Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Research article

Comparison of an AI-driven planning tool and manual radiographic measurements in total knee arthroplasty

Marie Theres Heller, Guenther Maderbacher, Marie Farina Schuster, Lina Forchhammer, Markus Scharf, Tobias Renkawitz, Stefano Pagano *

Department of Orthopedic Surgery, University of Regensburg, Asklepios Klinikum, Bad Abbach, Germany

ARTICLE INFO

Keywords:
Total Knee Arthroplasty
Artificial Intelligence
Automated Planning
Radiographic Measurement
Intraclass Correlation Coefficient
Measurement Efficiency

ABSTRACT

Background: Accurate preoperative planning in total knee arthroplasty (TKA) is essential. Traditional manual radiographic planning can be time-consuming and potentially prone to inaccuracies. This study investigates the performance of an AI-based radiographic planning tool in comparison with manual measurements in patients undergoing total knee arthroplasty, using a retrospective observational design to assess reliability and efficiency.
Methods: We retrospectively compared the Autoplan tool integrated within the mediCAD software (mediCAD Hectec GmbH, Altdorf, Germany), routinely implemented in our institutional workflow, to manual measurements performed by two orthopedic specialists on pre- and postoperative radiographs of 100 patients who underwent elective TKA. The following parameters were measured: leg length, mechanical axis deviation (MAD), mechanical lateral proximal femoral angle (mLPFA), anatomical mechanical angle (AMA), mechanical lateral distal femoral angle (mLDFA), joint line convergence angle (JLCA), mechanical medial proximal tibial angle (mMPTA), and mechanical tibiofemoral angle (mTFA).
Intraclass correlation coefficients (ICCs) were calculated to assess measurement reliability, and the time required for each method was recorded.
Results: The Autoplan tool demonstrated high reliability (ICC $>$ 0.90) compared with manual measurements for linear parameters (e.g., leg length and MAD). However, the angular measurements of mLPFA, JLCA, and AMA exhibited poor reliability (ICC $<$ 0.50) among all raters. The Autoplan tool significantly reduced the time required for measurements compared to manual measurements, with a mean time saving of 44.3 seconds per case (95 % CI: 43.5–45.1 seconds, $p < 0.001$).
Conclusion: AI-assisted tools like the Autoplan tool in mediCAD offer substantial time savings and demonstrate reliable measurements for certain linear parameters in preoperative TKA planning. However, the observed low reliability in some measurements, even amongst experienced human raters, suggests inherent challenges in the radiographic assessment of angular parameters. Further development is needed to improve the accuracy of automated angular measurements, and to address the inherent variability in their assessment.

1. Introduction

Meticulous preoperative planning constitutes an essential component of orthopedic surgery, particularly in total knee arthroplasty (TKA) and total hip arthroplasty (THA) [1]. This planning includes the accurate assessment of the leg axis, the evaluation of potential axis deviations, and the selection of the most appropriate implant prosthesis [2–4]. It also offers numerous advantages [1]. For instance, anticipating implant component sizes can streamline hospital logistics by optimizing implant inventory and sterilization costs [5]. Furthermore, conducting systematic preoperative planning can help surgeons avoid unforeseen intraoperative challenges and has the potential to reduce operative time [6]. Ultimately, rigorous planning contributes to an optimally prepared surgical team and procedure, which in turn facilitates the best possible postoperative outcomes for each patient [1]. Achieving such preparedness necessitates a precise and critical assessment of the leg axis, potential leg-length discrepancies, and other relevant limb characteristics [2,7]. Consequently, detailed and accurate analysis of preoperative radiographs is imperative [1]. Interactive software programs are routinely employed in clinical practice for surgical planning, enabling surgeons to

* Corresponding author.
E-mail address: stefano.pagano@ukr.de (S. Pagano).

https://doi.org/10.1016/j.csbj.2025.04.009
Received 31 December 2024; Received in revised form 7 April 2025; Accepted 8 April 2025
Available online 12 April 2025

perform detailed measurements of the affected limb [3]. However, this interactive approach presents limitations, such as inaccuracies in landmark placement, variability among different users, and the time commitment required for measurement execution [8–10].

In recent years, the increasing adoption of AI-assisted programs in the medical marketplace has been noted, reflecting a general trend toward progressively automated applications in healthcare [11,12]. Potential applications span various fields, including cardiovascular and neurological imaging [13–15]. Measurement accuracy may thus be increased, and human error potentially reduced, owing to the standardized, automated procedures of AI-driven software. Significant time savings are also frequently offered by such tools [7,8,16]. In orthopedic radiology, these automated measurement programs demonstrate particular promise for preoperative planning, especially given the extensive mechanical considerations inherent in arthroplasty surgery [17,18].

Numerous programs and applications currently assist in measuring lower-limb alignment angles for arthroplasty planning [3,19]. Many of these tools function with minimal user intervention, potentially mitigating human error and variability [8,10,20]. Beyond primary arthroplasty, AI-assisted planning has broadened into other orthopedic applications, such as revision arthroplasty—for example, in the identification of in-situ implants in preparation for revision surgery or in the assessment of dislocation risk following THA [21–24]. A more efficient preoperative planning process and improved patient care with better postoperative results are generally observed [8,25]. Nevertheless, despite these promising developments, rigorous evaluation is required before such AI-based tools can become standard practice [26–28].

A previous study by our team evaluated the performance of an AI-supported measurement tool (LAMA, IB Lab GmbH, Vienna, Austria) for radiographic planning in TKA, highlighting both the potential and limitations of automated lower limb alignment analysis [7]. Building upon this prior experience, the present study examines the performance of the Autoplan tool in mediCAD (mediCAD Hectec GmbH, Altdorf, Germany) in comparison with manual radiographic measurements performed by two orthopedic surgeons, applying the same methodological approach. The primary objective was to assess whether Autoplan can provide accurate and consistent measurements of key lower-limb alignment parameters on pre- and postoperative radiographs in TKA patients, while also improving workflow efficiency in routine clinical practice. The Autoplan tool in mediCAD's 2D planning software potentially offers rapid, automated landmark detection for preoperative radiographs and, consequently, swift endoprosthetic planning [29]. Underlying these tools is AI trained through techniques such as transfer learning, intersection-over-union, region-based convolutional neural networks (R-CNNs), and deep convolutional networks [30–33], enabling the efficient identification of anatomic landmarks.

2. Material and methods

2.1. Study data, inclusion, and exclusion criteria

We retrospectively reviewed 200 archived radiographs (pre- and postoperative) from 100 patients (50 women, 50 men) who had undergone TKA in the past seven years at our institution (Department of Orthopedic Surgery, University of Regensburg, Germany). All included patients had appropriate radiographic imaging both before and after surgery. Radiographs were retrieved from our PACS archive, beginning with the first available scan dated January 2, 2018. Each image was reviewed for eligibility based on predefined inclusion and exclusion criteria, as well as image quality. Only cases meeting all requirements were included in the final dataset. Included were patients who met the following criteria: aged 18 years or older, underwent primary TKA for gonarthrosis within the last seven years, had standardized full-length AP standing lower-extremity radiographs both pre- and postoperatively, and had digital radiographs obtained within the last seven years. Exclusion criteria included fractures of the operative leg visible on the radiograph, postoperative radiographs showing implant failure, and preoperative radiographs demonstrating the presence of implants in the knee (e.g. previous TKA, unicondylar knee arthroplasty, high tibial osteotomy, screws, or plates for periarticular fractures), poor image quality that prevented consistent identification of key landmarks, and cases where the indication for TKA was for reasons other than gonarthrosis (e.g. trauma, neoplastic lesion). The variables assessed included leg length measured according to the Mikulicz line (from the femoral head center to the middle point of the distal tibial joint line), mechanical axis deviation (MAD), mechanical lateral proximal femoral angle (mLPFA), anatomical mechanical angle (AMA), mechanical lateral distal femoral angle (mLDFA), joint-line convergence angle (JLCA), mechanical medial proximal tibial angle (mMPTA), and mechanical tibiofemoral angle (mTFA).

We also recorded the time required for measurements by both the Autoplan tool and by manual procedure. For the Autoplan tool, measurement time was defined strictly as the automated calculation period without human interaction. For manual measurements, time started after the calibration of the radiographic image in the software and ended upon completion of all required landmarks and measurements for that image.

2.2. Evaluation of the radiographs (Autoplan vs. Manual)

Each radiograph was analyzed using two methods: (1) the Autoplan tool by mediCAD (Version 7.0), and (2) manual measurements by two orthopedic specialists performed with the same MediCAD Software (Fig. 1). Rater A (SP) was a resident with less than 5 years of orthopedic experience, and Rater B (GM) was a senior surgeon with a decade of professional experience. Both orthopedic specialists independently assessed the same 200 radiographs, blinded to the Autoplan tool results. Rater A repeated the measurements on the same images four weeks later to assess intra-rater reliability.

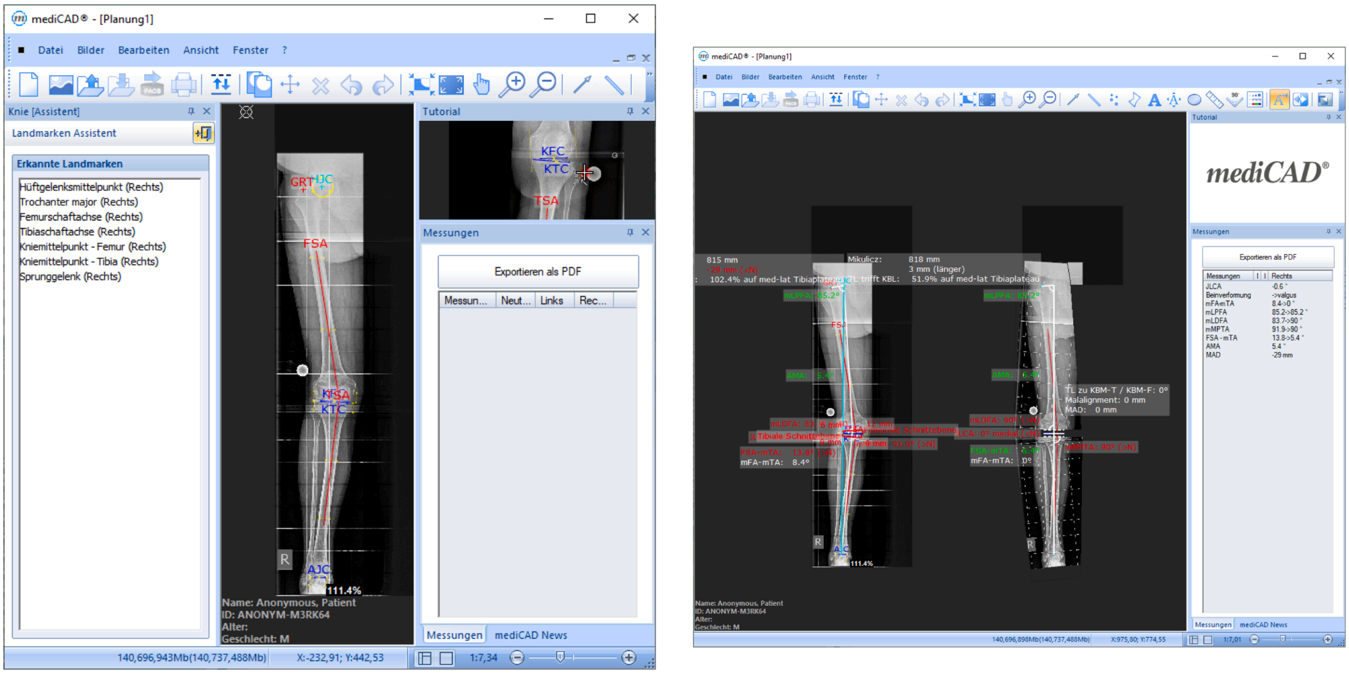
Our clinic routinely includes a calibration sphere in all preoperative radiographs as a standard reference. This radiopaque marker, placed at a known distance from the X-ray source, is essential for accurately scaling the images so that real-world distances can be reliably extracted from the radiograph. However, since postoperative radiographs do not include the calibration sphere, we applied indirect calibration by referencing known dimensions from the preoperative radiograph (e.g., femoral head size or hip prosthesis head diameter). These landmarks served as surrogate reference points for the postoperative measurements.

2.3. Statistical analysis

Intraclass correlation coefficients (ICCs) were used to evaluate the reliability of lower-limb alignment measurements obtained by the Autoplan tool and the two independent orthopedic specialists. ICCs were calculated using a two-way mixed-effects model (single measurement, absolute agreement) and interpreted per established guidelines [34]. Inter-rater reliability was determined by comparing Autoplan tool-derived values to each orthopedic specialist's measurements, as well as comparing the three measurers to one another. Intra-rater reliability was assessed by comparing Rater A's two sets of measurements. Pre- and postoperative readings were analyzed separately to examine reliability differences between these conditions.

A "gold standard" reference was established by calculating the mean of three measurements: Rater A's first and second measurements and Rater B's single measurement. The automated measurements from the Autoplan tool were then compared to this average. The Wilcoxon Signed-Rank Test was used to compare ICC values for preoperative versus postoperative conditions. Descriptive statistics (mean, standard deviation) were calculated for all measurements.

Bland-Altman plots were used to visualize agreement for both



Erkannte Landmarken:
Hüftgelenksmittelpunkt (Rechts)
Trochanter major (Rechts)
Femurschaftachse (Rechts)
Tibiaschaftachse (Rechts)
Kniemittelpunkt - Femur (Rechts)
Kniemittelpunkt - Tibia (Rechts)
Sprunggelenk (Rechts)

Messungen	I	I	Rechts
JLCA			-0.6 °
Beinverformung			->valgus
mFA-mTA			8.4->0 °
mLPFA			85.2->85.2 °
mLDFA			83.7->90 °
mMPTA			91.9->90 °
FSA - mTA			13.8->5.4 °
AMA			5.4 °
MAD			-29 mm

Fig. 1. Leg alignment measurement performed by the Autoplan tool (MediCAD). The upper window (A) displays the anatomical landmarks recognized by the AI. The lower screenshot (B) presents the AI's measurement results in the summary box located on the right of the screen.

angular measurements (AMA, JLCA) and linear parameters (MAD), plotting the mean of the two measurement methods against their difference. Limits of agreement (LoA) were defined according to established clinical thresholds: ± 2° for angular measurements and ± 5 mm for linear parameters [7,9]. Furthermore, the measurement completion times were explicitly compared between the Autoplan tool and each manual measurement instance. Paired t-tests were employed to evaluate statistically significant differences in measurement duration between the different methods and raters.

All statistical analyses were performed in SPSS (Version 29, IBM Corp., Armonk, NY, USA). Statistical significance was set at two-sided $p < 0.05$. Ethical approval for this study was obtained from the University of Regensburg's Ethics Committee (reference 24–3937–104, 29 October 2024).

3. Results

A total of 200 radiographic images (pre- and postoperative) from 100 patients were included in the study. Of these, proper measurement by the Autoplan tool was not feasible in three cases. In one patient case, automated measurement could not be conducted preoperatively and postoperatively because of an inlaying femoral nail. In a second case, automated measurement could not be executed postoperatively, as the femoral head was not identifiable due to excessive soft tissue density. As a result, 197 radiographs were ultimately available for performance evaluation and statistical analysis involving the Autoplan tool.

In ten additional cases, preoperative implanted foreign material was present. Eight cases involved an ipsilateral inlying endoprosthesis following THA, one case involved foreign material in the osseous pelvis, and one case involved a plate and cerclage in the proximal femur region together to an inlying hip endoprosthesis. However, these instances did not impact the practicability of the measurements; the Autoplan tool could still perform the measurements.

In three cases, the reference sphere could not be detected preoperatively by the software. Manual scaling of the sphere was then performed, allowing measurement to proceed automatically as intended. As the reference sphere was used solely preoperatively, this issue occurred exclusively in preoperative radiographs.

In two cases, a reference sphere was detected that was not present in the postoperative radiograph. The software then incorrectly identified certain osseous structures as scaling spheres. However, because postoperative measurement scaling was performed using the femur and its preoperative measurements as reference, this issue was not a significant concern.

ICCs demonstrated overall high reliability for leg length, MAD, mLDFA and mTFA in both pre- and postoperative conditions (Table 1). Notably, MAD exhibited excellent agreement (ICC > 0.90) between the Autoplan tool and both orthopedic specialists, indicating strong consistency for this parameter (Table 2). Conversely, mLPFA, JLCA, AMA exhibited poor reliability (ICC < 0.50) across all comparisons.

Comparisons between preoperative and postoperative measurements revealed generally somewhat lower ICCs following surgery for certain parameters, such as mLDFA and mMPTA (Table 2). However, the Wilcoxon Signed-Rank Test indicated no statistically significant differences in overall ICC values between pre- and postoperative conditions for Leg Length, MAD, and mTFA ($p > 0.05$).

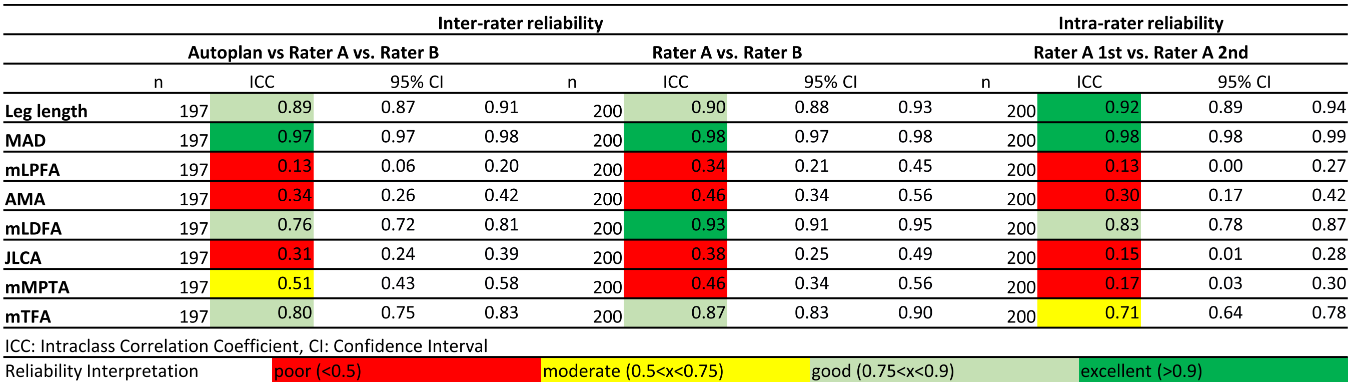
Bland-Altman plots (Fig. 2) further illustrate these findings. MAD measurements clustered tightly around the zero-difference line, while AMA and JLCA showed greater scatter, particularly at higher angle values (Figs. 2B and 2C).

Postoperative Bland-Altman plots (Fig. 3) revealed similar patterns. MAD demonstrated high agreement, with most data points falling within the limits of agreement (± 5 mm), confirming strong postoperative reliability (Fig. 3 A). In contrast, angular measurements like AMA and JLCA continued to show noticeable variability, especially at higher values (Figs. 3B and 3 C). Although variability decreased slightly for JLCA postoperatively, proportional biases remained evident, indicating discrepancies increased with larger measurements.

Marked differences were observed in the time required for measurement completion between manual measurements and the automated measurement by the Autoplan tool. The Autoplan tool required a significantly shorter duration for completion (mean: 2.4 seconds, SD: 1.6 seconds) compared to manual measurements by either orthopedic specialist (mean range: 43.2 – 48.3 seconds, $p < 0.001$) (Table 3).

Additional descriptive statistics, including variability in linear and angular measurements, are presented in Table 3.

Table 1
Intraclass correlation coefficients (ICC) for evaluating the reliability of measurements across raters. The table includes ICC values with 95 % confidence intervals, comparing the Autoplan tool with both physician raters, inter-rater reliability between both physician raters, and intra-rater reliability for repeated measurements by one rater. The total number of measurements is reported, combining preoperative and postoperative cases. Lower ICCs observed for angular parameters such as mLPFA, JLCA, and AMA reflect known variability in landmark-based measurement and interpretation [7].

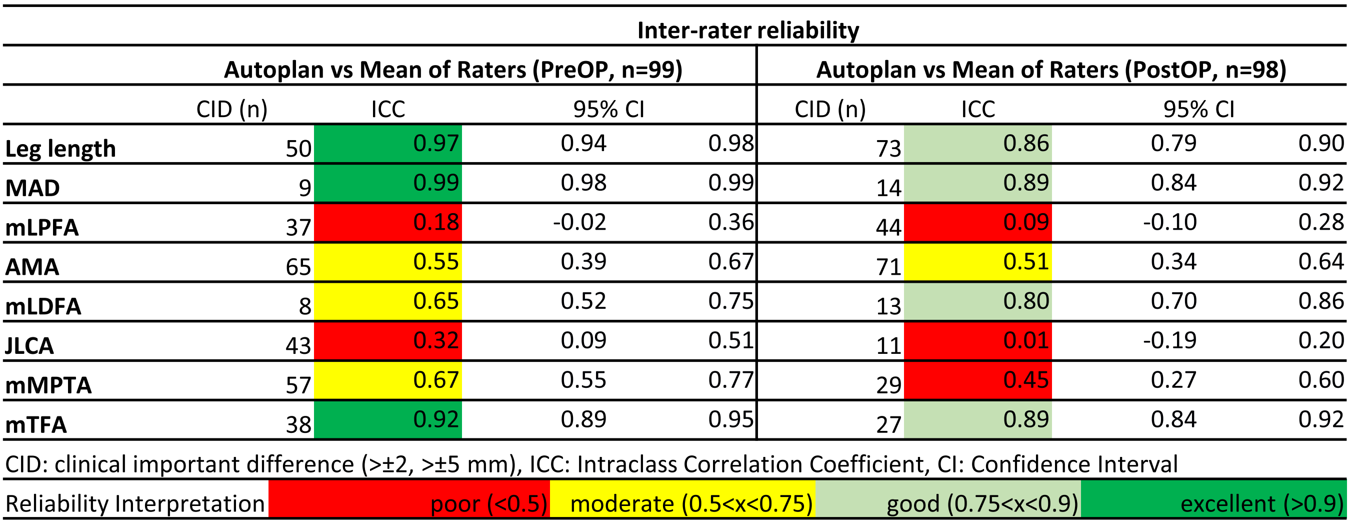


	Inter-rater reliability								Intra-rater reliability			
	Autoplan vs Rater A vs. Rater B				Rater A vs. Rater B				Rater A 1st vs. Rater A 2nd			
	n	ICC	95% CI		n	ICC	95% CI		n	ICC	95% CI	
Leg length	197	0.89	0.87	0.91	200	0.90	0.88	0.93	200	0.92	0.89	0.94
MAD	197	0.97	0.97	0.98	200	0.98	0.97	0.98	200	0.98	0.98	0.99
mLPFA	197	0.13	0.06	0.20	200	0.34	0.21	0.45	200	0.13	0.00	0.27
AMA	197	0.34	0.26	0.42	200	0.46	0.34	0.56	200	0.30	0.17	0.42
mLDFA	197	0.76	0.72	0.81	200	0.93	0.91	0.95	200	0.83	0.78	0.87
JLCA	197	0.31	0.24	0.39	200	0.38	0.25	0.49	200	0.15	0.01	0.28
mMPTA	197	0.51	0.43	0.58	200	0.46	0.34	0.56	200	0.17	0.03	0.30
mTFA	197	0.80	0.75	0.83	200	0.87	0.83	0.90	200	0.71	0.64	0.78

ICC: Intraclass Correlation Coefficient, CI: Confidence Interval

Reliability Interpretation	poor (<0.5)	moderate (0.5<x<0.75)	good (0.75<x<0.9)	excellent (>0.9)

Table 2
ICC values comparing measurements from the Autoplan tool with the gold standard derived from the mean values of orthopedic raters. Preoperative and postoperative measurements are analyzed separately. Clinically relevant thresholds (CID, ± 2° for angles and ± 5 mm for lengths) are applied to assess agreement and highlight key discrepancies across parameters.



	Inter-rater reliability							
	Autoplan vs Mean of Raters (PreOP, n=99)				Autoplan vs Mean of Raters (PostOP, n=98)			
	CID (n)	ICC	95% CI		CID (n)	ICC	95% CI	
Leg length	50	0.97	0.94	0.98	73	0.86	0.79	0.90
MAD	9	0.99	0.98	0.99	14	0.89	0.84	0.92
mLPFA	37	0.18	-0.02	0.36	44	0.09	-0.10	0.28
AMA	65	0.55	0.39	0.67	71	0.51	0.34	0.64
mLDFA	8	0.65	0.52	0.75	13	0.80	0.70	0.86
JLCA	43	0.32	0.09	0.51	11	0.01	-0.19	0.20
mMPTA	57	0.67	0.55	0.77	29	0.45	0.27	0.60
mTFA	38	0.92	0.89	0.95	27	0.89	0.84	0.92

CID: clinical important difference (>±2, >±5 mm), ICC: Intraclass Correlation Coefficient, CI: Confidence Interval

Reliability Interpretation	poor (<0.5)	moderate (0.5<x<0.75)	good (0.75<x<0.9)	excellent (>0.9)

4. Discussion

The findings of the present study indicate that the Autoplan tool provides notable advantages; however, limitations exist in the accuracy of certain measurements when compared to specialist assessments. Our results demonstrated high reliability and efficiency for linear parameters, such as MAD and leg length. Given these advantages, including the significant reduction in time required, automated measurement by the Autoplan tool clearly offers an advantage over conventional manual measurements for linear parameters. Interestingly, the poor reliability of mLPFA, JLCA, AMA (ICC $<$ 0.50) was not limited to the comparison between the Autoplan tool and manual measurements, but was also apparent among the two experienced orthopedic specialists. This suggests that the variability may stem from the inherent limitations in radiographic landmark visibility, particularly postoperatively where implants can obscure reference structures. Previous work has shown that even under standardized imaging conditions and across different levels of clinical experience, alignment measurements are subject to altered anatomy, difficulty in defining midpoints on prosthetic surfaces, or subtle changes in limb positioning [43]. Such variability likely contributes to the reduced agreement observed in our study and underscores the need for enhanced standardization in both imaging protocols and measurement algorithms [3].

Similar to our findings with the Autoplan tool, Hoffmann et al., who also evaluated this software, reported excellent accuracy for leg length (ICC $>$ 0.99) and overall alignment (ICC $>$ 0.97), but encountered difficulties with joint-level angles, particularly in postoperative measurements [9]. It should be noted, however, that their ICCs were calculated by comparing AI measurements to the average of two observers, whereas our analysis compared AI directly to each rater individually. This may partly account for the slightly higher ICC values reported in their study.

Our team's previous investigation by Pagano et al. evaluated another AI-supported automated measurement program (LAMA), which

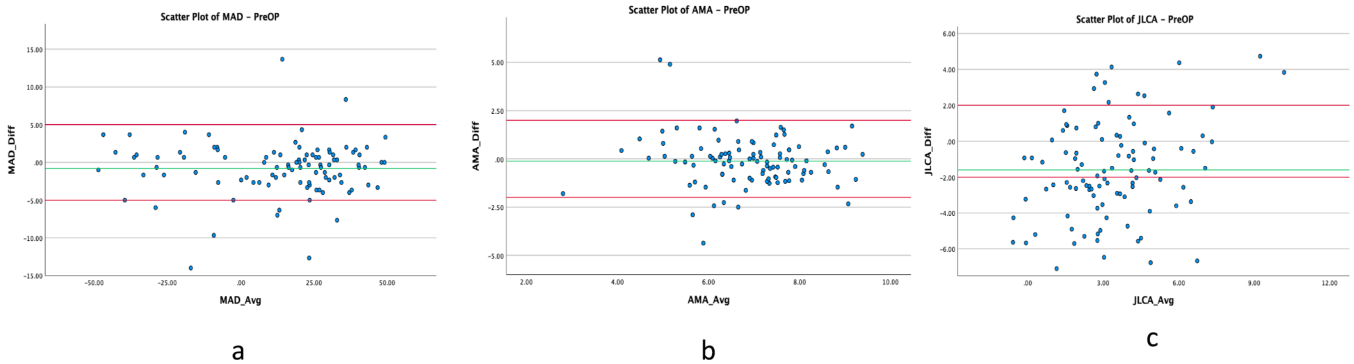


Fig. 2. Fig. 2 - Bland-Altman scatter plots for preoperative measurements of MAD (a), AMA (b), and JLCA (c). The Y-axis represents the differences between the Autoplan tool and the mean of all rater measurements, while the X-axis shows the averages of these measurements. The green line represents the mean difference, which is close to zero for most parameters, indicating minimal systematic bias. The red lines denote clinically relevant limits of agreement (± 2° for angles and ± 5 mm for lengths). Substantial disagreement is observed for JLCA, with many values exceeding the acceptable range.

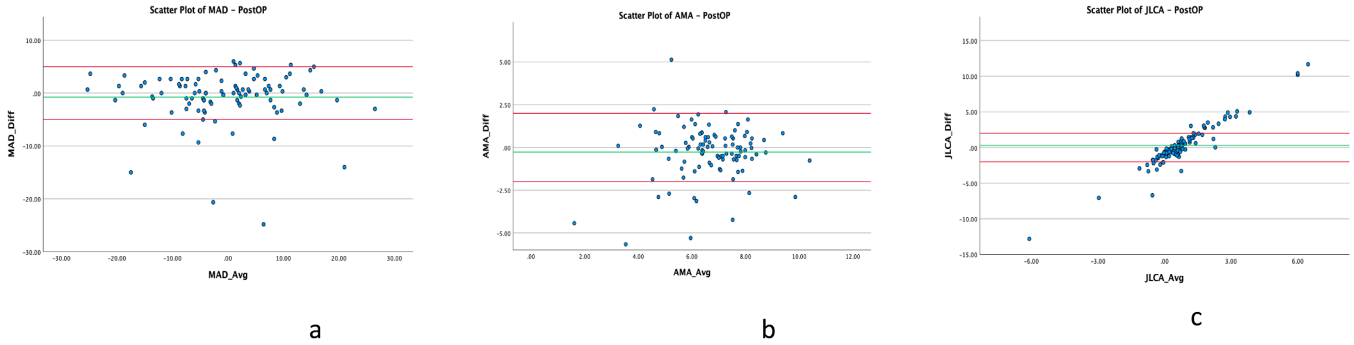


Fig. 3. Bland-Altman scatter plots for postoperative measurements of MAD (a), AMA (b), and JLCA (c). Agreement improves significantly for JLCA compared to preoperative measurements, with fewer outliers and reduced bias. However, a proportional trend suggests variability at higher values, likely due to measurement errors in the Autoplan tool.

Table 3
Overview of all preoperative and postoperative measurements conducted by the Autoplan tool, the less experienced rater (A) who repeated the same measurements (1st) after four weeks (2nd), and the more experienced rater (B). Mean and standard deviation values are presented for each parameter to summarize the variability across raters and measures.

	Autoplan (n = 197)		Rater A 1st (n = 200)		Rater A 2nd (n = 200)		Rater B (n = 200)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Leg length (mm)	790.6	54.8	792.5	53.8	789.3	53.9	792.0	60.6
MAD (mm)	5.8	20.1	6.6	20.2	6.6	20.0	6.5	20.7
mLPFA (°)	95.1	18.8	93.7	22.3	90.9	5.8	90.4	6.6
AMA (°)	6.7	1.5	6.8	2.6	7.2	2.0	6.7	1.1
mLDFA (°)	89.0	3.5	88.9	2.6	89.1	2.6	89.1	3.1
JLCA (°)	1.8	3.0	2.3	2.3	2.8	6.2	2.6	2.6
mMPTA (°)	89.2	4.3	88.8	3.2	88.2	6.0	88.6	3.1
mTFA (°)	–1.7	6.0	–1.7	6.0	–1.6	8.6	–1.9	5.9
Time (s)	2.4	1.6	48.3	6.6	44.9	7.4	43.2	4.2

n: total measurements, SD: Standard Deviation

demonstrated good to excellent agreement for several parameters (e.g., MAD with ICC = 0.98), but showed lower ICC values for angular measurements and particularly struggled with postoperative radiographs and in obese patients. Building on that work, the present study utilized the same radiographic dataset to compare the performance of the mediCAD Autoplan tool under identical imaging and statistical conditions. While both AI tools encountered challenges with angular parameters such as JLCA and AMA, the Autoplan tool exhibited improved measurement feasibility, with a lower failure rate and more consistent landmark detection across pre- and postoperative images (Table 4). These differences underscore the variability in performance among AI-based tools and the need for independent validation of each system before integration into clinical workflows [7].

Substantial differences were observed in the time required for measurement completion between manual measurements and automated results from the Autoplan tool. The Autoplan tool operated considerably faster throughout the study. While a trend of reduced time expenditure was observed for the orthopedic specialists due to a learning effect, the Autoplan tool presented a substantial advantage in terms of efficiency.

These results provide valuable insights into the potential benefits of integrating automated AI-driven measurement in routine clinical practice. With few exceptions, the Autoplan tool functioned effectively for most examined radiographs, both pre- and postoperatively, showing strong agreement with the results obtained by the orthopedic specialists,

Table 4
Intraclass correlation coefficients (ICCs) reported in the present study, Pagano et al. [7], and Hoffmann et al.[9] for evaluating the agreement between AI-based software (Autoplan from MediCAD and LAMA) and human raters in radiographic planning for TKA. Preoperative and postoperative values are reported separately when available. Sample sizes represent the number of radiographs or limbs analyzed per study. Across all studies, ICCs were consistently high for global alignment parameters such as mechanical tibiofemoral angle (mTFA) and leg length, while joint-level angular parameters (e.g., JLCA, mLDFA, mMPTA) showed lower agreement.

	Current study (Autoplan)			Pagano et al. (LAMA)			Hoffmann et al. (Autoplan)		
	n	ICC (PreOP)	ICC (PostOP)	n	ICC (PreOP)	ICC (PostOP)	n	ICC (PreOP)	ICC (PostOP)
Leg length	197	0.97	0.86	142	0.95	0.69	164	0.99	1.00
MAD	197	0.99	0.89	150	1.00	0.99	-	-	-
mLPFA	197	0.18	0.09	155	0.93	0.94	164	0.78	0.55
AMA	197	0.55	0.51	155	0.89	0.73	164	0.82	0.89
mLDFA	197	0.65	0.80	155	0.83	0.98	164	0.81	0.63
JLCA	197	0.32	0.01	150	0.49	0.47	164	0.88	0.11
mMPTA	197	0.67	0.45	155	0.89	0.93	164	0.89	0.34
mTFA	197	0.92	0.89	155	0.99	1.00	164	1.00	0.98

ICC: Intraclass Correlation Coefficient.

Reliability Interpretation	poor (<0.5)	moderate (0.5<x<0.75)	good (0.75<x<0.9)	excellent (>0.9)

particularly for linear parameters. Through automated planning, some potential sources of human error and investigator-dependent deviations could be reduced, and measurements could be performed more uniformly. In this way, a substantial increase in efficiency may be achieved, making the clinical workflow to become more standardized and faster, especially in a high-volume clinical setting.

In their studies, other authors drew similar conclusions, suggesting that automation of measurements, reduction of inter- and intra-observer variability, and time savings could be achieved through the integration of AI-based automated planning in clinical workflow [10,35,36]. For instance, Schock et al. [10] demonstrated significant time efficiency gained through automated analysis of lower limb alignment, a finding directly mirroring our observation. Furthermore, Bonnin et al. [35] provided evidence for the potential of AI to enhance standardization and reduce interobserver variability in radiographic assessments. Lambrechts et al. [36], also in the context of total knee arthroplasty planning, found that AI-generated plans were comparable to surgeon-approved plans, further supporting the idea that automation can streamline workflows without necessarily sacrificing accuracy, particularly in certain aspects of planning.

However, limitations and errors in the autonomous application of AI were also evident during the automated measurement by the Autoplan tool. The observed poor reliability of angular measurements, which was also present in our manual assessments, reflects the difficulty of consistently defining landmarks in complex anatomy or when interfering factors are present. Specific factors that interfered with the software, resulted in either erroneous results or an inability to perform measurements. For example, patient-specific factors, such as the presence of foreign material (e.g., an inlaying femoral nail) or obesity, could impede automated measurement by preventing accurate recognition of the femoral head as a reference point. Consistent with findings from prior investigations of other AI-based measurement software [7,9], the present study also encountered challenges in automatically measuring certain radiographs due to image complexities and observed lower reliability in angular measurements compared to linear measurements. The AI, in its current form, seems to reflect the inherent challenges that also exist in manual radiographic assessment of these specific angles. It is possible that the AI is consistently identifying and measuring landmarks, but the underlying variability in manual landmark identification contributes to the low agreement. This finding also raises questions about the ’gold standard’ itself, as significant variability even amongst experienced raters suggests that a single ’true’ value for these angular measurements might be difficult to ascertain from radiographs alone [44].

A potential explanation for the challenges in angular measurements lies in certain limitations of current algorithms. As highlighted by Lambrechts et al. [37], manufacturers’ default plans often necessitate significant surgeon modification. Their work demonstrates machine learning’s potential for generating more surgeon- and patient-specific plans, thus reducing subsequent need for corrections.

Several limitations of this study warrant consideration when interpreting the results. The automated assessment depends on the image quality of the radiographs. Various confounding factors can influence the automated measurement process, potentially leading to inaccurate results. These factors include image quality, contrast reduction due to soft tissue shadows, the method of image acquisition (e.g., rotation of the legs, affecting the position of landmarks such as the greater trochanter or center of the knee), or incomplete image capture (e.g., showing only a portion of the femoral head). When interpreting the results, particularly concerning their potential future application in clinical practice, the possibility of bias arising from the authors’ evaluation of new approaches compared to established standards should be acknowledged. Therefore, the available results limit the ability to draw broad conclusions about the general use of AI in preoperative planning. While the study included a postoperative scaling method based on anatomical references (e.g., femoral head diameter), the accuracy and repeatability of this indirect approach were not formally evaluated. We acknowledge that this may introduce variability in postoperative linear measurements. However, as the same anatomical reference was used consistently across both manual and automated measurements for each case, the relative comparison remains valid.

Further research is required to establish standardized usage of automated measurement tools in future routine clinical practice. Initially, the algorithms underlying these tools require further refinement through machine learning to better handle angle measurements and adapt to diverse anatomical conditions, as this was the most challenging aspect identified in our findings. This refinement can be achieved through further validation by collecting larger, diverse datasets, and conducting real-world clinical studies to assess generalizability and utility [37–39]. To enhance the accuracy and reliability of automated measurements, using complementary AI-based tools, where appropriate, seems advisable for a more comprehensive approach to preoperative planning [40].

5. Conclusion

While some caution against exclusive reliance on fully automatic algorithms due to accuracy concerns compared to manual measurements [41,42], our findings also highlight the limitations and inherent variability within manual measurements, particularly for angular parameters. The Autoplan tool demonstrates clear benefits in efficiency and reliability for linear measurements, but further development is needed to improve angular assessments, both for automated tools and to address the challenges in achieving consistent manual angular measurements.

CRediT authorship contribution statement

Renkawitz Tobias: Supervision. **Pagano Stefano:** Writing – review & editing, Validation, Methodology, Formal analysis, Conceptualization. **Schuster Marie Farina:** Investigation, Data curation. **Forchhammer Lina:** Investigation, Data curation. **Scharf Markus:** Investigation, Data curation. **Heller Marie:** Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation. **Maderbacher Guenther:** Formal analysis, Data curation.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Gemini Version 2.0 from Google AI Studio exclusively for grammar correction and text style refinement of the original manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of Competing Interest

Tobias Renkawitz declares the following financial interests: Research funding at personal disposal from DePuy, Zimmer, Aesculap, German Federal Ministry of Education and Research, Deutsche Arthrose-Hilfe, OttoBock-Stiftung, German Federal Ministry of Economic and Development, Oskar-Helene-Heim Foundation in Berlin, Vielberth Foundation, and Deutsche Forschungsgemeinschaft (DFG). He has received reimbursement of costs from DePuy, Zimmer, Aesculap, Federal Ministry of Education and Research, Deutsche Arthrose-Hilfe, OttoBock-Stiftung, Federal Ministry for Economic Cooperation and Development, Oskar-Helene-Heim Foundation in Berlin, Vielberth Foundation, DGOOC, BVOU, and DGOU. He has also received reimbursement of costs for training/lectures from DePuy, Zimmer, Aesculap, German Society for Endoprosthetics (AE), and Bavarian Association of General Practitioners. His non-financial interests include his position as Chair (W3) of Orthopedic Surgery, University of Regensburg, and Director of the Orthopedic Department for the University Hospital Regensburg. He holds memberships as Vice President of the Professional Association for Orthopedics and Trauma Surgery (BVOU), in the General Board of the DGOOC, as Head of the "Arbeitsgemeinschaft Evidenzbasierte Medizin" of the DGOU, on the editorial board of "Die Orthopädie" and "Die Unfallchirurgie" (SpringerMedicine), as editor of the "Orthopädie und Unfallchirurgie—Mitteilungen und Nachrichten" (OUMN), and on the International Advisory Board of the Journal of the American Academy of Orthopedic Surgeons (AAOS).

Acknowledgements

Conceptualization, S.P.; Methodology, M.H., S.P.; Validation, M.H., S.P.; Formal Analysis, M.H., G.M., S.P.; Investigation, M.H., M.S., L.F., M.S., S.P.; Data Curation, M.H., M.S., L.F., M.S., G.M., S.P.; Supervision, T.R.; Writing – Original Draft Preparation, M.H.; Writing – Review & Editing, S.P.

References

[1] Tanzer M, Makhdom AM. Preoperative planning in primary total knee arthroplasty. J Am Acad Orthop Surg 2016;24(4):220–30. https://doi.org/10.5435/JAAOS-D-14-00332.

[2] Marques Luís N, Varatojo R. Radiological assessment of lower limb alignment. EFORT Open Rev 2021;6:487–94.

[3] Khalifa AA, Mullaji AB, Mostafa AM, Farouk OA. A protocol to systematic radiographic assessment of primary total knee arthroplasty. Orthop Res Rev 2021;13:95–106.

[4] Kniesel B, Konstantinidis L, Hirschmüller A, Südkamp N, Helwig P. Digital templating in total knee and hip replacement: an analysis of planning accuracy. Int Orthop 2014;38(4):733–9. https://doi.org/10.1007/s00264-013-2157-1.

[5] Hafez MA, Moholkar K. Patient-specific instruments: advantages and pitfalls. SICOT-J 2017;3:66. https://doi.org/10.1051/sicotj/2017054.

[6] Rodrigues AST, Gutierres MAP. Patient-specific instrumentation in total knee arthroplasty. Should we adopt it? Rev Bras De Ortop 2016;52(3):242–50. https://doi.org/10.1016/j.rboe.2016.06.008.

[7] Pagano S, Müller K, Götz J, et al. The role and efficiency of an AI-powered software in the evaluation of lower limb radiographs before and after total knee arthroplasty. J Clin Med 2023;12(17):5498. https://doi.org/10.3390/jcm12175498.

[8] Chen K, Stotter C, Klestil T, Nehrer S. Artificial intelligence in orthopedic radiography analysis: a narrative review. Diagnostics 2022;12(9):2235. https://doi.org/10.3390/diagnostics12092235.

[9] Hoffmann C, Göksu F, Klöpfer-Krämer I, et al. High accuracy in lower limb alignment analysis using convolutional neural networks, with improvements needed for joint-level metrics. Knee Surg, Sports Traumatol, Arthrosc 2024. https://doi.org/10.1002/ksa.12481.

[10] Schock J, Truhn D, Abrar DB, Merhof D, Conrad S, Post M, Mittelstrass F, Kuhl C, Nebelung S. Automated analysis of alignment in long-leg radiographs by using a fully automated support system based on artificial intelligence. Radiol Artif Intell 2020;3(2):e200198. https://doi.org/10.1148/ryai.2020200198.

[11] Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. Gastrointest Endosc 2020;92(4):807–12. https://doi.org/10.1016/j.gie.2020.06.040.

[12] Crossnohere NL, Elsaid M, Paskett J, et al. Guidelines for artificial intelligence in medicine: literature review and content analysis of frameworks. J Med Internet Res 2022;24(8):e36823. https://doi.org/10.2196/36823.

[13] Al'Aref SJ, Anchouche K, Singh G, et al. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. Eur Heart J 2019;40(24):1975–86. https://doi.org/10.1093/eurheartj/ehy404.

[14] den Boer RB, de Jongh C, Huijbers WTE, et al. Computer-aided anatomy recognition in intrathoracic and -abdominal surgery: a systematic review. Surg Endosc 2022;36(12):8737–52. https://doi.org/10.1007/s00464-022-09421-5.

[15] Jiang J, Wang D, Song Y, et al. Computer-aided extraction of select MRI markers of cerebral small vessel disease: a systematic review. NeuroImage 2022;261:119528. https://doi.org/10.1016/j.neuroimage.2022.119528.

[16] Bousson V, Benoist N, Guetat P, Attané G, Salvat C, Perronne L. Application of artificial intelligence to imaging interpretations in the musculoskeletal area: where are we? Where are we going? Jt Bone Spine 2023;90(1):105493. https://doi.org/10.1016/j.jbspin.2022.105493.

[17] von Eisenhart-Rothe R, Hinterwimmer F, Graichen H, et al. Artificial intelligence and robotics in TKA surgery: promising options for improved outcomes? Knee Surg, Sports Traumatol, Arthrosc 2022;30:2535–7. https://doi.org/10.1007/s00167-022-07035-x.

[18] Mavrogenis AF, Megaloikonomos PD, Panagopoulos GN, Maffulli N. Biomechanics in orthopaedics. J Biomed 2017;2:89–93. https://doi.org/10.7150/jbm.19088.

[19] Tiefenboeck S, Sesselmann S, Taylor D, Forst R, Seehaus F. Preoperative planning of total knee arthroplasty: reliability of axial alignment using a three-dimensional planning approach. Acta Radiol 2022;63(8):1051–61. https://doi.org/10.1177/02841851211029076.

[20] Pei Y, Yang W, Wei S, et al. Automated measurement of hip–knee–ankle angle on the unilateral lower limb X-rays using deep learning. Phys Eng Med Biol 2021;44: 53–62. https://doi.org/10.1007/s13246-020-00951-7.

[21] Murphy M, Killen C, Burnham R, et al. Artificial intelligence accurately identifies total hip arthroplasty implants: a tool for revision surgery. Hip Int 2022;32(6): 766–70. https://doi.org/10.1177/1120700020987526.

[22] Loppini M, Gambaro FM, Chiappetta K, et al. Automatic identification of failure in hip replacement: an artificial intelligence approach. Bioengineering 2022;9(7): 288. https://doi.org/10.3390/bioengineering9070288.

[23] Rouzrokh P, Ramazanian T, Wyles CC, et al. Deep learning artificial intelligence model for assessment of hip dislocation risk following primary total hip arthroplasty from postoperative radiographs. J Arthroplast 2021;36(6): 2197–2203.e3. https://doi.org/10.1016/j.arth.2021.02.028.

[24] Rouzrokh P, Mickley JP, Khosravi B, et al. THA-AID: deep learning tool for total hip arthroplasty automatic implant detection with uncertainty and outlier quantification. J Arthroplast 2024;39(4):966–973.e17. https://doi.org/10.1016/j.arth.2023.09.025.

[25] Graves ML. The value of preoperative planning. J Orthop Trauma 2013;27 1: S30–4. https://doi.org/10.1097/BOT.0b013e3182a52626.

[26] Chea P, Mandell JC. Current applications and future directions of deep learning in musculoskeletal radiology. Skelet Radiol 2020;49(2):183–97. https://doi.org/10.1007/s00256-019-03284-z.

[27] Longo UG, De Salvatore S, Valente F, Villa Corta M, Violante B, Samuelsson K. Artificial intelligence in total and unicompartmental knee arthroplasty. BMC Musculoskelet Disord 2024;25(1):571. https://doi.org/10.1186/s12891-024-07516-9.

[28] Kurmis AP. A role for artificial intelligence applications inside and outside of the operating theatre: a review of contemporary use associated with total knee arthroplasty. Arthroplasty 2023;5(1):40. https://doi.org/10.1186/s42836-023-00189-0.

[29] *2D knee*. mediCAD. (2024, August 16). ⟨https://medicad.eu/produkte/2d-classic/2d-knie/?lang=en⟩.

[30] Mwiti D. *Transf Learn Guide: A Pract Tutor Ex Images Text Keras* Neptune ai 2024, April 22. ⟨https://neptune.ai/blog/transfer-learning-guide-examples-for-images-and-text-in-keras⟩.

[31] Rosebrock A. Intersection over union (IOU) for object detection. PyImageSearch; 2024, November 18. ⟨https://pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/⟩.

[32] Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2017;39(6): 1137–49. https://doi.org/10.1109/TPAMI.2016.2577031.

[33] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, editors. *Medical image computing and computer-assisted intervention – MICCAI 2015* (Lecture Notes in Computer Science, 9351. Cham: Springer; 2015. https://doi.org/10.1007/978-3-319-24574-4_28.

[34] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15(2):155–63. https://doi.org/10.1016/j.jcm.2016.02.012.

[35] Bonnin M, Müller-Fouarge F, Estienne T, Bekadar S, Pouchy C, Ait Si Selmi T. Artificial intelligence radiographic analysis tool for total knee arthroplasty. J Arthroplast 2023;38(7 2):S199–207.e2. https://doi.org/10.1016/j.arth.2023.02.053.

[36] Lambrechts A, Ganapathi M, Wirix-Speetjens R. Clinical evaluation of artificial intelligence-based preoperative plans for total knee arthroplasty. CAOS 2020 (EPiC Ser Health Sci, 4) 2020:169–73. https://doi.org/10.29007/9c6c.

[37] Lambrechts A, Wirix-Speetjens R, Maes F, Van Huffel S. Artificial intelligence based patient-specific preoperative planning algorithm for total knee arthroplasty. Front Robot AI 2022;9:840282. https://doi.org/10.3389/frobt.2022.840282.

[38] Amin A, Cardoso SA, Suyambu J, Abdus Saboor H, Cardoso RP, Husnain A, Isaac NV, Backing H, Mehmood D, Mehmood M, Maslamani ANJ. Future of artificial intelligence in surgery: a narrative review. Cureus 2024;16(1):e51631. https://doi.org/10.7759/cureus.51631.

[39] Hinterwimmer F, Lazic I, Langer S, Suren C, Charitou F, Hirschmann MT, Matziolis G, Seidl F, Pohlig F, Rueckert D, Burgkart R, von Eisenhart-Rothe R. Prediction of complications and surgery duration in primary TKA with high accuracy using machine learning with arthroplasty-specific data. Knee Surg, Sports Traumatol, Arthrosc: J ESSKA 2023;31(4):1323–33. https://doi.org/10.1007/s00167-022-06957-w.

[40] Batailler C, Shatrov J, Sappey-Marinier E, Servien E, Parratte S, Lustig S. Artificial intelligence in knee arthroplasty: current concept of the available clinical applications. Arthroplasty 2022;4(1):17. https://doi.org/10.1186/s42836-022-00119-6.

[41] Seaver T, McAlpine K, Garcia E, Niu R, Smith EL. Algorithm based automatic templating is less accurate than manual digital templating in total knee arthroplasty. J Orthop Res: Publ Orthop Res Soc 2020;38(7):1472–6. https://doi.org/10.1002/jor.24696.

[42] Farooq H, Deckard ER, Arnold NR, Meneghini RM. Machine learning algorithms identify optimal sagittal component position in total knee arthroplasty. J Arthroplast 2021;36(7S):S242–9. https://doi.org/10.1016/j.arth.2021.02.063.

[43] Bowman A, Shunmugam M, Watts AR, Bramwell DC, Wilson C, Krishnan J. Inter-observer and intra-observer reliability of mechanical axis alignment before and after total knee arthroplasty using long leg radiographs. Knee 2016;23(2):203–8. https://doi.org/10.1016/j.knee.2015.11.013.

[44] Maderbacher G, Matussek J, Greimel F, Grifka J, Schaumburger J, Baier C, Keshmiri A. Lower limb malrotation is regularly present in long-leg radiographs resulting in significant measurement errors. J Knee Surg 2021;34(1):108–14. https://doi.org/10.1055/s-0039-1693668.