



# Deep learning generated lower extremity radiographic measurements are adequate for quick assessment of knee angular alignment and leg length determination

Holden Archer<sup>1</sup> · Seth Reine<sup>1</sup> · Shuda Xia<sup>1</sup> · Louis Camilo Vazquez<sup>1</sup> · Oganesh Ashikyan<sup>1</sup> · Parham Pezeshk<sup>1</sup> · Ajay Kohli<sup>1</sup> · Yin Xi<sup>1</sup> · Joel E. Wells<sup>2</sup> · Allan Hummer<sup>3</sup> · Matthew Difranco<sup>3</sup> · Avneesh Chhabra<sup>1,4,5,6</sup>

Received: 27 June 2023 / Revised: 21 October 2023 / Accepted: 28 October 2023 / Published online: 15 November 2023  
© The Author(s), under exclusive licence to International Skeletal Society (ISS) 2023

## Abstract

**Purpose** Angular and longitudinal deformities of leg alignment create excessive stresses across joints, leading to pain and impaired function. Multiple measurements are used to assess these deformities on anteroposterior (AP) full-length radiographs. An artificial intelligence (AI) software automatically locates anatomical landmarks on AP full-length radiographs and performs 13 measurements to assess knee angular alignment and leg length. The primary aim of this study was to evaluate the agreements in LLD and knee alignment measurements between an AI software and two board-certified radiologists in patients without metal implants. The secondary aim was to assess time savings achieved by AI.

**Methods** The measurements assessed in the study were hip-knee-angle (HKA), anatomical-tibiofemoral angle (aTFA), anatomical-mechanical-axis angle (AMA), joint-line-convergence angle (JLCA), mechanical-lateral-proximal-femur-angle (mLPFA), mechanical-lateral-distal-femur-angle (mLDFA), mechanical-medial-proximal-tibia-angle (mMPTA), mechanical-lateral-distal-tibia-angle (mLDTA), femur length, tibia length, full leg length, leg length discrepancy (LLD), and mechanical axis deviation (MAD). These measurements were performed by two radiologists and the AI software on 164 legs. Intraclass-correlation-coefficients (ICC) and Bland–Altman analyses were used to assess the AI's performance.

**Results** The AI software set incorrect landmarks for 11/164 legs. Excluding these cases, ICCs between the software and radiologists were excellent for 12/13 variables (11/13 with outliers included), and the AI software met performance targets for 11/13 variables (9/13 with outliers included). The mean reading time for the AI algorithm and two readers, respectively, was 38.3, 435.0, and 625.0 s.

**Conclusion** This study demonstrated that, with few exceptions, this AI-based software reliably generated measurements for most variables in the study and provided substantial time savings.

**Keywords** Leg length discrepancy · Knee deformities · Radiographs · Deep learning · Artificial intelligence

**Summary statement** Our study showed that, with the exception of outlier cases, AI-based software demonstrated reliable radiographic assessment of knee angular alignment and leg length determination for most measurements along with significant time savings.

## Key points

1. The artificial intelligence algorithm demonstrated excellent agreement (intraclass correlation coefficients > 0.75) for 12/13 measurements used in the study when outlier cases were excluded.
2. Artificial intelligence met the performance targets based on Bland-Altman analysis for 11/13 variables when outlier cases were excluded.
3. The artificial intelligence system generated measurements more than 90% faster than the expert readers.

Extended author information available on the last page of the article

## Introduction

Leg length discrepancy (LLD) is a condition in which the paired lower extremities are of different lengths. The prevalence of LLD is estimated to be about 90% in the general population [1]. However, most LLD is asymptomatic until it reaches a large enough or a clinically significant value (the specific value constituting “large enough” is debated) [2, 3]. Large LLD is problematic because it has been linked to altered gait patterns, low back pain, hip and knee pain, scoliosis, and osteoarthritis [3–5]. Although there is some disagreement about when LLD becomes clinically significant, LLD above 2.0 cm is a criterion used for surgical correction [6, 7]. Symptomatic LLD below 2.0 cm may be treated conservatively [6, 7].

Apart from longitudinally measured LLD, leg and knee alignment are also important factors in pre-operative assessment. The mechanical axis of the lower extremity (representing the weight-bearing axis) is the line from the center of the femoral head to the center of the ankle joint [8]. In a varus knee (bowed leg), the mechanical axis is medial to the knee center, and in a valgus knee, the mechanical axis is lateral of the knee center. In one study from China with all participants above the age of 50 years, the prevalence of knees with normal, valgus, or varus alignments was estimated to be 29.9%, 56.5%, and 13.7%, respectively [9]. Varus and valgus deformities alter how the force is distributed across the knees and can promote the progression of osteoarthritis (OA) [10].

LLD and knee alignment deformities can be measured radiographically on the anteroposterior (AP) view of the

lower extremity using a variety of parameters, notably hip-knee-ankle angle (HKA), anatomical-mechanical axis angle (AMA), mechanical lateral distal femoral angle (mLDFA), and joint line convergence angle (JLCA) among others (Table 1) [8]. The landmarks used for manual measurement are described in Table 1 and illustrated in Figs. 3 and 4.

Measuring these different parameters for a comprehensive assessment of lower extremity anatomy and deformity is time-consuming, effort-intensive, and inconsistent among readers. If a machine can be trained to perform these measurements, it may save considerable effort on the part of the radiologist and treatment-planning physician. Using artificial intelligence (AI) may also potentially allow for rapid, consistent, and reproducible measurements of the lower extremity [11, 12]. The primary aim of this study was to evaluate the agreements in LLD and knee alignment measurements

**Table 1** Parameters and landmarks used for manual and AI-based measurements on full leg standing radiographs

| Measurement                                     | Landmarks  |
|---|--|
| Hip-knee-ankle angle (HKA)                      | The HKA is defined as the angle between the mechanical axis of the femur (center of the femoral head to femoral-knee center) and the mechanical axis of the tibia (tibial-knee center to ankle joint center at the level of the tibia) [19]. Positive values indicate varus and negative values indicate valgus  |
| Anatomical tibiofemoral angle (aTFA)            | The aTFA is the angle between the anatomical tibia axis (tibial bone shaft axis) and anatomical femur axis (femoral bone shaft axis) [20]. Positive values indicate varus and negative values indicate valgus  |
| Anatomical-mechanical axis angle (AMA)          | The AMA is the angle between the mechanical femur axis and anatomical femur axis [19]  |
| Joint line convergence angle (JLCA)             | The JLCA is the angle between the distal femoral-knee joint orientation line (defined as a tangent to the two most convex points of the femoral condyles) and proximal tibial knee joint orientation line (defined as a tangent to the most concave aspect of the tibial plateau subchondral line) [19]. It is classified as positive for medial joint space narrowing and negative for lateral joint space narrowing  |
| Mechanical lateral proximal femur angle (mLPFA) | The mLPFA is defined as the proximal angle between the hip joint orientation line (a line connecting the greater trochanter to the center of the femoral head) and mechanical femoral axis [21]  |
| Mechanical lateral distal femur angle (mLDFA)   | The mLDFA is defined as the lateral distal angle between the mechanical femoral axis and the distal femoral-knee joint orientation line [19, 21]   |
| Mechanical medial proximal tibia angle (mMPTA)  | The mMPTA is defined as the medial proximal angle between the mechanical tibial axis and the proximal tibial-knee joint orientation line [19]  |
| Mechanical lateral distal tibia angle (mLDTA)   | The mLDTA is defined as the lateral distal angle between the mechanical tibial axis and ankle joint orientation line (a line connecting both ends of the tibial plafond) [22]  |
| Femur length                                    | The femur length is defined as the distance from the most cranial point of the femoral head to the most distal point of the medial femoral condyle   |
| Tibia length                                    | The tibia length is defined as the distance between the most distal point of the medial femoral condyle to the ankle joint center (determined by the midpoint of the medial and lateral points of the talus at the level of the distal tibia)  |
| Full leg length                                 | The full leg length is defined as the distance from the most cranial point of the femoral head to the ankle joint center   |
| Leg length discrepancy (LLD)                    | The LLD is defined as the full leg length difference between the two legs. The LLD is only calculated for bilateral images   |
| Mechanical axis deviation (MAD)                 | The MAD is defined as the distance between the mechanical axis of the leg (the center of the femoral head to the ankle joint center) and the knee joint center. Positive values indicate that the knee joint center is lateral of the mechanical axis of the leg [21]. Negative values indicate that the knee joint center is medial of the mechanical axis of the leg. In patients with total knee arthroplasty (TKA), the femoral knee center (rather than the knee joint center) is used for this calculation |

between a Conformance Européenne (CE) certified AI software and two board-certified experienced radiologists in patients without metal implants. The secondary aim was to assess time savings achieved by AI.

## Materials and methods

This study received institutional review board (IRB) approval for retrospective cross-sectional analysis of a consecutive sample from our institutional database. The HIPAA regulations were followed, and an IRB waiver was in place for evaluation of patient images.

### Patients

Our tertiary care center's institutional database was queried for all full leg standing radiographs from January 2014 to February 2022 at two hospitals regardless of the type of machine used. The inclusion criteria were age 18–100 years, all genders, and full leg standing radiographs. Exclusion criteria were (1) failure to meet the quality criteria of the AI algorithm and (2) presence of any metal implant, which was applied on a per leg basis; if only one leg of a bilateral image was excluded, then the image remained in the dataset. Patient demographic data including race, sex, and age were extracted from the database.

### Imaging parameters

All images were taken using anteroposterior (AP) full leg standing radiographs with feet facing forward. The tube-to-film distance was 120 cm using 80–90 kilovoltage peak (kVp) and 20–30 milli-ampere-second (mAs) depending on the patient's size.

### AI algorithm

A vendor-provided deep-learning based software (IB Lab LAMA, IB Lab GmbH, Vienna, Austria) automatically locates anatomical landmarks on AP full leg standing radiographs and performs 13 measurements: hip-knee-ankle angle (HKA), anatomical tibiofemoral angle (aTFA), anatomical-mechanical axis angle (AMA), joint line convergence angle (JLCA), mechanical lateral proximal femur angle (mLPFA), mechanical lateral distal femur angle (mLDFA), mechanical medial proximal tibia angle (mMPTA), mechanical lateral distal tibia angle (mLDTA), femur length, tibia length, full leg length, leg length discrepancy (LLD), and mechanical axis deviation (MAD). An example AI report is shown in Fig. 1.

The AI software accepts images in DICOM format as input and returns a DICOM compatible AI report. A software failure is considered when the output of the software is an error report or when no report at all is returned.

All images used in the study were securely transferred to our institutional research PACS server. From there, the images were pushed to a local installation of the vendor AI software to obtain the automated measurements.

### Image quality assurance

The AI software has a set of imaging requirements to ensure proper functioning, as listed in Table 2. An initial quality assurance (QA) process was performed through a Redcap form by three medical students to exclude images not meeting those requirements. The radiologist readers carried out QA with the same requirements when performing the study reads. The QA criteria were applied on a per-image basis and some images had multiple violations. The number of images that were in violation of each criterion are included in Table 2.

A total of 1172 legs (586 images) entered the quality assurance process, and 526 legs (from 263 images) passed the process. In total, 370 of these 526 legs did not contain implants. From this group, 164 legs (the adequate sample size determined from Bland–Altman analysis described below) were randomly sampled to obtain the final cohort (Fig. 2).

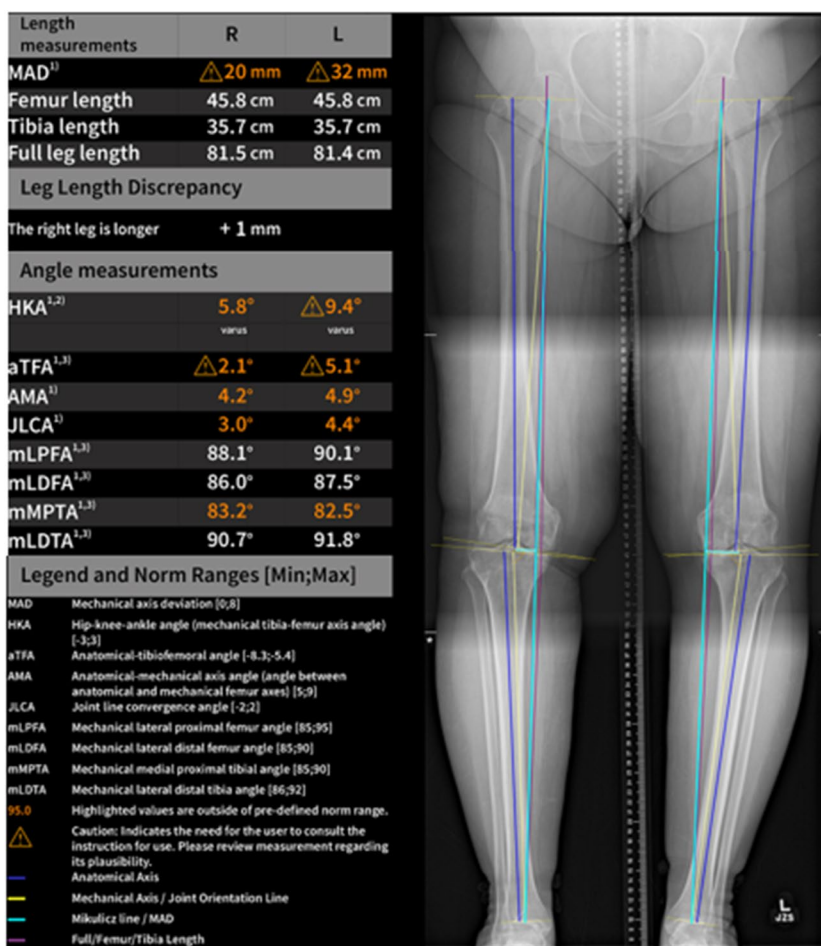
### Clinically meaningful maximum allowed differences

Clinically meaningful maximum allowed differences were defined as the threshold values for which any difference, as large or larger, was considered meaningful to patients, clinicians, or both [13]. Clinically meaningful maximum allowed differences for each variable were assigned prospectively and confirmed by a senior musculoskeletal fellowship trained radiologist and fellowship-trained hip preservation orthopedic surgeon (Table 3).

### Reference standard

Manual measurements were performed by two expert radiologist readers, both more than 5 years after musculoskeletal radiology fellowship. Both readers were trained in all the measurements using an external set of 10 images which were not included in the final study. They were also given a presentation guide on a PowerPoint (Microsoft, Redwood, WA) describing the measurements with pictures to use as examples during the image evaluations (Figs. 3 and 4).

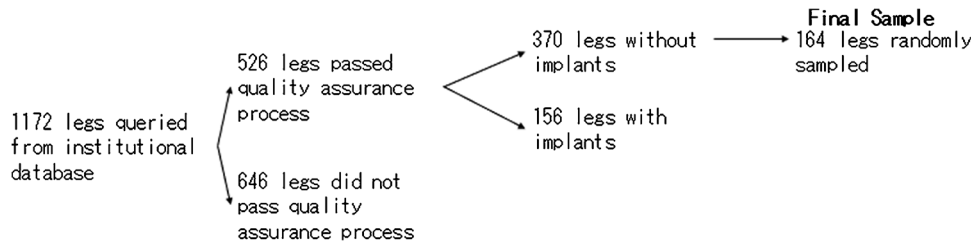
**Fig. 1** Example artificial intelligence (AI) output. This figure shows an example of the output produced by the AI. The colorful lines represent the landmarks and axis used for each measurement. When a value is outside of the normal range, based on selected literature, it is highlighted on the output table in orange



**Table 2** Criteria to fit AI software image quality requirements and how many exclusions there were for each criterion

| Criteria   | Exclusions |
|--|------------|
| AI software image quality exclusions   |            |
| 1) Total ankle arthroplasty present  | 2          |
| 2) Bone contours of the tibia, femur, talus and acetabular pelvic region are not fully visible and/or are overlapped by, e.g., calibration devices, radiographic protections, implants or image artifacts  | 137        |
| 3) Critical areas including landmarks used for angle and length measurements are not fully visible and are overlapped by, e.g., calibration devices, radiographic protections, implants or image artifacts | 138        |
| 4) The image stitching is not continuous, and anatomical features are not clearly and completely visible   | 19         |
| 5) The contrast across stitched images is not homogenous   | 55         |
| 6) The image cropping has not been done correctly (additional regions are included)  | 47         |
| 7) The calibration ball, if present, is positioned close to the region of interest or not fully visible  | 0          |
| 8) The radiograph is not of adequate quality (e.g., too noisy, poor contrast on all or part of the image) as subjectively assessed by the radiologists   | 33         |

**Fig. 2** Final study cohorts with inclusions and exclusions



**Table 3** Maximum allowed differences prospectively assigned for each variable

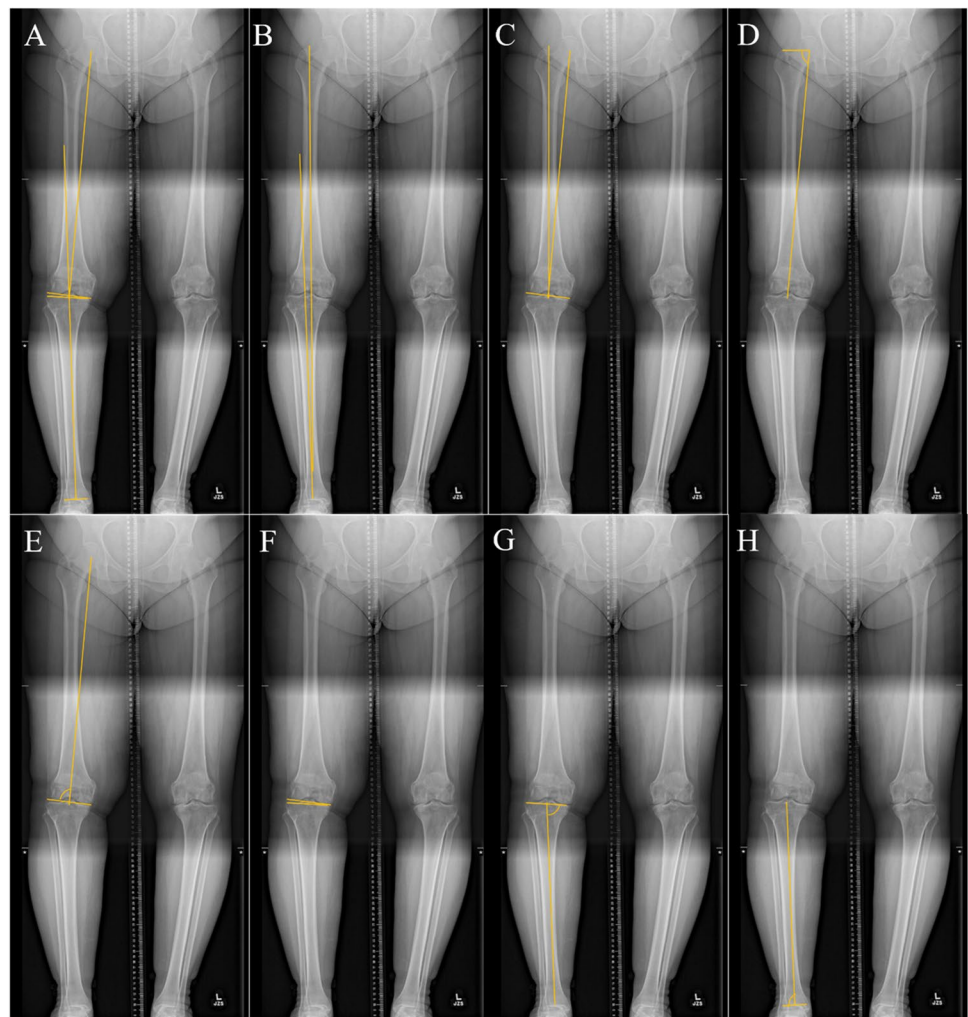
| Angles (°)   | HKA | aTFA       | AMA          | JLCA         | mLPFA | mLDFA | mMPTA | mLDTA |
|--------------|-----|------------|--------------|--------------|-------|-------|-------|-------|
|              | 3   | 4          | 3            | 4            | 8     | 4     | 4     | 5     |
| Lengths (mm) | MAD | Leg length | Femur length | Tibia length | LLD   |       |       |       |
|              | 6   | 6          | 8            | 8            | 8     |       |       |       |

Following the training process, the measurements were performed by each reader on IntelliSpace Picture Archiving and Communication System (IPACS, Philips, Best, the Netherlands) with the built-in measurement tools. The measurements were performed independently, and the readers were blinded to each other's measurements and the AI measurements. Reader discrepancies for measurements that had different signs (positive vs. negative) for the parameters that required sign in addition to magnitude (HKA, aTFA, JLCA, MAD) or obvious recording errors (e.g., writing 950° instead of 9.50° or using cm instead of mm) were repeated.

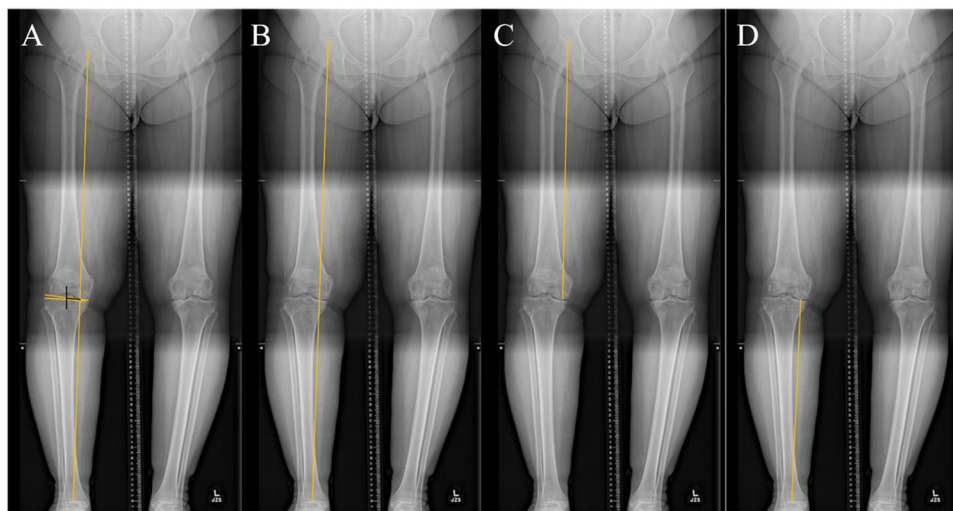
### Definition of reading time

The reading time was automatically captured in Redcap form, where the measurements were stored. There were separate start and stop buttons built into the Redcap form so that the time for each leg could be recorded individually. The reading time for each leg was calculated as the difference between the starting time and ending time. The AI “reading time” was recorded with a stopwatch; it measured the time from which the image was sent via cloud to the vendor for analysis to the time the AI report returned.

**Fig. 3** Landmarks used for manual angle measurements. **A** Hip knee angle (HKA). **B** Anatomical tibiofemoral angle (aTFA). **C** Anatomical mechanical axis angle (AMA). **D** Mechanical lateral proximal femur angle (mLPFA). **E** Mechanical lateral distal femur angle (mLDFA). **F** Joint line convergence angle (JLCA). **G** Mechanical medial proximal tibia angle (mMPTA). **H** Mechanical lateral distal tibia angle (mLDTA)



**Fig. 4** Landmarks used for manual length measurements. **A** Mechanical axis deviation (MAD). **B** Full leg length. **C** Femur length. **D** Tibia length



## Statistical methods

Summary statistics including counts and percentages were tabulated for the scanners used in the study and for the patient characteristics of race, site, sex, and age.

All analyses were done on a per-leg basis except for LLD, where analyses were conducted on a per-image basis on images which contained two non-implant legs.

The failure rate of the AI algorithm was defined as the percentage of leg readings returning an error report or no report due to algorithm-based errors. Each completed report from the AI algorithm was visually inspected, and error rates due to missed image QA measures, incorrect meta data requirements, and misplaced measurement landmarks were recorded. Images containing these errors were removed from the study if the errors were deemed to be easily identifiable (i.e., misplaced landmark) upon manual inspection of the report.

Inter-reader agreement between the two expert readers was assessed using intraclass correlation coefficients (ICC, single-rating, absolute-agreement, two-way random-effects model). Bland–Altman analyses as well as ICC (single-rating, absolute-agreement, two-way mixed-effects model) were conducted to assess the agreement between the reference standard and the AI algorithm measurements for all 13 variables [14]. Agreement between the AI algorithm and the reference standard was deemed acceptable if the 95% confidence interval upper limit for the upper limit of agreement (LOA) and the 95% confidence interval lower limit for the lower LOA were both smaller in magnitude than the predefined clinically meaningful maximum allowed difference for the given leg variable as defined in Table 3.

To calculate the percentage time reduction offered by the AI algorithm for a given patient, a linear mixed model was fit with log-transformed time as the dependent variable and a three-level categorical variable indicating the reader as the

independent variable. Random intercepts were included for each leg. Linear contrasts were estimated and exponentiated to calculate the percentage time reduction produced by the AI algorithm for a given patient relative to each of the two readers.

Equivalence index, root mean square error [15], and mean absolute deviation (along with 95% confidence intervals) were calculated for all variables. Equivalence index is a metric to assess interchangeability using the concept from Obuchowski et al. [16]. It is meant to demonstrate that when tests are compared to each other (such as the AI product versus the expert readers), both assessments produce the same results in a given patient. An equivalence index above 0 indicates that deviation between the AI and radiologists is larger than the deviation between the two radiologists. Therefore, if the equivalence index is equal to or below 0, there is evidence that the AI measurements are interchangeable with radiologist measurements in this study.

Ninety-five percent confidence intervals were reported for all ICC and time reduction estimates. Benchmarks for the ICC estimates were as follows: 0.00 to 0.40 poor; 0.41 to 0.59 fair; 0.60 to 0.74 good; and 0.75 to 1.00 excellent [17]. Agreement analyses were performed in R (R Core Team, 2022) using the irr package (ICC) and Python 3.8 (Python Software Foundation, 2019) using in-house code (Bland–Altman). The timing mixed model analysis was performed in the SAS v.9.4 Mixed Procedure (SAS Institute, USA).

## Sample size determination

Sample sizes were determined based on the Bland–Altman analysis. For each variable, the method of Lu et al. was used which required the clinically meaningful maximum allowed difference values as well as estimated values for the mean and standard deviation of the differences between the AI algorithm

**Table 4** Sample sizes to achieve 80% power in the Bland–Altman analysis for each variable

| Angles (°)   | HKA | aTFA       | AMA          | JLCA         | mLPFA | mLDFA | mMPTA | mLDTA |
|--------------|-----|------------|--------------|--------------|-------|-------|-------|-------|
|              | 16  | 58         | 10           | 43           | 168   | 99    | 23    | 84    |
| Lengths (mm) | MAD | Leg length | Femur length | Tibia length | LLD   |       |       |       |
|              | 131 | 10         | 8            | 22           | 60    |       |       |       |

and the reference standard [18]. The significance level and statistical power for the study were set to 0.05 and 0.80, respectively. To determine estimates for the mean and standard deviation of the differences, a pilot study using 50 images external to the final study was performed by three trained medical student readers. The resulting sample sizes required to achieve 80% power for each variable are provided in Table 4.

## Results

### Patients

Summary statistics of the study population characteristics in Fig. 5 show that patients were majority female (58%), between the ages of 45 and 64 (45%), and of white race (53%). There were more patients from site 2 (58%) than site 1 (42%), and the majority of scanners (71%) used across images were from Siemens, Erlangen, Germany.

### AI algorithm computational performance

Of the available 164 legs from 103 images, 132 legs from 85 images were successfully processed by the AI algorithm and served as the final study population for the agreement and time reduction analyses. The failure rate due to internal errors of the AI algorithm was calculated at 1.8% (3/164). An additional 11.0% (18/164) of legs led to no output by the AI algorithm due to quality assurance criteria not being met upon secondary inspection ( $n=2$ ), missing pixel spacing DICOM metadata tags ( $n=11$ ), and incorrect pixel scaling of length measurements ( $n=5$ ). Lastly, 6.7% (11/164) legs had incorrect landmarks set by the AI algorithm. Visual inspection of the reports for these images with missed landmarks deemed their output as obvious errors that would be caught easily (Supplemental.Fig. 1); hence, these images were removed from the final agreement and time reduction analyses. However, results for ICC calculation and Bland–Altman analysis including the images removed upon visual inspection are available in the Supplementary Materials (Supplemental.Table.1, Supplemental.Table.2).

Because the analyses assume independence, which may be unlikely given that some patients contributed two legs to the analysis, calculations were performed again with a single leg randomly chosen. There were minor differences in the results (Supplemental. Table.3, Supplemental.Table.4, Supplemental. Table.5, Supplemental.Table.6).

### Inter-reader agreement

ICC estimates between the two radiologist readers demonstrated fair to excellent agreement across all variables (Table 6), with excellent agreement being demonstrated by 12 of the 13 variables. The lowest agreement between readers occurred for AMA (ICC=0.55). Similarly, in the analysis that included images previously removed for erroneous landmark placement, 12/13 variables demonstrated excellent (Supplemental.Table.2). AMA agreement was fair (ICC=0.57).

### AI algorithm agreement with reference standard

The results of the Bland–Altman agreement analysis are shown in Table 5. The AI algorithm met the performance targets for agreement for the following measurements: Femur length, tibia length, leg length, LLD, HKA, aTFA, AMA, JLCA, mLPFA, mLDFA, and mMPTA. The AI algorithm did not meet the performance targets for agreement for MAD and mLDTA. In the analysis that included images with erroneously placed landmarks, performance targets were not met for 4/13 variables: mLDTA, mLPFA, mMPTA, and MAD (Supplemental.Table.1).

Given that 132 legs were available for the analysis, 80% prospective power was maintained for all variables except mLPFA and LLD. Since images consisting of two native legs were required for the LLD variable, only 47 measurements were available for this measure.

Equivalence index, root mean square error (RMSE), and mean absolute deviation (MAD) are shown in Table 7. A corresponding table including outliers is shown in Supplemental.Table.7.

ICC estimates between the AI algorithm and the reference standard demonstrated excellent agreement across 12/13 variables (Table 6). The lowest agreement occurred for AMA (ICC=0.64) and mLDTA (ICC=0.80). In the analysis that included images previously removed for erroneous landmark placement, 11/13 variables demonstrated excellent ICC. AMA (ICC=0.64) and mLPFA (ICC=0.73) had good agreement (Supplemental.Table.2).

### Time reduction analysis

A subset of 48 images of the total 85 images contained two native legs and were thus used for the time reduction analysis. The mean (standard error in parenthesis) reading

| <b>Race</b>               | <b>n (%)</b> |
|---------------------------|--------------|
| White                     | 45 (53%)     |
| Black or African American | 10 (12%)     |
| Asian                     | 2 (2%)       |
| Other/Unknown             | 28 (33%)     |
| Total                     | 85 (100%)    |
| <b>Site</b>               | <b>n (%)</b> |
| Site 1                    | 36 (42%)     |
| Site 2                    | 49 (58%)     |
| Total                     | 85 (100%)    |
| <b>Vendors</b>            | <b>n (%)</b> |
| CANON Inc.                | 11 (13%)     |
| KODAK                     | 11 (13%)     |
| KONICA MINOLTA            | 3 (4%)       |
| SIEMENS                   | 60 (71%)     |
| Total                     | 85 (100%)    |
| <b>Patient Sex</b>        | <b>n (%)</b> |
| Female                    | 49 (58%)     |
| Male                      | 36 (42%)     |
| Total                     | 85 (100%)    |
| <b>Age Group</b>          | <b>n (%)</b> |
| 18 to 24                  | 2 (2%)       |
| 25 to 44                  | 15 (18%)     |
| 45 to 64                  | 38 (45%)     |
| 65 to 79                  | 16 (19%)     |
| 80 and over               | 5 (6%)       |
| Unknown                   | 9 (11%)      |
| Total                     | 85 (100%)    |

**Fig. 5** Patient demographics

time per image for the AI algorithm and the two radiologist readers were 38.3 (0.7), 435.0 (25.0), and 625.0 (35.8) seconds, respectively. For a given image, the AI algorithm performed reads 91.0% (90.0 to 91.8%) and 93.9% (93.2 to 94.5%) faster than reader 1 and reader 2, respectively.

## Discussion

To the authors' knowledge, there are no commercially available software products that comprehensively perform all the measurements evaluated in this study. With outlier cases excluded, the AI algorithm demonstrated excellent agreement ( $ICC > 0.75$ ) for 12/13 of these measurements ( $AMA = 0.64$ ). The lowest ICC values were for  $AMA$  ( $ICC = 0.64$ ) and  $mLDTA$  ( $ICC = 0.80$ ), but the expert readers also showed more variability with these measurements ( $ICC = 0.55$  and  $0.84$ , respectively).

In general, the AI was successful at performing these measurements with a few exceptions. With outlier cases excluded, the Bland–Altman analysis showed that the AI met the performance target on 11/13 variables (with the exception of  $MAD$  and  $mLDTA$ ). Though not being accepted in the Bland–Altman analysis, the  $MAD$  showed excellent agreement in the ICC analysis ( $ICC = 0.99$ ). Notably, in the analysis that excluded outlier cases, 5/13 variables ( $AMA$ ,  $HKA$ ,  $mLDFA$ , leg length, tibia length) had equivalence indices with upper limits below 0, which suggests interchangeability between AI measurements and reader measurements in those variables (Table 7). With all images included, only  $AMA$  and  $HKA$  were deemed “interchangeable” based on the equivalence index (Supplemental Table 7). Therefore, using equivalence index, it appears most measurements are not interchangeable despite meeting Bland–Altman performance targets. Femur length, leg length, and tibia length had the lowest RMSE and  $MAD$  values, indicating less deviation from the radiologists' measurements (Table 7). This is consistent with ICC and Bland–Altman analysis, which showed good performance in these length measurements compared to the angular measurements. The highest RMSE and  $MAD$  values were seen with  $mLPFA$ ,  $MAD$ , and  $mLDTA$ , indicating more deviation from radiologists' measurements.  $MAD$  and  $mLDTA$  were not accepted in Bland–Altman analysis (Table 5), so poor performance in these variables as captured by RMSE and  $MAD$  is expected.  $mLPFA$ , although accepted by Bland–Altman, had the third lowest ICC in the study and thus expectedly had large RMSE and  $MAD$  values as well.

Although Bland–Altman analysis was the primary tool to assess AI software performance in this study, equivalence index, RMSE, and  $MAD$  provide additional insight confirming poor performance of  $mLPFA$ ,  $MAD$ , and  $mLDTA$  and



**Table 5** Bland–Altman analysis results between AI algorithm and the reference standard

| Variable          | <i>n</i> | Mean differences (lower 95% CI of lower LOA, upper 95% CI of upper LOA) | SD of differences | Acceptance |
|-------------------|----------|---|-------------------|------------|
| AMA (°)           | 132      | 0.08 (−1.63, 1.79)  | 0.76              | Yes        |
| aTFA (°)          | 132      | −0.59 (−3.3, 2.11)  | 1.2               | Yes        |
| HKA (°)           | 132      | −0.2 (−1.51, 1.41)  | 0.71              | Yes        |
| JLCA (°)          | 132      | 0.22 (−3.01, 3.46)  | 1.44              | Yes        |
| mLDFA (°)         | 132      | −0.3 (−2.21, 1.61)  | 0.85              | Yes        |
| mLDTA (°)         | 132      | −0.65 (−5.61, 4.3)  | 2.2               | No         |
| mLPFA (°)         | 132      | 2.43 (−3.05, 7.9)   | 2.43              | Yes        |
| mMPTA (°)         | 132      | −0.4 (−3.39, 3.32)  | 1.49              | Yes        |
| Femur Length (mm) | 132      | 0.08 (−0.2, 0.36)   | 0.12              | Yes        |
| Leg Length (mm)   | 132      | 0.05 (−0.2, 0.3)  | 0.11              | Yes        |
| LLD (mm)          | 47       | 0.03 (−3.33, 3.39)  | 1.36              | Yes        |
| MAD (mm)          | 132      | −1.53 (−7.03, 3.97)   | 2.44              | No         |
| Tibia Length (mm) | 132      | −0.02 (−0.26, 0.22)   | 0.11              | Yes        |

**Table 6** Intraclass correlation coefficient (ICC) estimates for AI algorithm vs. reference standard and for reader 1 vs. reader 2

| Variable     | Reader 1 vs. reader 2 ICC (95% CI) | AI algorithm vs. reference standard ICC (95% CI) |
|--------------|------------------------------------|--|
| AMA          | 0.55 (0.42, 0.66)                  | 0.64 (0.56, 0.72)                                |
| aTFA         | 0.98 (0.98, 0.99)                  | 0.98 (0.98, 0.99)                                |
| HKA          | 0.99 (0.99, 0.99)                  | 0.99 (0.99, 0.99)                                |
| JLCA         | 0.89 (0.82, 0.93)                  | 0.89 (0.85, 0.92)                                |
| mLDFA        | 0.91 (0.88, 0.94)                  | 0.93 (0.91, 0.95)                                |
| mLDTA        | 0.84 (0.78, 0.89)                  | 0.8 (0.74, 0.85)                                 |
| mLPFA        | 0.94 (0.9, 0.96)                   | 0.88 (0.72, 0.93)                                |
| mMPTA        | 0.89 (0.83, 0.92)                  | 0.89 (0.85, 0.92)                                |
| Femur Length | > 0.99 (> 0.99, 1)                 | > 0.99 (> 0.99, 1)                               |
| Leg Length   | > 0.99 (> 0.99, 1)                 | > 0.99 (> 0.99, 1)                               |
| LLD          | 0.99 (0.98, 0.99)                  | 0.99 (0.98, 0.99)                                |
| MAD          | 0.99 (0.99, 1)                     | 0.99 (0.99, 1)                                   |
| Tibia Length | > 0.99 (> 0.99, 1)                 | > 0.99 (> 0.99, 1)                               |

confirming relatively good performance of the length measurements (particularly leg length).

There were 13 variables used in this study, and for physicians who do not perform these measurements regularly, it is burdensome to remember the intricacies and exact landmarks of each measurement. If AI software could accurately measure these variables, radiologists, physical therapists, and orthopaedic surgeons would no longer need to focus on remembering “how to” perform these measurements and could instead devote time prudently to more difficult tasks not suited for AI.

The results of this study show that, regardless of which cases are included or excluded, AI can measure femur length, leg length, and tibia length very quickly and reliably. This tool could provide preliminary leg length measurements to clinicians for quick evaluation of the lower extremity. Difficult cases, determined by quickly looking at the AI output, could be measured manually.

**Table 7** Equivalence index, root mean square error, and mean absolute deviation between AI and readers

| Variable     | Equivalence Index (95% CI) | RMSE (95% CI)     | MAD (95% CI)      |
|--------------|----------------------------|-------------------|-------------------|
| AMA          | −0.57 (−0.83, −0.38)       | 0.76 (0.66, 0.87) | 0.58 (0.5, 0.67)  |
| aTFA         | 0.4 (0.07, 1.27)           | 1.34 (1.18, 1.49) | 1.06 (0.93, 1.2)  |
| HKA          | −0.43 (−0.79, −0.25)       | 0.74 (0.65, 0.84) | 0.58 (0.51, 0.66) |
| JLCA         | 0.17 (−0.66, 1.22)         | 1.45 (1.2, 1.78)  | 1.11 (0.97, 1.29) |
| mLDFA        | −0.82 (−1.01, −0.66)       | 0.9 (0.79, 1.01)  | 0.7 (0.61, 0.8)   |
| mLDTA        | 1.48 (1.07, 2.02)          | 2.29 (1.96, 2.61) | 1.78 (1.54, 2.03) |
| mLPFA        | 2.81 (2.55, 3.09)          | 3.43 (3.15, 3.71) | 2.96 (2.66, 3.25) |
| mMPTA        | −0.27 (−1.59, 0.43)        | 1.49 (1.14, 1.91) | 1.05 (0.89, 1.24) |
| Femur length | 0.08 (0.06, 0.11)          | 0.15 (0.12, 0.18) | 0.11 (0.1, 0.13)  |
| Leg length   | −0.05 (−0.14, −0.02)       | 0.12 (0.1, 0.14)  | 0.1 (0.08, 0.11)  |
| LLD          | −0.43 (−1.56, 0.09)        | 1.36 (1.14, 1.57) | 1.12 (0.92, 1.34) |
| MAD          | 1.32 (0.72, 2.96)          | 2.88 (2.51, 3.25) | 2.25 (1.95, 2.56) |
| Tibia length | −0.1 (−0.13, −0.08)        | 0.11 (0.09, 0.13) | 0.08 (0.07, 0.09) |

Additionally, there are significant time savings associated with using AI to perform these measurements. The LAMA AI system generated reads more than 90% faster than the expert readers. Although there were some errors made by the AI with regard to landmark detection, these errors are readily detectable by viewing the AI output. Those who look at the AI-generated reports for quality control do not need to be fellowship-trained MSK radiologists to see clear errors, such as those shown in the supplemental figures section. The interpreter is always in control of the output data to make clinically meaningful decisions.

### Limitations

One limitation of this study is that images went through a thorough QA process, during which the majority of the initial images were excluded, to determine if they were adequate to send to the AI. However, if prospective imaging can be standardized, the output data success rates from the software would be improved.

Importantly, 11 cases were excluded in the primary analysis due to the AI placing incorrect landmarks. This was determined by a visual inspection of every report. It is important to note that in the analysis including cases with incorrect landmarks, the AI software was still useful for most measurements; 11/13 rather than 12/13 ICCs were in the excellent range, and 9/13 variables were accepted in Bland–Altman analysis rather than 11/13. However, because errors could be significant, any future use of this product should include visual inspection of the reports to ensure adequate quality. Furthermore, software updates to ensure adequacy or to prevent erroneous report generation should be implemented before it is widely used.

Another limitation is that for the time savings analysis, different processes were used for radiologist and AI reading time; radiologists used start and stop buttons, whereas a stopwatch was used to time AI processing. We also did not record the time that radiologists used to look at each AI report. However, the magnitude of the savings is so substantial that this limitation has little bearing on the study results.

Although we performed a sensitivity analysis choosing a single leg randomly which showed only minor differences in results, this analysis did not fully account for potential data dependencies and clusters in the data.

Additionally, there were no clear, universal thresholds found in the literature for clinically relevant differences in all of these measurements. Because of this, we used conservative estimates of maximum allowed differences. The AI performed well on most parameters, notably including the most frequently used measurements — HKA and leg length.

### Future directions

This study assessed the AI's ability to perform LLD and angular measurements on radiographs without any implants. Future studies may evaluate the AI's performance on radiographs with lower-extremity implants and more standardized prospective leg-length radiographs for more generalizability.

### Conclusion

In conclusion, our study showed that, with few exceptions, AI-based leg-length and angular measurements on legs without implants are reliably and accurately obtained with significant time savings as compared to expert radiologists.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s00256-023-04502-5>.

### Declarations

**Conflict of interest** AC: consultant: ICON Medical and TREACE Medical Concepts Inc.; book royalties: Jaypee, Wolters; speaker: Siemens; medical advisor: ImageBiopsy Lab Inc.; research grant: ImageBiopsy Lab Inc.

OA: consultant: ImageBiopsy Lab.

PP: consultant: ImageBiopsy Lab.

JW: consultant: Ethicon.

MD: employee: ImageBiopsy Lab GmbH.

AH: employee: ImageBiopsy Lab GmbH.

### References

- Gordon JE, Davis LE. Leg length discrepancy: the natural history (and what do we really know). *J Pediatr Orthop*. 2019;39(Issue 6, Supplement 1 Suppl 1):S10-s13.
- Siffert RS. Lower limb-length discrepancy. *J Bone Joint Surg Am*. 1987;69(7):1100–6.
- Gurney B. Leg length discrepancy. *Gait Posture*. 2002;15(2):195–206.
- Gofton JP, Trueman GE. Studies in osteoarthritis of the hip. II. Osteoarthritis of the hip and leg-length disparity. *Can Med Assoc J*. 1971;104(9):791–9.
- Applebaum A, Nessim A, Cho W. Overview and spinal implications of leg length discrepancy: narrative review. *Clin Orthop Surg*. 2021;13(2):127–34.
- Rackowski JW, Daniszewska B, Zolynski K. Functional scoliosis caused by leg length discrepancy. *Arch Med Sci*. 2010;6(3):393–8.
- Ghanem I, Karam JA, Widmann RF. Surgical epiphysiodesis indications and techniques: update. *Curr Opin Pediatr*. 2011;23(1):53–9.
- Thienpont E, et al. Bone morphotypes of the varus and valgus knee. *Arch Orthop Trauma Surg*. 2017;137(3):393–400.
- Wang B, et al. Valgus malalignment and prevalence of lateral compartmental radiographic knee osteoarthritis (OA): the Wuchuan OA study. *Int J Rheum Dis*. 2018;21(7):1385–90.

10. Sharma L, et al. Varus and valgus alignment and incident and progressive knee osteoarthritis. *Ann Rheum Dis*. 2010;69(11):1940–5.
11. Schwarz GM, 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*. 2022;30(8):2538–47.
12. Simon S, et al. Fully automated deep learning for knee alignment assessment in lower extremity radiographs: a cross-sectional diagnostic study. *Skeletal Radiol*. 2022;51(6):1249–59.
13. Kallogjeri D, Spitznagel EL Jr, Piccirillo JF. Importance of defining and interpreting a clinically meaningful difference in clinical research. *JAMA Otolaryngol-Head Neck Surg*. 2020;146(2):101–2.
14. Martin Bland J, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;327(8476):307–10.
15. Hodson TO. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geosci Model Dev*. 2022;15(14):5481–7.
16. Obuchowski NA, Subhas N, Schoenhagen P. Testing for interchangeability of imaging tests. *Acad Radiol*. 2014;21(11):1483–9.
17. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess*. 1994;6:284–90.
18. Lu MJ, Zhong WH, Liu YX, Miao HZ, Li YC, Ji MH. Sample size for assessing agreement between two methods of measurement by Bland-Altman method. *Int J Biostat*. 2016;12(2):/j/ijb.2016.12.issue-2/ijb-2015-0039/ijb-2015-0039.xml. <https://doi.org/10.1515/ijb-2015-0039>.
19. Lin YH, Chang FS, Chen KH, Huang KC, Su KC. Mismatch between femur and tibia coronal alignment in the knee joint: classification of five lower limb types according to femoral and tibial mechanical alignment. *BMC Musculoskelet Disord*. 2018;19(1):411. <https://doi.org/10.1186/s12891-018-2335-9>.
20. Marques Luis N, Varatojo R. Radiological assessment of lower limb alignment. *EFORT Open Rev*. 2021;6(6):487–94.
21. Bellemans J, et al. The Chitranjan Ranawat Award: is neutral mechanical alignment normal for all patients?: the concept of constitutional varus. *Clin Orthop Relat Res*. 2012;470(1):45–53.
22. Shimizu M, et al. Examination of the changes in lower extremities related to progression of adult spinal deformity: a longitudinal study of over 22 years. *Sci Rep*. 2020;10(1):11605.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

## Authors and Affiliations

Holden Archer<sup>1</sup> · Seth Reine<sup>1</sup> · Shuda Xia<sup>1</sup> · Louis Camilo Vazquez<sup>1</sup> · Oganesh Ashikyan<sup>1</sup> · Parham Pezeshk<sup>1</sup> · Ajay Kohli<sup>1</sup> · Yin Xi<sup>1</sup> · Joel E. Wells<sup>2</sup> · Allan Hummer<sup>3</sup> · Matthew Difranco<sup>3</sup> · Avneesh Chhabra<sup>1,4,5,6</sup>

✉ Avneesh Chhabra  
avneesh.chhabra@utsouthwestern.edu

<sup>1</sup> University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390, USA

<sup>2</sup> Baylor, Scott, & White, Dallas, TX, USA

<sup>3</sup> IB Lab GmbH, Zehetnergasse 6/2/2, 1140 Vienna, Austria

<sup>4</sup> Adjunct Faculty, Johns Hopkins University, Baltimore, MD, USA

<sup>5</sup> University of Dallas, Richardson, TX, USA

<sup>6</sup> Walton Centre for Neurosciences, Liverpool, UK