manual measurements are prone to mistakes and inter- and intra-reader variability. Artificial intelligence (AI) could mitigate such issues by providing accurate and reproducible measurements. The aim of this study was to compare the performance of BoneMetrics (Gleamer, Paris, France) in measuring pelvic and hip parameters on anteroposterior (AP) and false profile radiographs to expert manual measurements.
Materials and Methods: This retrospective study included AP and false profile pelvic radiographs collected from private practices in France. Pelvic and hip measurements included the femoral neck shaft angle, lateral center edge angle, acetabular roof angle, pelvic obliquity, and vertical center anterior angle. AI measurements were compared to a ground truth established by two expert radiologists. Performance metrics included mean absolute error (MAE), Bland-Altman analysis, and intraclass correlation coefficients (ICC).
Results: AI measurements were performed on AP views from 88 patients and on false profile views from 60 patients. They demonstrated high accuracy, with MAE values inferior to 0.5 mm for pelvic obliquity and inferior to 4.2° for all pelvic angles on both views. ICC values indicated good to excellent agreement between AI measurements and the ground truth (0.78–0.99). Notably, no significant differences were found in AI measurement accuracy between patients with normal and abnormal acetabular coverage.
Conclusion: The application of AI in measuring pelvic and hip parameters on AP and false profile radiographs demonstrates promising outcomes. The results reveal that these AI-powered measurements provide accurate estimations and show strong agreement with expert manual measurements. Ultimately, the use of AI has the potential to enhance the reproducibility of measurements as part of comprehensive hip assessments, thereby improving diagnostic accuracy.

1. Introduction

The hip is a complex three-dimensional joint whose primary weight-bearing and load distribution functions allow for the body's stability and mobility. Radiographic examination is the gold standard for assessing the anatomical morphology of the joint which is required in both pre- and post-operative monitoring of hip surgeries, as well as for diagnosis of hip pathology [1]. Timely detection of structural hip disorders is of paramount importance in the young adult population for effective pain management and the formulation of adequate treatment strategies [2]. For example, conditions characterized by abnormal acetabular rim loading, such as hip dysplasia and femoroacetabular impingement, can precipitate chondrolabral damage and increase the risk of early osteoarthritis if left untreated [3–7].

* Corresponding author.
E-mail address: llassalle@risf.fr (L. Lassalle).

https://doi.org/10.1016/j.ejrad.2024.111911
Received 11 September 2024; Received in revised form 12 December 2024; Accepted 30 December 2024
Available online 31 December 2024

Conventional pelvic radiographs are the preferred imaging modality for visualizing the relationship between the pelvis and the proximal femur. This is primarily achieved using standing anteroposterior (AP) views and Lequesne's false profile views [1,8], hereafter referred to as false profile radiographs. Multiple radiographic measurements have been used as indicators of abnormal hip joint morphology. The most frequently examined are the lateral center edge angle (LCEA) [9,10], femoral neck shaft angle (NSA) [11], acetabular roof angle or Tönnis angle [12,13], and pelvic obliquity [14] on AP radiographs. Recent advancements highlight the importance of integrating measurements from false profile pelvic views, such as the vertical center anterior angle (VCA) [15], to achieve a more comprehensive evaluation of femoral head coverage [16].

The accurate placement of key points is critical to obtain hip measurements and to diagnose structural hip abnormalities in a reproducible and standardized manner. However, manual measurements are prone to error, as demonstrated by the substantial inter-reader variability reported in the literature [17,18]. Reliability ranges from poor to moderate with the highest correlation coefficients typically seen in measurements of the LCEA and the lowest in measurements of the VCA. Furthermore, measuring hip parameters is a time-consuming and laborious task for clinicians, adding to the complexity of diagnosing hip abnormalities [19].

Given the limitations associated with manual readings, artificial intelligence (AI) emerges as a robust alternative. The implementation of AI in various aspects of pelvic and hip health has already demonstrated its efficacy in diagnosing developmental dysplasia of the hip in children [20], detecting hip fractures [21–23], and assessing hip osteoarthritis [24]. Data on AI-powered measurements are still sparse, two commercial AI software solutions have revealed promising results in automating measurements on pelvic radiographs [19,25–27]. Moreover, the effectiveness of AI in analyzing views other than the standard AP view remains unexplored.

This study aimed to assess the performance of BoneMetrics, a CE-certified commercial AI solution (Gleamer, Paris, France), in measuring pelvic and hip parameters on AP and false profile views on a real-world consecutive dataset. The primary objectives were to compare the precision of AI-powered measurements with a ground truth established by two senior musculoskeletal radiologists and to evaluate inter- and intra-reader variability. The secondary objectives of this study were three-fold: (1) to evaluate the AI's performance on AP views and, notably, on false profile views, which have never been investigated in prior research; (2) to analyze measurement variability across different degrees of acetabular coverage; and (3) to broaden the scope of evaluation for BoneMetrics (Gleamer, Paris) to include additional body parts, specifically the pelvic and hip regions.

2. Materials and methods

2.1. Study design

The present study evaluated the performance of AI-driven automated pelvic and hip measurements against a reference standard established by two expert radiologists, referred to as the ground truth. Acquisition of consecutive patient data was conducted retrospectively from three participating institutions. Each institution informed patients about the use of their anonymized data for research purposes and provided explicit instructions on how to decline participation.

2.2. Study population

Consecutive pelvic radiographs acquired between January 2015 and February 2018 were obtained from three private practices in France. A custom-built natural language processing (NLP) algorithm was used to search radiologists' reports within these databases for AP and false profile views acquired to measure pelvic and hip parameters. The NLP was designed to detect specific textual patterns related to measurements on weight-bearing pelvic radiographs using tailored regular expressions related to pelvic and hip measurements (e.g., "measurement(s)", "angle (s)", "degree(s)", "lequesne", "false profile"). Data on patient sex and age were extracted from the DICOM tags. Inclusion criteria were AP and false profile pelvic radiographs involving patients over the age of 10 years old. Radiographs were excluded based on the following predefined criteria: image not adhering to quality standards, incorrect views, non-weight-bearing patient, measurements visible on the image, and rejection by the AI algorithm. In addition, any images showing hip prostheses were discarded as the AI software does not generate measurements under such conditions.

2.3. Radiologists' manual measurements

To establish the ground truth, a two-phase process was employed. Initially, an expert musculoskeletal radiologist with 35 years of experience (JDL) reviewed the radiographs, excluding those that did not meet the predefined criteria, and placed the key points on the accepted images. In the second phase, these radiographs were independently annotated by another musculoskeletal radiologist with 7 years of experience (AT). The ground truth was defined as the mean of their measurements, in line with similar studies [28–30]. Annotation was performed on Kili, a dedicated web-based platform equipped with tools to facilitate precise labeling such as zoom, pan, contrast adjustment, and a circle drawer. The labeling task involved positioning key points on each radiograph, which were subsequently used to compute the pelvic and hip parameters. Specifically, on AP pelvic radiographs, the annotators identified the center of the femoral head, the center of the femoral neck, the center of the femoral proximal diaphysis, the center of the femoral distal diaphysis, the top of the acetabular roof, the lateral edge of the acetabular roof, and the medial edge of the acetabular roof (**Fig. A.1**). On false profile views, key points were placed at the center of the femoral head and at the anterior-most aspect of the acetabulum. To assess intra-reader reliability, one of the two annotators who established the ground truth re-annotated random subsamples of 28 AP and 29 false profile pelvic radiographs after a 1-month washout period. Furthermore, inter-reader reliability was evaluated by having two other radiologists with 13 and 12 years of experience label these subsamples (NER, LL).

2.4. Anatomic definitions of pelvic and hip parameters

Measurements of interest on AP pelvic radiographs were the femoral NSA or caput-collum-diaphyseal angle, the LCEA, the acetabular roof angle, and the pelvic obliquity (**Fig. A.1**) [19,25]. For all measurements but pelvic obliquity, the measurements were taken on the left and right sides independently. Hip deformities such as coxa vara and coxa valga were defined as a femoral NSA $< 120^\circ$ and a femoral NSA $> 140^\circ$, respectively. On AP pelvic radiographs, normal coverage was defined as a LCEA between 21° and 33°, overcoverage as a LCEA $> 33^\circ$, and undercoverage as a LCEA $< 21^\circ$ [26]. On false profile pelvic radiographs, normal coverage was defined as having a VCA between 0° and 10°, overcoverage as a VCA $> 10^\circ$ and undercoverage as a VCA $< 0^\circ$ [1].

2.5. AI software

Automated analyses were performed by the AI-powered software BoneMetrics (version 2.3.1, Gleamer, Paris, France). BoneMetrics is a CE-certified image processing tool that automates musculoskeletal measurements on conventional radiographs and EOS images. The algorithm was trained on over 5,000 images collected from more than 20 medical centers across Europe. The images were annotated by ten radiographers and radiologists who were distinct from the readers involved in the present study. To ensure the quality of the training data, an expert musculoskeletal radiologist with 14 years of experience reviewed all manual annotations.

BoneMetrics relies on multiple convolutional neural networks and leverages diverse architectures such as a top-down model implemented with detectron2, a lightweight HRNet (litehrnet), and a bottom-up approach. The AI first executes a key point detection task, wherein it identifies anatomical landmarks as points of interest and assigns a confidence score ranging from 0 to 100 to each detected point. It then uses the points that exceed a predefined threshold of 50 to compute angles and lengths according to established clinical measurement protocols. Key points scoring below this threshold are discarded, preventing any comparison between the AI and the ground truth for the corresponding measurement.

2.6. Statistical analyses

The sample size for this study was determined based on calculations from a similar study [27]. The authors computed the sample sizes required to evaluate automated measurements against manual expert measurements of the pelvis and the hip, using Bland-Altman analyses. At a significance level of 5 % and a power of 80 %, they established the required sample size to be 176 individual hips.

Left and right hips were considered independently for all analyses except for pelvic obliquity which yields a unique value per patient. Root mean square error (RMSE) and mean absolute error (MAE) were computed for each pelvic or hip parameter to assess the performance of the AI software. Bootstrapped 95 % confidence intervals (n = 1000 samples) were calculated, and patient resampling was applied to all parameters but pelvic obliquity in order to account for data dependencies. Dependencies within the dataset relate to left and right measurements from a single radiograph being considered independently. The MAE was also computed according to age, sex, and the degree of acetabular coverage. Differences between patients over and under 60 years old, male and female patients, and normal and abnormal acetabular coverage were assessed using the Mann-Whitney *U* test for pelvic obliquity and linear mixed models with patient as a random intercept for all other parameters. Bonferroni correction for multiple comparisons was applied, setting statistical significance at $p = 0.01$ (= 0.05 / 5 measurements).

Bland-Altman analyses were conducted to assess the agreement between the AI algorithm and the ground truth. While the conventional Bland-Altman method was employed for pelvic obliquity, a mixed effects approach was applied to all other parameters to account for data dependencies [31]. A mixed effects regression model was therefore implemented to calculate limits of agreement, and it modeled patient as a random intercept and side of the measurement (left, right) as a fixed effect. To further explore the performance of the AI software in comparison with the ground truth, intraclass correlation coefficients (ICC) between the two were computed based on a two-way mixed effects model with absolute agreement for multiple raters. The ICC values contrasting the AI with the ground truth were compared statistically to the ICC between the two radiologists who established the ground truth with a z-test using Fisher's Z transformation.

To address the question of inter-reader reliability, there were the two expert radiologists who established the ground truth and two other independent radiologists who annotated a subsample of the dataset. ICC values between all four readers were calculated based on two-way random effects models with absolute agreement for multiple raters. Moreover, the MAE of each of the two independent radiologists was computed for each pelvic and hip parameter. It was compared to the MAE of the AI using Mann-Whitney U tests. Bonferroni correction for multiple comparisons was applied, setting statistical significance at $p = 0.01$ (= 0.05 / 5 measurements). The intra-reader reliability was calculated on a subset of radiographs using a two-way mixed effects model with absolute agreement for a single rater [32]. ICC values were interpreted as follows: poor reliability, ICC < 0.5; moderate reliability, $0.5 \leq$ ICC < 0.75; good reliability, $0.75 \leq$ ICC < 0.9; excellent reliability, ICC $\geq$ 0.9 [32]. Statistical analyses were conducted using R (v4.3.2) in RStudio (v2023.09.1 + 494) with the "irr", "nlme", "boot", and "psych" packages. This study received approval from the Institutional Review Board (ethics approval number CRM-2209–306).

3. Results

3.1. Dataset description

Overall, 88 AP pelvic radiographs (176 individual hips) from 88 patients and 80 false profile pelvic radiographs (88 individual hips) from another 60 patients were included in the final analysis (Table 1). The discrepancy between the number of hips and the number of false profile views can be explained by the combination of both left and right false profile views into single images for 8 patients. Following the first phase of establishing the ground truth, 21 AP pelvic radiographs (18.4 %) and 27 false profile pelvic radiographs (24.8 %) were excluded. In total, 10 radiographs (4.5 %) were excluded due to visible measurements, 11 (4.9 %) due to poor quality, 25 (11.2 %) due to a wrong pelvic view, 1 (0.4 %) due to the presence of a prosthesis and 1 (0.4 %) due to the patient being non-weight-bearing (Fig. 1). A further 5 AP (2.2 %) and two false profile (0.9 %) pelvic radiographs were discarded as the AI algorithm didn't yield any measurement. The absence of measurements from the AI was due to the presence of hip implants, incorrect view acquisition, or poor image quality.

Patient data were sourced from three centers, with 44 patients from Center 1 (29.7 %), 54 from Center 2 (36.5 %), and 50 from Center 3 (33.8 %). Additionally, the radiographs were obtained from 9 different manufacturers. In the final dataset, there were 60 women (68.2 %) and 28 men (31.8 %) with AP pelvic radiographs, and their mean age was 59.0 years (± 17 years). Likewise, there were 50 women and 30 men with false profile pelvic radiographs, and their mean age was 63.9 years (± 14.3 years). Ages ranged from 14 to 90 years. The pelvic radiographs were acquired for various clinical indications, including pain, trauma, post-operative follow-up and routine monitoring (Table 1).

Hips with different degrees of acetabular coverage were included. Among both AP and false profile pelvic radiographs, there were 10 images of patients with acetabular undercoverage (6.0 %), 94 images of

Table 1
Demographic and clinical characteristics of patients.

	Anteroposterior pelvic radiographs	False profile pelvic radiographs
Sample size (n)	88	80
Unique patients (n)	88	60
Patient age		
Mean ± SD (y)	61.9 ± 16.0	63.9 ± 14.3
Range (y)	[14.0 – 90.0]	[30.0 – 90.0]
Sex		
Women (%)	60 (68.2 %)	50 (62.5 %)
Men (%)	28 (31.8 %)	30 (37.5 %)
Hip alignment		
Femoral NSA < 120° (%)	6 (6.8 %)	
Femoral NSA [120° – 130°[(%)	54 (61.4 %)	
Femoral NSA [130° – 140°[(%)	27 (30.7 %)	
Femoral NSA ≥ 140° (%)	1 (1.1 %)	
Acetabular coverage		
Under coverage (%)	3 (3.4 %)	7 (8.75 %)
Normal coverage (%)	36 (40.9 %)	58 (72.5 %)
Over coverage (%)	49 (55.7 %)	15 (18.75 %)
Clinical indication		
Pain (%)	29 (33.0 %)	33 (55.0 %)
Trauma (%)	17 (19.3 %)	15 (25.0 %)
Post-operative (%)	3 (3.4 %)	0 (0.0 %)
Routine monitoring (%)	4 (4.5 %)	0 (0.0 %)
Other (%)	35 (39.8 %)	12 (20.0 %)

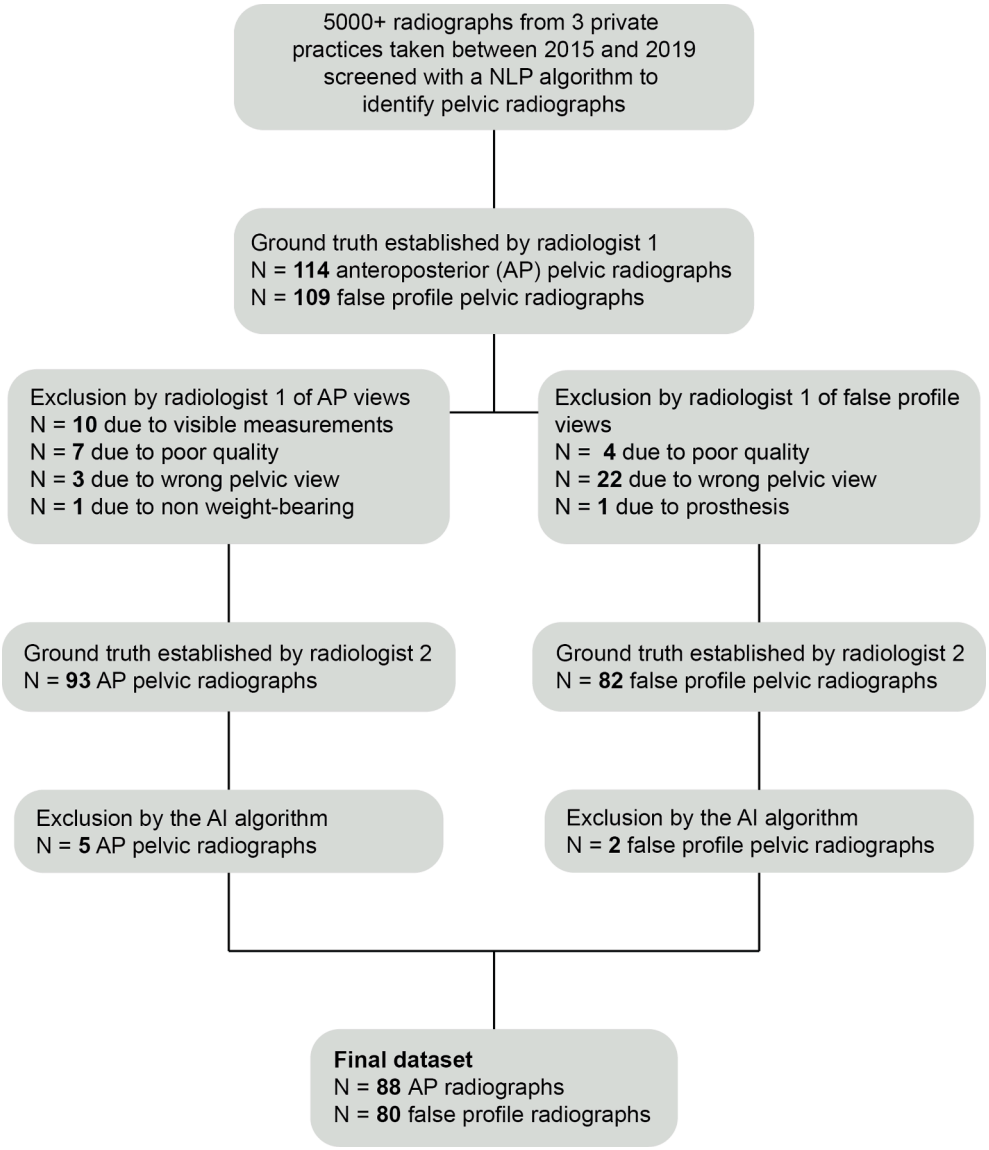

Fig. 1. Flowchart outlining the step-by-step process used to constitute the final dataset.

patients with normal coverage (56.0 %) and 64 images of patients with overcoverage (38.0 %). The majority of radiographs were from patients who did not exhibit hip deformities (95.8 %). There was, however, some variability, with 54 patients exhibiting femoral NSA values between 120° and 130° and 27 patients displaying values between 130° and 140°. Six patients could be classified as having coxa vara (6.8 %) and one as having coxa valga (1.1 %).

3.2. AI software measurement performance

In total, 32 measurements were missing and thus not considered in the analyses. Specifically, 7 hips lacked measurements for the femoral NSA (3.9 %), 6 hips for the LCEA (3.4 %), 10 hips for the acetabular roof angle (5.6 %), and 10 hips for pelvic obliquity (5.6 %). Measurements for the VCA on false profile pelvic radiographs were successfully obtained for all 88 hips (**Table A.1**). Missing measurements were primarily due to the presence of a hip prosthesis, as the AI does not generate results in such instances. In a few cases, the absence of pixel spacing information in the DICOM tags prevented the algorithm from processing pelvic obliquity.

First, the RMSE and MAE values for each pelvic and hip parameter were examined (Table 2). Notably, the acetabular roof angle on AP pelvic radiographs demonstrated the lowest errors with a RMSE of 2.40° (95 % CI [2.16, 2.62]) and a MAE of 1.79° (95 % CI [1.60, 1.97]). Discrepancies between the AI software and the ground truth were also minimal for pelvic obliquity with a RMSE of 0.57 mm (95 % CI [0.46, 0.64]) and a MAE of 0.42 mm (95 % CI [0.36, 0.49]). In contrast, the largest discrepancies were noted for the VCA on false profile pelvic radiographs with a RMSE of 5.61° (95 % CI [4.60, 6.40]) and a MAE of 4.16° (95 % CI [3.60, 4.73]).

Table 2
Performance assessment of the AI algorithm on pelvic radiographs.

Angles and lengths	N	RMSE [95 % CI]	MAE [95 % CI]
Anteroposterior radiographs			
Femoral NSA (°)	169	3.58 [3.17, 3.92]	2.86 [2.61, 3.11]
LCEA (°)	170	3.24 [2.88, 3.52]	2.58 [2.34, 2.81]
Acetabular roof angle (°)	166	2.40 [2.16, 2.62]	1.79 [1.60, 1.97]
Pelvic obliquity (mm)	78	0.57 [0.46, 0.64]	0.42 [0.36, 0.49]
False profile radiographs			
VCA (°)	88	5.61 [4.60, 6.40]	4.16 [3.60, 4.73]

Second, Bland-Altman analyses were conducted, and results are summarized in Fig. 2. To further evaluate the deviation of the AI

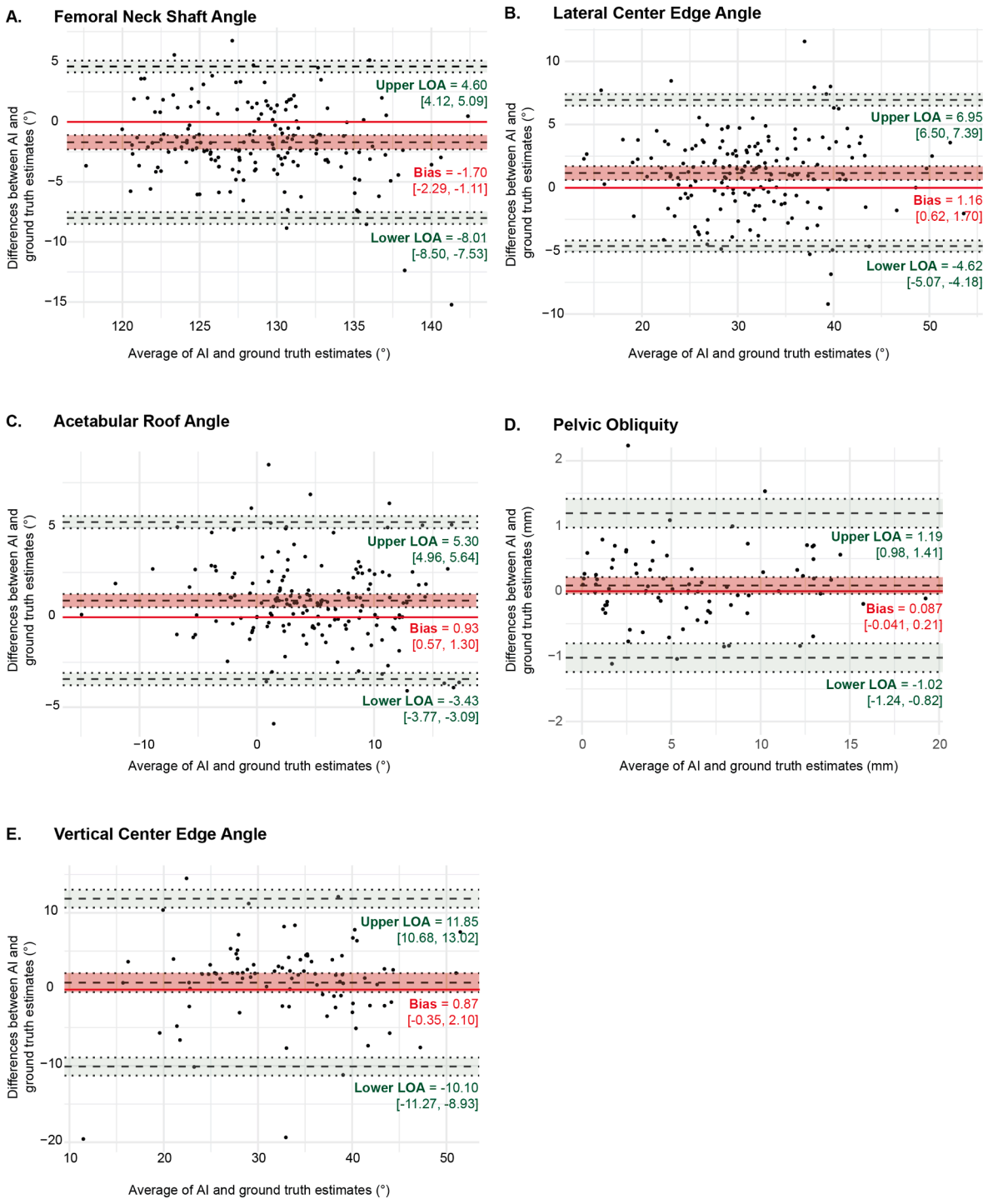

Fig. 2. Comparison of AI predictions and ground truth measurements using Bland-Altman plots. The latter are displayed for the femoral neck shaft angle (A), the lateral center edge angle (B), the acetabular roof angle (C) and the pelvic obliquity (D) on anteroposterior views, and the vertical center edge angle (E) measured on false profile views. The red line highlights perfect agreement between the AI and the ground truth. The area shaded in red indicates the 95 % confidence interval, that is centered around the black dotted mean difference line. The black dotted lines at the boundaries of the plots mark the upper and lower limits of agreement along with their 95 % confidence intervals in light green.

measurements from the ground truth, the ICC was assessed for all pelvic and hip parameters (Table 3). The agreement was found to be good for the femoral NSA (0.81, 95 % CI [0.75, 0.86]) and VCA (0.78, 95 % CI [0.69, 0.85]) and excellent for the LCEA (0.92, 95 % CI [0.89, 0.94]), acetabular roof angle (0.93, 95 % CI [0.91, 0.95]), and pelvic obliquity (0.99, 95 % CI [0.99, >0.99]). The agreement between the AI and the ground truth and the agreement between the two radiologists who established the ground truth were statistically equivalent for the femoral NSA ($p = 0.72$), the pelvic obliquity ($p = 0.074$) and the VCA ($p = 0.67$). The agreement between the AI and the ground truth was significantly higher than the agreement between the two radiologists for the LCEA ($p < 0.001$) and the acetabular roof angle ($p < 0.001$).

3.3. *AI performance across age, sex, and various hip morphologies*

The MAE between AI and ground truth measurements was first

Table 3
Results of the agreement analyses.

Lengths and angles	ICC between AI and ground truth [95 % CI]	ICC between expert radiologists [95 % CI]	Statistical difference
Anteroposterior radiographs			
Femoral NSA	0.81 [0.75, 0.86]	0.80 [0.73, 0.85]	$p = 0.72$
LCEA	0.92 [0.89, 0.94]	0.73 [0.38, 0.86]	$p < 0.001$*
Acetabular roof angle	0.93 [0.91, 0.95]	0.80 [0.50, 0.90]	$p < 0.001$*
Pelvic obliquity	0.99 [0.99, >0.99]	>0.99 [0.99, >0.99]	$p = 0.074$
False profile radiographs			
VCA	0.78 [0.69, 0.85]	0.76 [0.65, 0.83]	$p = 0.67$
	Intra-reader reliability [95 % CI]	ICC between all four radiologists [95 %CI]	
Anteroposterior radiographs			
Femoral NSA	0.86 [0.78, 0.92]	0.68 [0.51, 0.79]	
LCEA	0.89 [0.82, 0.93]	0.62 [0.41, 0.75]	
Acetabular roof angle	0.95 [0.91, 0.97]	0.63 [0.48, 0.75]	
Pelvic obliquity	>0.99 [0.99, >0.99]	>0.99 [0.99, >0.99]	
False profile radiographs			
VCA	0.96 [0.92, 0.98]	0.75 [0.56, 0.87]	

Table 4
Performance of the AI algorithm across hip morphologies.

Lengths and angles	MAE [95 % CI]		Statistical test
	Abnormal coverage	**Normal coverage**	
Anteroposterior radiographs			
Femoral NSA (°)	2.78 [2.40, 3.13]	2.99 [2.58, 3.38]	$F(1, 86) = 1.02$, $p = 0.32$
LCEA (°)	2.86 [2.49, 3.18]	2.16 [1.80, 2.49]	$F(1, 86) = 0.63$, $p = 0.43$
Acetabular roof angle (°)	1.89 [1.62, 2.15]	1.64 [1.38, 1.90]	$F(1, 85) = 0.59$, $p = 0.44$
Pelvic obliquity (mm)	0.39 [0.33, 0.45]	0.47 [0.33, 0.60]	$W = 707$, $p = 0.83$
False profile radiographs			
VCA (°)	4.94 [3.79, 6.08]	3.81 [3.17, 4.49]	$F(1, 27) = 1.10$, $p = 0.30$

compared across age and sex. There were no statistically significant differences between male and female patients for the NSA ($p = 0.40$), LCEA ($p = 0.97$), acetabular roof angle ($p = 0.19$), pelvic obliquity ($p = 0.51$), and VCA ($p = 0.12$). In addition, there were no statistically significant differences between patients over and under 60 years old for the NSA ($p = 0.49$), LCEA ($p = 0.38$), acetabular roof angle ($p = 0.43$), pelvic obliquity ($p = 0.21$), and VCA ($p = 0.053$). The MAE was then compared for patients with abnormal (over- or under-coverage) versus

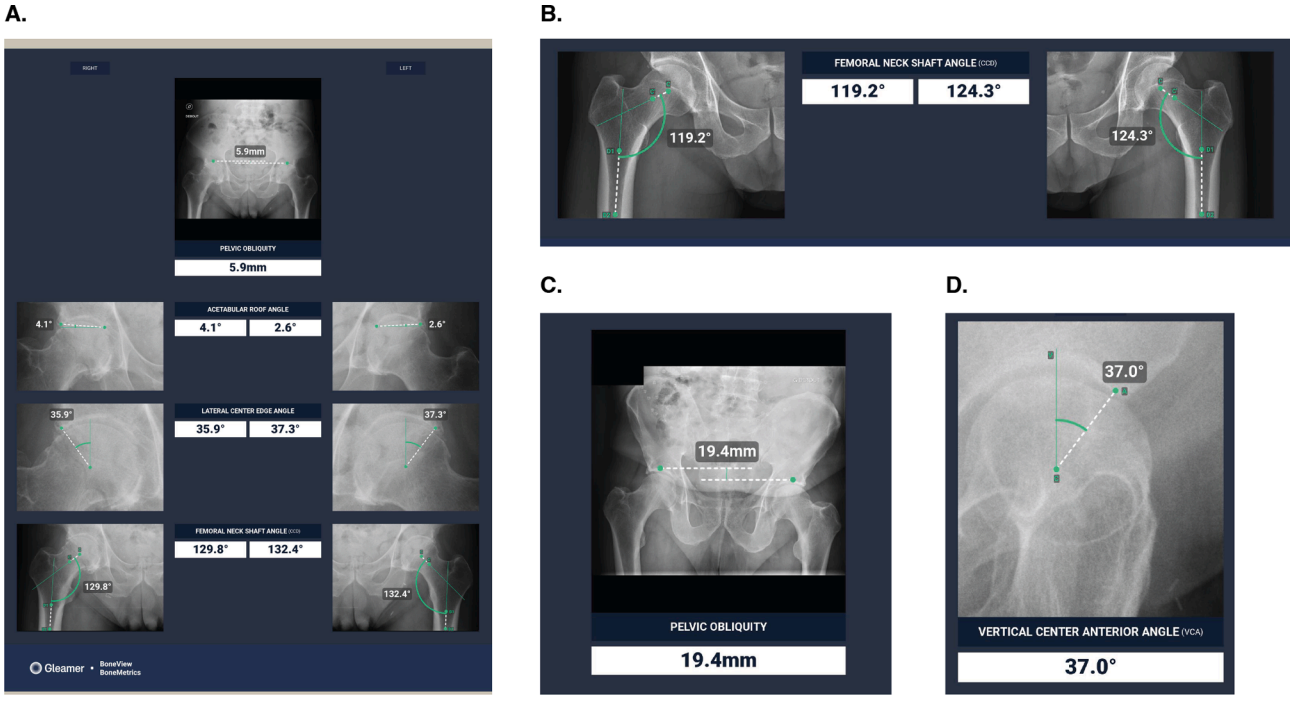

Fig. 3. Examples of measurements on anteroposterior and false profile radiographs performed by the AI algorithm. (A) Full output of the AI analysis on the anteroposterior (AP) radiographic view of a 78-year-old man. (B) A 53-year-old man with a femoral neck shaft angle (NSA) < 120°, indicative of coxa vara. (C) A 71-year-old man showing a pelvic obliquity > 10 mm, indicative of leg length discrepancy. (D) False profile radiograph of a 70-year-old woman with AI measurement of the vertical center edge angle (VCA).

normal acetabular coverage (Table 4). There were no statistically significant differences between the two groups for the NSA ($p = 0.64$), LCEA ($p = 0.039$), acetabular roof angle ($p = 0.34$) and pelvic obliquity ($p = 0.83$) on AP pelvic radiographs. Similarly, no statistically significant difference was found for the VCA measurement ($p = 0.18$) on false profile pelvic radiographs. Fig. 3 provides illustrative examples of radiographs analyzed by the AI in normal and pathological contexts.

3.4. Intra- and inter-reader analyses

Additional analyses were conducted to assess both intra- and inter-reader reliability (Table 3) using a subset of the dataset. The intra-reader reliability for one of the radiologists who established the ground truth was found to be good for the NSA with an ICC of 0.86 and for the LCEA with an ICC of 0.89. Excellent reliability was observed for the acetabular roof angle (ICC = 0.95), pelvic obliquity (ICC > 0.99), and VCA (ICC = 0.96). Inter-reader agreement, assessed among the four radiologists (two who established the ground truth and two independent radiologists), was moderate for the NSA (ICC = 0.68), LCEA (ICC = 0.62) and acetabular roof angle (ICC = 0.63). However, the agreement was found to be good for the VCA (ICC = 0.75) on false profile pelvic radiographs and excellent for pelvic obliquity (ICC > 0.99) on AP pelvic radiographs. Finally, the performance of the two independent radiologists was compared to that of the AI. The results demonstrated that both radiologists either matched the MAE of the AI or exhibited significantly higher MAE values (**Table A.2**). Specifically, the first radiologist's MAE for the acetabular roof angle was significantly higher than that of the AI ($p = 0.009$). Likewise, the second radiologist recorded significantly greater MAE values for the femoral NSA ($p < 0.001$) and the LCEA ($p < 0.001$).

4. Discussion

The present study evaluated the performance of AI-based measurements for multiple pelvic and hip parameters on AP and false profile pelvic radiographs, comparing them to a ground truth established by two senior musculoskeletal radiologists. The AI algorithm demonstrated high accuracy with MAE values inferior to 0.6 mm for pelvic obliquity and ranging from 1.79° to 4.16° for angles. Notably, the AI - ground truth reliability values for the femoral NSA (ICC = 0.81), pelvic obliquity (ICC > 0.99), LCEA (ICC = 0.92), acetabular roof angle (ICC = 0.93), and VCA (ICC = 0.78) were either statistically equivalent to or higher than the reliability between the two musculoskeletal radiologists.

The present study highlighted the robustness of AI-based pelvic and hip measurements in comparison to prior studies on automated measurements. While previous research reported a range of ICC values from moderate to excellent [19], our findings showed more consistently high agreement between the AI and the ground truth. Second, the AI - ground truth agreement was greater than the agreement between radiologists who established the ground truth. This finding aligns with previous reports on the poor reproducibility of manual measurements, likely due to variability in clinician expertise. Indeed, a review of the literature revealed that ICC values ranged from 0.58 to 0.94 for pelvic obliquity compared to > 0.99 in the present study [19,33–35], from 0.54 to 0.72 for the femoral NSA compared to 0.81 [19,35,36], from 0.73 to 0.93 for the LCEA compared to 0.92 [19,28,35–37], and from 0.45 to 0.89 for the acetabular roof angle compared to 0.93 [19,28,35–37]. This study highlights how an AI algorithm could facilitate the implementation of standardized workflows that mitigate differences in clinician expertise.

This study presents notable strengths. First, it reveals the consistent reliability of AI measurements on radiographs across a diverse patient population. This includes patients with various types of hip morphology, such as those exhibiting acetabular over and undercoverage. These findings stand in marked contrast with a prior study that indicated poorer AI - clinician agreement in cases of acetabular overcoverage [26]. In clinical practice, comprehensive evaluation of the hip often involves multiple radiographic views [38]. To our knowledge, this is the first study to test how AI would behave in the assessment of an angle on false profile views. We show good agreement between the AI-powered measurement of the VCA and the ground truth, strengthening the potential of AI to extend to all radiographic views of the pelvis and thus fill critical gaps in current clinical practice. Indeed, the VCA on false profile views can prove useful in detecting osteoarthritic changes [8].

While AI-based measurements of the pelvis and hip offer numerous advantages, it is essential to acknowledge the limitations of this study. First, the study design was retrospective, and the sample size comprised a limited number of radiographs. Second, the AI algorithm was not able to process images with prostheses, a limitation that may affect the clinicians' workflow. Moreover, this version of the algorithm did not correct for pelvic tilt and rotation, although they have been documented to influence radiographic parameters of the hip [39]. Finally, it should be noted that the study focused on a single AI solution, which could limit the broader generalizability of the findings.

In conclusion, this study demonstrates that AI offers a promising alternative to traditional manual measurements on AP and false profile pelvic radiographs. By providing highly accurate and more standardized measurements of the pelvic obliquity, femoral NSA, LCEA, acetabular roof angle, and VCA, AI stands to be a significant contributor in the diagnosis of structural disorders of the hip. Further studies that examine the integration of such algorithms into clinical workflows are essential to precisely estimate their impact on patient care and outcomes as well as on clinician efficiency and workload.

5. Financial support

This study was funded by Gleamer (Paris, France).

CRediT authorship contribution statement

Louis Lassalle: Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Data curation, Conceptualization. **Nor-eddine Regnard:** Supervision, Project administration, Methodology, Investigation, Data curation, Conceptualization. **Marion Durteste:** Writing – original draft, Visualization, Investigation, Formal analysis. **Jeanne Ventre:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **Vincent Marty:** Resources, Investigation. **Lauryane Clovis:** Resources, Investigation. **Zekun Zhang:** Software, Resources. **Nicolas Nitche:** Software, Resources. **Alexis Ducarouge:** Supervision, Project administration, Methodology, Conceptualization. **Alexia Tran:** Investigation, Data curation. **Jean-Denis Laredo:** Writing – review & editing, Methodology, Investigation, Data curation, Conceptualization. **Ali Guermazi:** Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ali Guermazi reports a relationship with Boston Imaging Core Lab LLC that includes: equity or stocks. Ali Guermazi reports a relationship with Pfizer that includes: consulting or advisory. Ali Guermazi reports a relationship with Novartis Pharmaceuticals Corporation that includes: consulting or advisory. Ali Guermazi reports a relationship with TrialSpark Inc that includes: consulting or advisory. Ali Guermazi reports a relationship with Coval that includes: consulting or advisory. Ali Guermazi reports a relationship with ICM that includes: consulting or advisory. Ali Guermazi reports a relationship with TissueGene Inc that includes: consulting or advisory. Ali Guermazi reports a relationship with Medipost that includes: consulting or advisory. Corresponding author and co-authors are employees of Gleamer (Paris, France) - LL, MD, JV, VM, LC, ZZ, NN, and JDL. Co-author is a co-founder and chief medical officer of Gleamer (Paris, France) - NER. Co-author is a co-

founder and chief technical officer of Gleamer (Paris, France) - AD. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejrad.2024.111911.

References

[1] J.C. Clohisy, et al., A systematic approach to the plain radiographic evaluation of the young adult hip, J. Bone Joint Surg. Am. 90 (2008) 47–66, https://doi.org/10.2106/JBJS.H.00756.

[2] L. Gala, J.C. Clohisy, P.E. Beaulé, Hip dysplasia in the young adult, JBJS 98 (2016) 63, https://doi.org/10.2106/JBJS.O.00109.

[3] S. James, et al., MR imaging findings of acetabular dysplasia in adults, Skeletal Radiol. 35 (2006) 378–384, https://doi.org/10.1007/s00256-006-0082-8.

[4] T.M. Ecker, M. Tannast, M. Puls, K.A. Siebenrock, S.B. Murphy, Pathomorphologic alterations predict presence or absence of hip osteoarthrosis, Clin. Orthop. Relat. Res. 465 (2007) 46, https://doi.org/10.1097/BLO.0b013e318159a998.

[5] C.C. Wyles, et al., The John Charnley Award: redefining the natural history of osteoarthritis in patients with hip dysplasia and impingement, Clin. Orthop. Relat. Res. 475 (2017) 336–350, https://doi.org/10.1007/s11999-016-4815-2.

[6] S.L. Weinstein, Natural history of congenital hip dislocation (CDH) and hip dysplasia, Clin. Orthop. Relat. Res. (1987) 62–76.

[7] J. Morvan, et al., Relationship between hip dysplasia, pain, and osteoarthritis in a cohort of patients with hip symptoms, J. Rheumatol. 40 (2013) 1583–1589, https://doi.org/10.3899/jrheum.121544.

[8] M. Lequesne, S. de Seze, False profile of the pelvis. a new radiographic incidence for the study of the hip. its use in dysplasias and different coxopathies, Rev. Rhum. Mal. Osteoartic. 28 (1961) 643–652.

[9] G. Wyberg, Studies on dysplastic acetabula and congenital subluxation of the hip joint with special reference to the complication of osteo-arthritis, J. Am. Med. Assoc. 115 (1940) 81, https://doi.org/10.1001/jama.1940.02810270083038.

[10] C.B. Lee, Y.-J. Kim, Imaging hip dysplasia in the skeletally mature, Orthop. Clin. North Am. 43 (2012) 329–342, https://doi.org/10.1016/j.ocl.2012.05.007.

[11] C.K. Boese, et al., The femoral neck-shaft angle on plain radiographs: a systematic review, Skeletal Radiol. 45 (2016) 19–28, https://doi.org/10.1007/s00256-015-2236-z.

[12] Tönnis Dietrich. *Congenital Dysplasia and Dislocation of the Hip in Children and Adults*. Berlin, Heidelberg: Springer Berlin Heidelberg (1987).

[13] F. Pereira, A. Giles, G. Wood, T.N. Board, Recognition of minor adult hip dysplasia: which anatomical indices are important? Hip Int. 24 (2014) 175–179, https://doi.org/10.5301/hipint.5000119.

[14] J. Dubousset, Pelvic obliquity: a review, Orthopedics 14 (1991) 479–481, https://doi.org/10.3928/0147-7447-19910401-13.

[15] E. Chosa, N. Tajima, Anterior acetabular head index of the hip on false-profile views. new index of anterior acetabular cover, J. Bone Joint Surg. Br. 85 (2003) 826–829, https://doi.org/10.1302/0301-620X.85B6.14146.

[16] R. Sinha, et al., Radiographic evaluation of the painful adolescent and young adult hip, JPOSNA 7 (2024), https://doi.org/10.1016/j.jposna.2024.100039.

[17] J.C. Carlisle, et al., Reliability of various observers in determining common radiographic parameters of adult hip structural anatomy, Iowa Orthop. J. 31 (2011) 52–58.

[18] D.S. Carreira, B.R. Emmons, The reliability of commonly used radiographic parameters in the evaluation of the pre-arthritic hip: a systematic review, JBJS Reviews 7 (2019) e3, https://doi.org/10.2106/JBJS.RVW.18.00048.

[19] H. Archer, et al., Artificial intelligence-generated hip radiological measurements are fast and adequate for reliable assessment of hip dysplasia, Bone Jt Open 3 (2022) 877–884, https://doi.org/10.1302/2633-1462.311.BJO-2022-0125.R1.

[20] M. Fraiwan, N. Al-Kofahi, A. Ibnian, O. Hanatleh, Detection of developmental dysplasia of the hip in X-ray images using deep transfer learning, BMC Med. Inform. Decis. Mak. 22 (2022) 216, https://doi.org/10.1186/s12911-022-01957-9.

[21] N.-E. Regnard, et al., Assessment of performances of a deep learning algorithm for the detection of limbs and pelvic fractures, dislocations, focal bone lesions, and elbow effusions on trauma X-rays, Eur. J. Radiol. 154 (2022), https://doi.org/10.1016/j.ejrad.2022.110447.

[22] A. Guermazi, et al., Improving radiographic fracture recognition performance and efficiency using artificial intelligence, Radiology 302 (2022) 627–636, https://doi.org/10.1148/radiol.210937.

[23] L. Duron, et al., Assessment of an AI aid in detection of adult appendicular skeletal fractures by emergency physicians and radiologists: a multicenter cross-sectional diagnostic study, Radiology 300 (2021) 120–129, https://doi.org/10.1148/radiol.2021203886.

[24] C.E. von Schacky, et al., Development and validation of a multitask deep learning model for severity grading of hip osteoarthritis features on radiographs, Radiology 295 (2020) 136–145, https://doi.org/10.1148/radiol.2020190925.

[25] C. Stotter, et al., Deep learning for fully automated radiographic measurements of the pelvis and hip, Diagnostics 13 (2023) 497, https://doi.org/10.3390/diagnostics13030497.

[26] G.M. Schwarz, et al., Can an artificial intelligence powered software reliably assess pelvic radiographs? Int. Orthop. (SICOT) 47 (2023) 945–953, https://doi.org/10.1007/s00264-023-05722-z.

[27] J. Jensen, et al., A deep learning algorithm for radiographic measurements of the hip in adults—a reliability and agreement study, Diagnostics 12 (2022) 2597, https://doi.org/10.3390/diagnostics12112597.

[28] W. Yang, et al., Feasibility of automatic measurements of hip joints based on pelvic radiography and a deep learning algorithm, Eur. J. Radiol. 132 (2020), https://doi.org/10.1016/j.ejrad.2020.109303.

[29] L. Lassalle, et al., Automated weight-bearing foot measurements using an artificial intelligence–based software, Skeletal Radiol. (2024), https://doi.org/10.1007/s00256-024-04726-z.

[30] G.M. Schwarz, et al., Artificial intelligence enables reliable and standardized measurements of implant alignment in long leg radiographs with total knee arthroplasties, Knee Surg. Sports Traumatol. Arthrosc. 30 (2022) 2538–2547, https://doi.org/10.1007/s00167-022-07037-9.

[31] R.A. Parker, et al., Application of mixed effects limits of agreement in the presence of multiple sources of variability: exemplar from the comparison of several devices to measure respiratory rate in COPD patients, PLoS One 11 (2016) e0168321, https://doi.org/10.1371/journal.pone/0168321.

[32] T.K. Koo, M.Y. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research, J. Chiropr. Med. 15 (2016) 155–163, https://doi.org/10.1016/j.jcm.2016.02.012.

[33] M. Tannast, et al., Tilt and rotation correction of acetabular version on pelvic radiographs, Clin. Orthop. Relat. Res. 438 (2005) 182, https://doi.org/10.1097/01.blo.0000167669.26068.c5.

[34] F. Kalberer, R.J. Sierra, S.S. Madan, R. Ganz, M. Leunig, Ischial spine projection into the pelvis, Clin. Orthop. Relat. Res. 466 (2008) 677–683, https://doi.org/10.1007/s11999-007-0058-6.

[35] N.H. Mast, F. Impellizzeri, S. Keller, M. Leunig, Reliability and agreement of measures used in radiographic evaluation of the adult hip, Clin. Orthop. Relat. Res. 469 (2011) 188, https://doi.org/10.1007/s11999-010-1447-9.

[36] M. Nelitz, K.P. Guenther, S. Gunkel, W. Puhl, Reliability of radiological measurements in the assessment of hip dysplasia in adults, BJR 72 (1999) 331–334, https://doi.org/10.1259/bjr.72.856.10474491.

[37] M. Tannast, et al., Radiographic analysis of femoroacetabular impingement with Hip2norm—reliable and validated, J. Orthop. Res. 26 (2008) 1199–1205, https://doi.org/10.1002/jor.20653.

[38] Steppacher, S. D., Albers, C. E., Stetzelberger, V. M., Tannast, M. & Siebenrock, K. A. Plain Radiographic Evaluation of the Hip. in *Hip Arthroscopy and Hip Joint Preservation Surgery* (eds. Nho, S. J., Bedi, A., Salata, M. J., Mather III, R. C. & Kelly, B. T.) 27–46 (Springer International Publishing, Cham, 2022). 10.1007/978-3-030-43240-9_3.

[39] M. Tannast, S. Fritsch, G. Zheng, K.A. Siebenrock, S.D. Steppacher, Which radiographic hip parameters do not have to be corrected for pelvic rotation and tilt? Clin. Orthop. Relat. Res. 473 (2015) 1255, https://doi.org/10.1007/s11999-014-3936-8.