



Accuracy evaluation of an automatic landmarking method based on 3D segmented CT scans of full lower limbs with knee osteoarthritis

Clément Horteur^{a,b}, Nolwenn Fougeron^a, Benoît Gaulin^{a,b}, Ysé Roch^{a,c}, Marek Bucki^c, Estelle Palluel^a, Yohan Payan^a, Antoine Perrier^{a,c,d,*} 

^a Univ. Grenoble Alpes, CNRS UMR 5525, VetAgro Sup, Grenoble INP, TIMC, 38000 Grenoble, France

^b Department of Orthopedics and Sport Surgery, Grenoble-Alpes CHU South Teaching Hospital, Kimberley Avenue, BP 338, 38434 Échirolles Cedex, France

^c Twinsight, Grenoble, France

^d Department of Orthopedics and Traumatology, Groupe Hospitalier Diaconesses Croix St-Simon, Paris, France

ARTICLE INFO

Keywords:

Lower limb

Knee

Automatic registration

Landmarking

Statistical evaluation

ABSTRACT

In preoperative planning, a number of techniques for registration of anatomical landmarks on bones, or “bones landmarking”, have been developed for automatic measurement of lower limbs. However, the evaluation of these techniques is often neglected, reducing confidence in results and limiting clinical applications. We propose a protocol to evaluate landmarking techniques applied to thirty 3D images of osteoarthritic lower limbs. The protocol includes the evaluation of the ground truth based on manual landmarking performed by two experienced clinicians. Intra- and inter-operator variability were quantified. Automatic landmarking was evaluated using the same procedure and compared to manual landmarking by computing distances with bounding boxes including manual landmarks. The method was applied to 87 landmarks on the femur, tibia, fibula, and patella. For intra-operator variability, the mean median error was 2.0 mm, with a maximum of 42.1 mm. For inter-operator variability, errors ranged from 0.0 to 43.5 mm, with a mean median error of 2.3 mm across all landmarks. The mean median error between manual and automatic landmarking methods was 2.4 mm, with a range from 0.0 mm to 22.8 mm. Confidence intervals of automatic landmarking errors were consistent with intra- and inter-operator variability confidence intervals. Approximately 42 % of the automatic landmarks were inside their bounding boxes, with a mean distance of 1.4 mm from the box for landmarks on the outside. Results showed different errors depending on landmark location and coordinates. Intra- and inter-operator variability shed light on the difficulty of defining ground truth based on manual segmentation. This finding emphasizes the need for a rigorous evaluation protocol for landmarking validation, including the verification of statistical test hypotheses and population size influence.

1. Introduction

Standards of care in orthopaedic lower limb surgery have evolved substantially over recent decades. Improvements in surgical tools, implants and accurate pre-operative planning have enhanced surgeon precision. Navigation systems[1,2], patient-specific instrumentation[3] and 3-Dimensional (3D) image-based planning[4] were developed in the late 1990 s and have been used primarily for joint replacements and osteotomies. Moreover, knee arthroplasty (KA) has seen a paradigm shift over the last decade, moving from conventional “one fits all” techniques to more personalised implantation that matches the patient’s native anatomy[5,6], thus requiring precise surgical planning and

instrumentation. Despite these technological advances, many surgical procedures are still planned using 2-Dimensional (2D) radiograph images and performed with conventional instrumentation. The potential cost and insufficient evidence for the superiority of these new tools compared to conventional techniques represent barriers that must be eliminated for these technologies to be widely adopted worldwide. For knee arthroplasty, several personalized alignment philosophies utilizing these technologies have been proposed to improve patient’ satisfaction. These approaches aim to enhance gap balancing and patellar tracking (functional alignment[7,8] and inverse kinematic alignment[9]) or achieve pure knee resurfacing that replicates the patient’s native anatomy with (restricted kinematic alignment[10]) or without (kinematic

* Corresponding author at: Department of Orthopedics and Traumatology, Groupe Hospitalier Diaconesses Croix St-Simon, Paris, France.

E-mail addresses: aperrier@hopital-dcss.org, antoine.perrier@univ-grenoble-alpes.fr (A. Perrier).

<https://doi.org/10.1016/j.ejrad.2025.112292>

Received 3 October 2024; Received in revised form 13 May 2025; Accepted 6 July 2025

Available online 8 July 2025

0720-048X/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

alignment[11]) specific boundaries to prevent the reproduction of highly abnormal anatomies.

The cornerstone of reproducing or correcting a patient’s anatomy with high precision is the pre-operative quantification of morphological and alignment parameters. Computed tomography (CT) provides superior lower limb evaluation compared to 2D radiographs, avoiding errors related to X-ray incidence variabilities[12]. CT scans offer 3D shape visualisation of the bone and add informations in the transverse plane [13]. The development of low-dose CT scans has considerably reduced patient irradiation[14], making them a reasonable medical investigation tool prior to orthopaedic lower limb surgery. In clinical routine, lower limb alignment must be estimated based on these medical images to inform preoperative planning[15]. While it is possible to perform 3D CT-based bone landmarking (manually or automatically) and compute lower limb metrics correctly, this practice remains challenging[16,17].

Many authors have proposed semi-automatic and automatic methods for bony landmark placement, especially to assess coronal lower limb alignment. Methods using bone surface curvature[18], cross-section models, shape-fitting models or axes of inertia[19,20] are currently considered semi-automatic because they require manual segmentations, making them impractical for daily medical practice. Automatic methods can be categorised into knowledge-based or artificial intelligence (AI)-based techniques[21]. Few researchers have developed methods specifically for lower limb landmarking. In lower limb studies, such methods have achieved up to 15.8 mm mean error for femoral landmarks and 8.7 mm for tibial landmarks compared to expected landmark positions[22].

The evaluation of semi-automatic or automatic landmarking methods appears to lack a clearly defined procedure for computing errors and evaluating their statistical significance. In many cases, significant methodological limitations jeopardize the relevance of the results. The sample size used to evaluate the method in some studies is insufficient to determine the significance of the results[18,23,24]. The quality of the ground truth and the process of obtaining it can be insufficiently detailed[18,23,24]. Statistical methodology issues must be highlighted in some studies[4,22,25] such as the lack of error significance calculation or the use of a statistical test without meeting its requirements.

Moreover, when intended for clinical purposes, the evaluation results should be compared with clinical standards. In cases where manual landmarking serves as the ground truth, an assessment of repeatability and reproducibility is necessary to establish acceptable error tolerance thresholds for semi-automatic or automatic methods. Finally, to the best of our knowledge, no study has yet evaluated automatic lower limb landmarking from CT scans in patients with degenerative knee deformations (e.g., cartilage wear, osteophytes) resulting from advanced osteoarthritis.

This paper proposes a procedure for evaluating 3D automatic landmarking methods applied to segmented osteoarthritic knee bones. We compared an automatic landmarking method to a ground truth based on manual landmarking performed by expert clinicians. Repeatability and reproducibility errors of manual landmarking were first computed to define acceptable orders of magnitude for errors.

2. Material and methods

2.1. Description of the database

2.1.1. Study participants and datasets

Thirty patients (11 men 65.6 +/- 9.8 years old and 19 women 71.2 +/- 7.8 years old) were included in this study following ethical committee approval (2019-A02124-53). Informed consent was obtained from all patients prior to data extraction. All patients underwent Total Knee Arthroplasty (TKA) at the Grenoble Alpes University Hospital, France. Mean hip-knee-ankle (HKA) angle was 178.3 ± 4.5°. All patients presented degenerative bony deformations of the knee. The proximal femur was free of wear and deformation in all patients. Femoro-tibial,

patellofemoral and ankle wear were reported according to radiographs-based classifications (Ahlbäck[26], Iwano[27] and Takakura[28], respectively). The presence, size and location of osteophytes were described using the method proposed by Hunter et al[29]. (Table 1 and supplementary material E).

Patients who underwent previous geometry altering surgeries of the ipsilateral limb, including the femur, tibia, fibula and patella bones, were excluded. For all patients, presurgical pseudonymised CT scans were available and used to segment and reconstruct the 3D geometries of the femur, tibia, fibula and patella bones of the ipsilateral limb. The CT scans were performed with a Siemens multidetector scanner. The scanning parameters were 140 kV, 49–266 mA, with a slice thickness of

Table 1
Population characteristics (osteoarthritis wear and osteophytes). FT: femoro-tibial, OA: osteoarthritis, PF: patellofemoral,

	Classification Method	Classification: N (%)				
		No lesion	Grade 1	Grade 2	Grade 3	Grade 4
Radiographs analysis (wear)						
Medial FT OA	AHLBÄCK [26]	0 (0 %)	1 (3.3 %)	8 (26.7 %)	15 (50 %)	6 (20 %)
Lateral FT OA	AHLBÄCK [26]	2 (6.7 %)	7 (23.3 %)	11 (36.7 %)	8 (26.7 %)	2 (6.7 %)
PF AO	IWANO[27]	6 (20 %)	7 (23.3 %)	2 (6.7 %)	10 (33.3 %)	5 (16.7 %)
Ankle OA	TAKAKURA [28]	12 (40 %)	6 (20 %)	8 (26.7 %)	4 (13.3 %)	0 (0 %)
CT Scans analysis (presence of osteophytes)						
Medial tibia	MOAKS[29]	7 (23.3 %)	13 (43.3 %)	6 (20 %)	4 (13.3 %)	NA
Lateral tibia		19 (63.3 %)	9 (30 %)	2 (6.7 %)	0 (0 %)	
Medial trochlea		11 (36.7 %)	7 (23.3 %)	8 (26.7 %)	4 (13.3 %)	
Lateral trochlea		14 (46.7 %)	11 (36.7 %)	3 (10 %)	2 (6.7 %)	
Medial posterior condyle		15 (50 %)	6 (20 %)	6 (20 %)	3 (10 %)	
Lateral posterior condyle		20 (66.7 %)	6 (20 %)	3 (10 %)	1 (3.3 %)	
Medial condyle		7 (23.3 %)	10 (33.3 %)	4 (13.3 %)	9 (30 %)	
Lateral condyle		13 (43.3 %)	8 (26.7 %)	7 (23.3 %)	2 (6.7 %)	
Superior patella		23 (76.7 %)	4 (13.3 %)	3 (10 %)	0 (0 %)	
Inferior patella		20 (66.7 %)	9 (30 %)	1 (3.3 %)	0 (0 %)	
Medial patella		19 (63.3 %)	7 (23.3 %)	4 (13.3 %)	0 (0 %)	
Lateral patella		11 (36.7 %)	14 (46.7 %)	5 (16.7 %)	0 (0 %)	
Anterior distal tibia		20 (66.7 %)	8 (26.7 %)	2 (6.7 %)	0 (0 %)	
Posterior distal tibia		29 (96.7 %)	1 (3.3 %)	0 (0 %)	0 (0 %)	

1 mm. After anonymisation, all images were exported according to the left-posterior-superior (LPS) coordinate system normalisation. Consequently, the z-axis corresponds to the proximo-distal axis, the x-axis to the medio-lateral axis and the y-axis to the antero-posterior axis.

2.1.2. Manual landmarking

We defined 87 landmarks distributed on anatomical features of the femur, tibia, fibula and patella. These landmarks were chosen according to their use in the computation of lower limb metrics usually done in orthopaedic surgery by clinicians on X-rays and CT[30]. A description of these landmarks is provided in [supplementary materials A](#).

Two experienced orthopaedic surgeons were trained to position the landmarks based on CT scan slices and 3D reconstructions of the bones. Each landmark was manually placed by operators on 3D CT scans using the axial, coronal, and sagittal planes. Any change in the position of a landmark in one plane was automatically reflected in the other two planes, ensuring spatial consistency. All landmarks were annotated directly on the CT scan slices, and none were placed exclusively on the 3D surface models. A 3D reconstruction of the lower limb was available throughout the process, serving as a visual aid to verify and refine the landmark positioning, but not as the primary annotation interface. Both operators received the same instructions regarding the landmarks' definitions before the landmarking process in a training session. Operator 1 did a preliminary landmarking session of half the subject's database. Then a second training session was performed with both operators to discuss the preliminary landmarking of operator 1. The landmarking used for this study was performed after both training sessions for both operators. The database of thirty patients was randomly divided into two groups of 15 patients: Group (1) for Operator 1 and Group 2 for Operator 2. Each operator performed landmarking twice on their assigned group, with at least seven days between the first (Measure 1) and second (Measure 2) landmarking sessions. Additionally, each operator performed landmarking once on the group assigned to the other operator (Measure 3). Thus, Operator 1 landmarked Group 2, and Operator 2 landmarked Group (1) once. Errors in landmark labelling were corrected retrospectively. Two landmarks were excluded from the analysis due to mispositioning: one on the patella of one patient and another on the femur of a different patient.

Measures 1 and 2 were used to perform a repeatability analysis of both operators, and measures 1 and 3 combined were used for the reproducibility analysis between operators. To evaluate the performance of automatic landmarking, the ground truth was defined as the barycentre of measures 1 and 3 of operator 1 and operator 2. Errors were calculated for each landmark's coordinates from differences between each measure.

2.1.3. Automatic landmarking

To illustrate the evaluation protocol described in this study, automatic landmarking was performed by establishing a mathematical one-to-one correspondence between the 3D triangular meshes of the reference (atlas) and target (patient) bones reconstructed from patient medical images. While this method can be implemented in various ways, ranging from statistical shape modelling to the inference of AI-based landmarks, we provide a brief technical description of our chosen geometrical approach below for context.

The 3D mapping in our automatic landmarking is represented by a 3D displacement (vectors) field that superimposes the 3D mesh of each reference bone onto its counterpart in the patient's space. Computing anatomical shape correspondence is an ill-posed problem as it has no unique well-defined solution, meaning many equally valid mappings can be chosen. Our implementation produces a meaningful space warping that superimposes morphological singularities: concavities and convexities corresponding to bony landmarks such as ligament or tendon insertion points.

The computation of the 3D warping function is performed in 3 steps, each progressively introducing new degrees of freedom. The first step

aims to compensate for the coarse difference in rigid pose between the bones. To this end, an inertial analysis of the reference and target bones is carried out, and the bones are aligned by superimposing their mass centres and aligning their principal inertial axes according to the patient's body orientation derived from the medical images' metadata. The second step anisotropically compensates for disparities in sizes along the principal inertial directions. An Iterative Closest Point (ICP) algorithm [31] coupled with a scaling optimization is iteratively performed to refine the initial rigid Arun[32] registration while accounting for scaling disparities. The third and final step unleashes the remaining degrees of freedom by computing displacement vectors at each vertex of the reference 3D mesh. The final and accurate warping is computed by iteratively applying orthogonal projection and relaxation operators to the 3D warping field until convergence (i.e., until the desired level of mapping accuracy has been reached).

Our method requires a few seconds to warp a 50,000-vertex mesh onto another with similar resolution. The final accuracy (mean and maximal Hausdorff distance) is less than a millimetre. Computations were performed on a workstation equipped with an Intel Core i7 processor and 32 GB of RAM, using only the CPU (no GPU acceleration).

2.2. Statistical analyses

2.2.1. Scaling and positioning of patient datasets

As the size, position and orientation of bones in the CT scanner may differ among all patients, we ensured that these differences did not influence the computation of landmark positioning errors. We used the landmarks defined during Measure 1 (i.e., the first landmarking of both groups of patients) to perform a generalised Procrustes analysis between all patients[33]. Before this analysis, left lower limb reconstructions were symmetrised to enable the comparison with right lower limbs. Transform matrices were defined for all patients according to the calculated Procrustes distances. Considering that patients may have various knee angles and relative bone dimensions, Procrustes distances were computed independently for each bone. For each bone in each patient, the same transform matrix, computed for measure 1, was applied to all measures to obtain similar positions, orientations and scales among patients but also to keep the same relative errors among landmarking measures. The Procrustes distance and the resulting patients' bones transform matrices were computed following the procedure described hereafter:

- 1- Arbitrarily choose a reference patient in measure 1;
- 2- Superimpose all patients of measure 1 by applying a translation equal to the distance between each patient's landmark centroid and the reference patient's centroid.
- 3- Normalise patients' size by applying a scaling factor so that the root mean square distance of the points to the centroid of the landmarks is one for each patient (Eq. (1)). N is the total number of landmarks, s is the scaling factor of the cloud of points, x, y, z are the coordinates of each landmark, and $\bar{x}, \bar{y}, \bar{z}$ are the mean coordinates of all the landmarks of the patient.

$$s = \frac{1}{\sum_{i=1}^N (x_i - \bar{x})^2 + (y_i - \bar{y})^2 + (z_i - \bar{z})^2} \quad (1)$$

- 4- Minimise distances between landmarks of each patient and the reference patient with a rotation matrix in an iterative process;

- 5- Compute the mean patient, defined as the average of the patients' landmarks positions.

- 6- Rescale both the mean and reference patients with the reference patient scaling factor and compute the root mean square error of the position of the landmarks between both patients.

- 7- If the root mean square error is larger than an arbitrarily chosen convergence value, then define the mean patient as the new reference patient and go over step 2 to step 6 again. Otherwise, save the transform matrix, including translation, scaling and rotation for the bone and

patient of interest.

8- Rescale all patients' landmarks using the mean patient scaling factor.

9- Apply the same transform matrix and rescaling for the same bone and patient in measures 2 and 3.

Steps 6 and 8 were included to obtain results that could be compared in millimetres to facilitate the interpretation. The convergence value was set to 1 mm, which resulted in 2 to 3 iterations of the generalised Procrustes analysis for all bones. Various patients were tested to define the reference patient in step 1, but resulting differences in the computation of the errors were negligible. Eventually, patient 6 was chosen as the reference patient since the generalised Procrustes analysis required fewer iterations to complete with this patient. Considering that the algorithm required the same landmarks for all patients, both discarded landmarks detailed in §1.2 were removed to compute the transform matrices. Except for the two patients for which these landmarks were discarded, the transformation matrices were applied to the complete set of landmarks for each patient.

2.2.2. Repeatability of the manual landmarking

Based on statistical principles, a sample size of thirty subjects is usually considered sufficient to ensure the validity of parametric analyses. This sample size is consistent with commonly accepted thresholds in statistical practice for studies involving normal distribution and Student t-tests. Consequently, this threshold was used in this study, however the normality and homoscedasticity of the distribution were checked in data processing to apply the appropriate tests. A repeatability analysis was performed for both operators across the three coordinates,

x, y, and z, of all landmarks. Consequently, a total of 522 (i.e., 2 operators x 3 directions x 87 landmarks) repeatability analyses were conducted to assess the repeatability confidence intervals of each coordinate of each landmark independently. The normality and homoscedasticity of each error distribution between measures 1 and 2 were evaluated before each repeatability analysis to determine if the paired Student's *t*-test parametric test or the Wilcoxon signed rank non-parametric test should be applied (Fig. 1). In both cases, a significance level of 5 % was applied. A Shapiro-Wilk test was used to assess the normality of the error distributions with a significance level of 5 %. The conclusion of the Shapiro-Wilk tests was also qualitatively compared to Quantile-Quantile (Q-Q) plots of the same error distributions. The homoscedastic property was evaluated for errors showing a normal distribution using the Levene test with a significance level of 5 %. In addition to the error confidence intervals for each operator and coordinate of the landmarks, the minimum and maximum limit and median values were also computed.

2.2.3. Reproducibility of the manual landmarking

A reproducibility analysis was performed to compare the landmarking of both operators. This procedure was identical to the one described for the repeatability analysis except that errors were computed between measure 1 and measure 3 (Fig. 1) for all the dataset (thirty patients).

2.2.4. Evaluation of the automatic landmarking

The final ground truth for algorithm validation was defined as the average of combined measures 1 and 3 from both operators, as no single

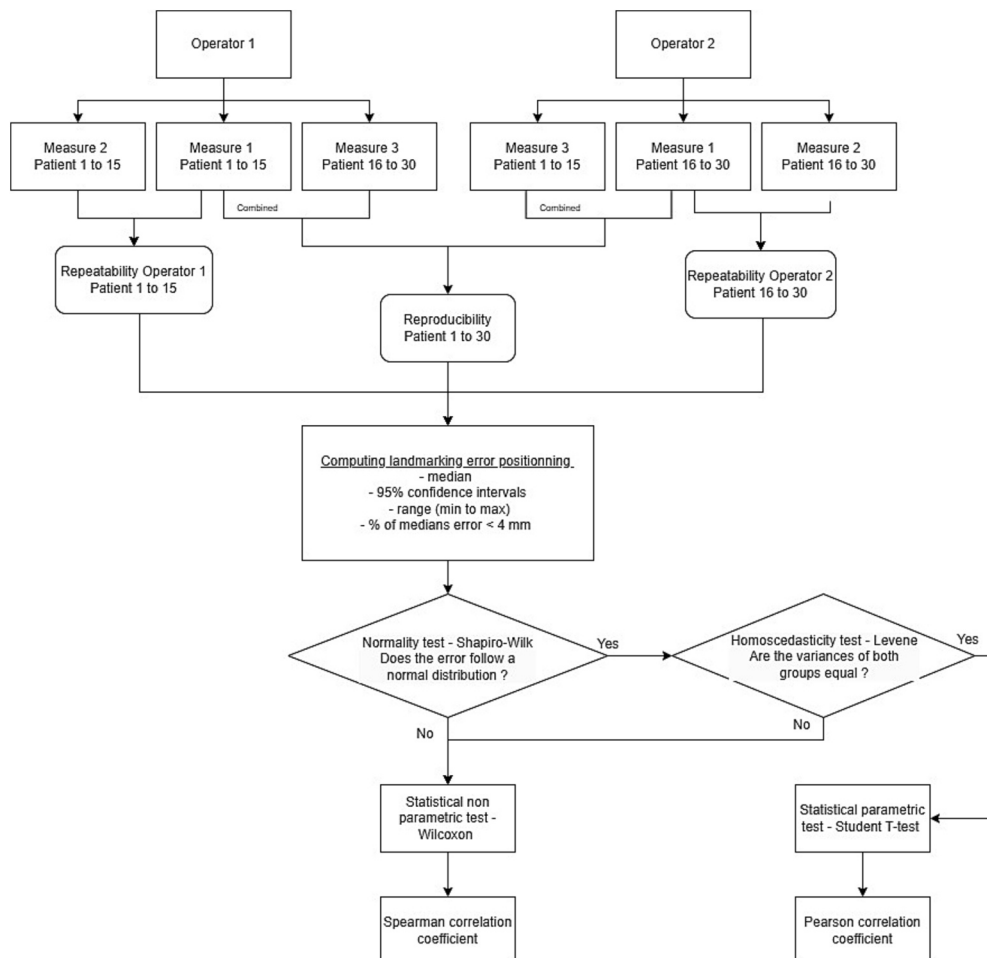


Fig. 1. Summary diagram of repeatability and reproducibility analyses.

operator can be considered the definitive reference due to the inherent subjectivity and variability in anatomical landmark placement. This procedure was identical to the one described previously for the repeatability and reproducibility analyses. The confidence intervals of the errors between the automatic and manual landmarking were compared to those obtained from the repeatability and reproducibility analyses. To further evaluate the performance of automatic landmarking in comparison to manual landmarking, we employed bounding boxes. This approach is more suitable than bounding spheres as it enables understanding of the influence of individual axes on the error. For each landmark in each patient, a bounding box was created with the absolute minimum and maximum coordinates of the three landmark coordinates of the three manual measures of both operators (measures 1, 2, and 3). Each patient's automatic landmarks were then tested to determine if their coordinates fell within these bounding boxes. If an automatic landmark was outside the bounding box, we determined the distances beyond the box per coordinate by measuring from the automatic landmark to the closest edge of the bounding box along each direction.

3. Results

3.1. Repeatability and reproducibility of the manual landmarking

Presenting the results for all analyses for all 41 femoral, 39 tibial, and 7 patellar bone landmarks across three directions (anterior-posterior, medio-lateral, and proximo-distal) would be cumbersome. Therefore, to enhance the clarity of our findings, this section is organized into two parts: an overall summary of results of the 87 critical landmarks and an in-depth review of selected landmarks, serving as examples to better illustrate and understand the statistical flow. Detailed results of both repeatability studies and the reproducibility studies for all the 87 landmarks are available in [Supplementary Material B](#).

3.1.1. Global overview of manual landmarking errors

An example of final manual landmarking performed during measures

1, 2 and 3 is presented in [Fig. 2](#). Specifically, landmarks for patients 1 and 25 are shown in [Fig. 2a](#) for the repeatability analysis of operator 1 (referred to as repeatability 1), and the landmarks for patients 6 and 9 are presented in [Fig. 2b](#) for the repeatability analysis of operator 2 (referred to as repeatability 2). Additionally, [Fig. 2c](#) illustrates the reproducibility analysis results between operator 1 and operator 2 for patients 1, 25, 6, and 9. Since not all distributions of errors in repeatability 1 and 2 and the reproducibility analyses were normal and homoscedastic, results are expressed in terms of the median of the error for each landmark across all patients. To simplify the presentation of the results, the mean of these medians was computed. [Table 2](#) provides a comprehensive summary of the mean results across bones and coordinates, including median errors and maximal values with their corresponding location. To provide further insight into the distributions of errors, the percentage of the median of the errors below 4 mm with 95 % Confidence Intervals (CI) were computed. Further detailed results for repeatability 1, repeatability 2, and reproducibility analyses across all landmarks and coordinates are respectively available in the [supplementary materials B. a\), B. b\), and B. c\)](#).

Each bone's results are summarized by calculating the mean of its landmark-specific measurements. This process involves several steps:

1. For each landmark, the differences between landmark positions gathered from measure 1 and measure 2 are computed in all 3 spatial directions for all 15 subjects in Group (1).
2. The median of these 15 differences is then calculated in x, y, and z for each landmark.
3. Finally, the mean of all per-landmark median differences is computed by considering all landmarks of a given bone.

For example, in the case of the femoral bone along the medio-lateral axis, the reported value of 1.2 mm represents the mean of the median errors across all 41 femoral landmarks in the repeatability analysis of operator 1.

For both repeatability analyses, femoral and tibial landmarks on the

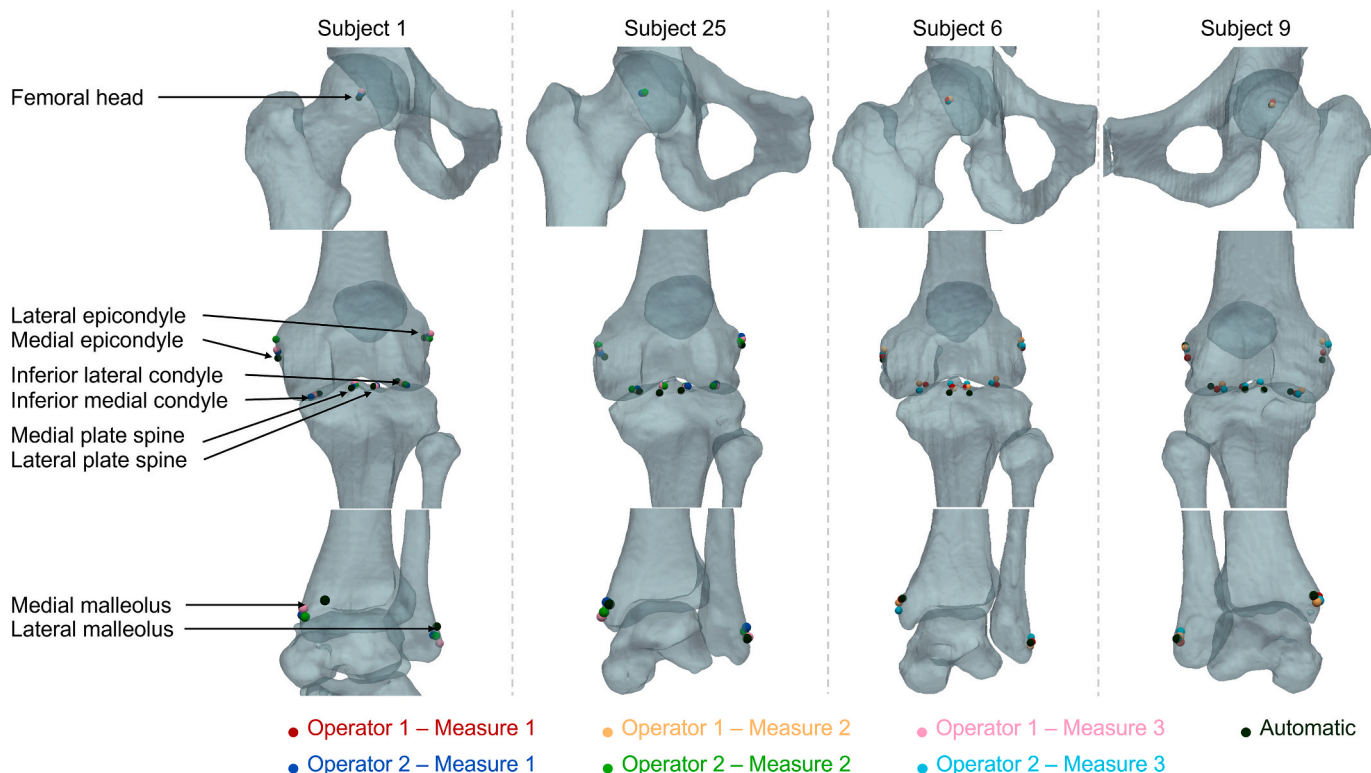


Fig. 2. Positioning of the 9 landmarks of interest done by both operators for all three measures and by the automatic algorithm.

Table 2

Summary results for repeatability 1, repeatability 2 and reproducibility analysis by bone and coordinate.

Analysis		Mean of the median errors (mm)	95 % CI (mm)	Maximum Error Location (mm)	Medians < 4 mm (%)
Femur					
X	Repeatability 1	1.2	[0.9; 2.1]	Femoral Inf Post Shaft (9.3 mm)	100.0
	Repeatability 2	1.5	[1.1; 2.7]	Femoral Sup Sup Ant Shaft (14.4 mm)	100.0
	Reproducibility	1.5	[1.3; 2.5]	Femoral Sup Sup Ant Shaft (14.7 mm)	100.0
Y	Repeatability 1	1.2	[0.8; 2.1]	Femoral Lateral Post Condyle (14.5 mm)	100.0
	Repeatability 2	1.6	[1.2; 2.9]	Femoral Lateral Inferior Condyle (16.1 mm)	95.1
	Reproducibility	1.6	[1.4; 2.7]	Femoral Neck Inf (17.9 mm)	92.7
Z	Repeatability 1	3.2	[2.5; 5.9]	Femoral Mid Ant Shaft (36.5 mm)	68.3
	Repeatability 2	3.8	[2.7; 6.6]	Femoral Inf Inf Med Shaft (32.7 mm)	51.2
	Reproducibility	3.5	[3.2; 5.8]	Femoral Inf Inf Med Shaft (31.6 mm)	61.0
Tibia					
X	Repeatability 1	1.9	[1.3; 3.0]	Tibial Plate Medial Sup-Post (13.5 mm)	89.7
	Repeatability 2	1.6	[1.3; 3.1]	Tibial Inf Post Shaft (14.5 mm)	97.4
	Reproducibility	2.2	[1.8; 3.1]	Tibial Inf Post Shaft (15.0 mm)	89.7
Y	Repeatability 1	1.5	[1.1; 2.6]	Tibial Sup Med Shaft (12.5 mm)	100.0
	Repeatability 2	1.3	[1.0; 2.7]	Tibial Medial Plate (12.6 mm)	100.0
	Reproducibility	1.8	[1.6; 2.9]	Tibial Plate Spine Medial (23.3 mm)	92.3
Z	Repeatability 1	3.4	[2.9; 7.1]	Tibial Inf Lat Shaft (28.6 mm)	59.0
	Repeatability 2	3.6	[2.7; 7.4]	Tibial Inf Med Shaft (42.1 mm)	59.0
	Reproducibility	4.1	[3.9; 7.5]	Tibial Lat Inf Lat (43.5 mm)	56.4
Patella					
X	Repeatability 1	1.2	[0.8; 2.0]	Patellar Apex Tendon Sup (7.9 mm)	100.0
	Repeatability 2	1.1	[0.7; 1.7]	Patellar Apex Tendon Sup (7.6 mm)	100.0
	Reproducibility	1.2	[1.1; 1.9]	Patellar Apex Cartilage Inf (7.7 mm)	100.0
Y	Repeatability 1	0.7	[0.5; 1.2]	Patellar Apex Inf (4.6 mm)	100.0
	Repeatability 2	0.7	[0.6; 1.5]	Patellar Apex Tendon Sup (7.7 mm)	100.0
	Reproducibility	0.8	[0.8; 1.6]	Patellar Apex Tendon Sup (8.9 mm)	100.0
Z	Repeatability 1	1.0	[0.9; 1.7]	Patellar Apex Ant (4.1 mm)	100.0
	Repeatability 2	1.7	[1.2; 2.8]	Patellar Apex Lat (6.8 mm)	100.0
	Reproducibility	1.1	[1.1; 2.1]	Patellar Apex Lat (7.5 mm)	100.0

z-axis (proximo-distal) exhibited means of median errors between 3.0 and 4.1 mm. Results of repeatability 1 showed that most errors were below 10.0 mm, with notable exceptions in the x- and y-axes (1 out of 87 and 5 out of 87 landmarks, respectively). However, the z-axis displayed the highest number of landmarking errors above 10.0 mm (12 out of 87 landmarks), particularly near the femoral and tibial shafts (9 out of 12 landmarks). Repeatability 2 revealed similar results, although with a higher proportion of errors exceeding 10.0 mm compared to repeatability 1. Specifically, errors larger than 10.0 mm occurred in 5 out of 87, 9 out of 87, and 37 out of 87 landmarks in the x-, y-, and z-axes, respectively. Among the 37 landmarks exhibiting an error above 10.0 mm on the z-axis, 28 were located on the bone shafts. Errors for all landmarks ranged from 0.0 to 36.5 mm for repeatability 1 and from 0.0 to 42.1 mm for repeatability 2.

Considering the reproducibility analysis, errors ranged from 0.0 to 43.5 mm, with a mean median of 2.3 mm for all landmarks. Comparing results between operators, reproducibility errors above 10.0 mm were slightly higher than repeatability errors, with 14 out of 87, 18 out of 87, and 30 out of 87 landmarks in the x-, y- and z-axes.

Statistical tests (Student's *t*-test for normal and homoscedastic landmarks, Wilcoxon test for non-normal or non-homoscedastic landmarks) revealed that 25.6 % of repeatability 1 errors, 17.8 % of repeatability 2 errors, and 47.6 % of reproducibility errors were significant. Operator 2 had fewer significant errors but larger measurement errors overall compared to operator 1.

3.1.2. Representative example: Manual landmarking errors

To illustrate in greater detail the process and results above, a representative example of the studies conducted is presented. Specifically, the findings focus on a subset of clinically relevant landmarks used in the calculation of surgical metrics, such as frontal alignment. These include landmarks of interest: femoral head, distal femoral condyles (medial and lateral) and epicondyles, tibial spines, and medial and lateral malleolus. The analysis highlights the reproducibility between operators for these key landmarks. Fig. 3 shows the error distribution for each of these selected landmarks along all coordinate axes.

Additionally, the Quantile-Quantile (Q-Q) plots in Fig. 4 illustrate the differences between normal and non-normal error distributions for two landmarks, the femoral head and tibial medial spine.

Table 3 shows the specific landmark results for reproducibility. None of the normal and homoscedastic landmarks of interest errors were significant except the tibial lateral plate spine in the z-axis. Considering non-normal or non-homoscedastic patellar landmarks, the errors for the tibial lateral plate spine in the x-axis, for the femoral medial and lateral condyle and tibial lateral malleolus in the y-axis, and for the femoral medial epicondyle and the tibial medial plate spine in the z-axis were significant, as well as the lateral malleolus on all axes.

3.2. Evaluation of automatic landmarking

Similarly to the evaluation of the manual landmarking, our findings will be presented in two subsections for clarity: the global results and a detailed analysis of some landmarks of interest as an example. Detailed results for all landmarks are available in [Supplementary materials C, D and F](#).

3.2.1. Global overview of automatic landmarking errors

To evaluate the performance of automatic landmarking, comparisons were made between ground truth measurements and the results of automatic landmarking. Similarly to the repeatability and reproducibility analyses, since not all distributions of errors were normal and homoscedastic, results are expressed in terms of the median or the mean of the error for each landmark across all patients. To simplify the presentation of the results, the means of these medians were computed. The detailed results for each landmark are included in [supplementary material C](#). Table 3 summarizes the errors between manual and automatic

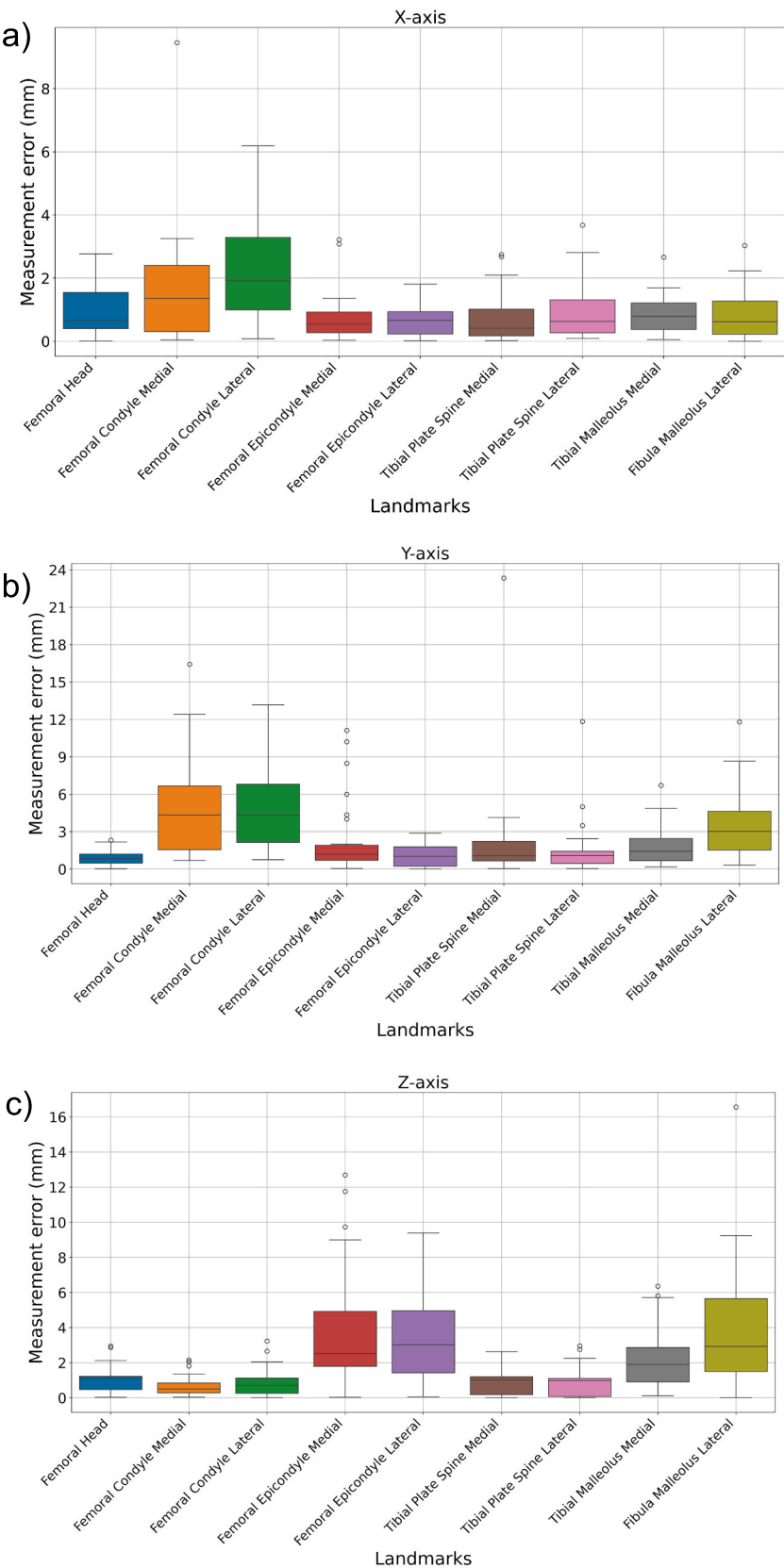


Fig. 3. Error distributions for each landmark of interest in the a) x-axis, b) y-axis and c) z-axis.

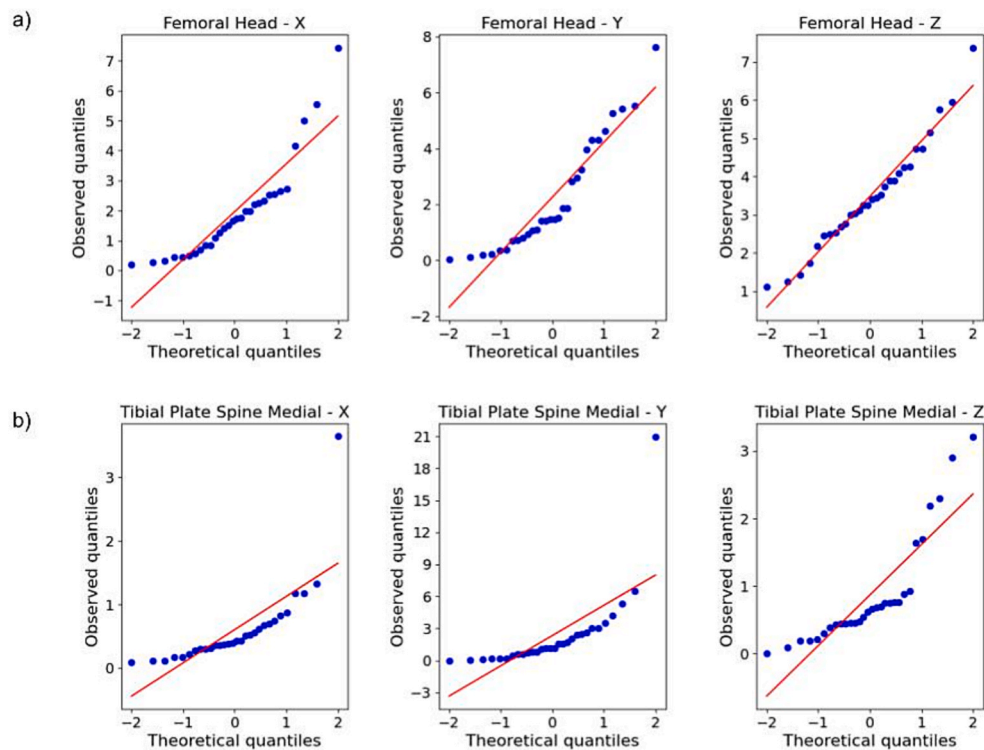


Fig. 4. Q-Q plot of the difference between two series of measurements by operator 1 and 2 for a) the femoral head showing a normal distribution in the Y-axis and for b) the tibia medial spine showing a non-normal distribution for all axis.

Table 3

Detailed landmarks of interest test results for reproducibility analysis. P-values with an (*) indicate significant inter-operator differences. SD: standard deviation. CI: confidence interval.

	Coordinate	Normal (Shapiro-Wilk)	Homoscedastic (Levene)	Mean error +/- SD error (mm)	Median (mm)	CI 95 % (mm)	P-value (Student)	P-value (Wilcoxon)
Femoral Head	X	x	x	–	0.7	[0.7; 1.2]	–	0.2
	Y	✓	✓	0.9 +/- 0.6	–	[0.7; 1.1]	0.4	–
	Z	x	x	–	1.1	[0.8; 1.3]	–	0.9
Femoral Condyle Medial	X	x	x	–	1.4	[1.0; 2.3]	–	0.1
	Y	x	x	–	4.3	[3.4; 6.1]	–	<0.05*
	Z	x	x	–	0.5	[0.5; 0.9]	–	0.1
Femoral Condyle Lateral	X	✓	✓	2.2 +/- 1.5	–	[1.6; 2.7]	0.4	–
	Y	x	x	–	4.3	[3.7; 6.1]	–	<0.05*
	Z	x	x	–	0.7	[0.6; 1.1]	–	0.1
Femoral Epicondyle Medial	X	x	x	–	0.5	[0.5; 1.0]	–	0.4
	Y	x	x	–	1.2	[1.3; 3.3]	–	0.8
	Z	x	x	–	2.5	[2.9; 5.2]	–	<0.05*
Femoral Epicondyle Lateral	X	✓	✓	0.7 +/- 0.5	–	[0.5; 0.9]	0.5	–
	Y	✓	✓	1.1 +/- 0.8	–	[0.8; 1.3]	0.2	–
	Z	x	x	–	3.0	[2.6; 4.4]	–	0.8
Tibial Plate Spine Medial	X	x	x	–	0.4	[0.4; 1.0]	–	0.1
	Y	x	x	–	1.1	[0.7; 3.7]	–	0.8
	Z	x	x	–	1.0	[0.7; 1.2]	–	<0.05*
Tibial Plate Spine Lateral	X	x	x	–	0.6	[0.6; 1.2]	–	<0.05*
	Y	x	x	–	1.1	[0.7; 2.3]	–	0.7
	Z	x	x	–	1.0	[0.6; 1.2]	<0.05*	0.1
Tibial Medial Malleolus	X	x	x	–	0.8	[0.6; 1.0]	–	0.2
	Y	x	x	–	1.4	[1.3; 2.4]	–	<0.05*
	Z	x	x	–	1.9	[1.6; 2.9]	–	0.1
Fibula Lateral Malleolus	X	x	x	–	0.6	[0.6; 1.1]	–	<0.05*
	Y	x	x	–	3.0	[2.5; 4.3]	–	<0.05*
	Z	x	x	–	2.9	[2.7; 5.0]	–	<0.05*

landmarking. The analysis method and the presented results mirror those used for manual landmarking: differences between automatic measures and the landmarking reference (mean of measures 1 and 3 of operators 1 and 2) were calculated for each landmark across all thirty patients. The median of these differences was then computed for each

landmark, followed by averaging these medians across all landmarks of a given bone. For instance, the 1.8 mm value reported in Table 4 for the femoral bone in the medio-lateral axis represents the mean of median errors across all 41 femoral landmarks. For patella landmarks, the mean of median errors was less than 2.0 mm, while femoral and tibial

Table 4

Summary of mean errors and statistical measures between manual and automatic landmarking.

	Mean of the median errors (mm)	95 % CI (mm)	Maximum Error Location (mm)	Medians < 4 mm (%)
Femur				
X	1.8	[1.5; 2.6]	Femoral Condyle Metaphysis Post Med (11.1 mm)	100.0
Y	1.7	[1.5; 2.6]	Femoral Great Trochanter Tip (17.0 mm)	97.6
Z	4.0	[3.3; 5.2]	Femoral Sup Lat Shaft (19.8 mm)	51.2
Tibia				
X	2.1	[1.8; 2.9]	Tibial Distal Ant (11.6 mm)	89.7
Y	2.5	[2.2; 3.4]	Tibial Plate Spine Medial (21.0 mm)	74.4
Z	3.3	[2.8; 4.8]	Tibial Sup Sup Ant Shaft (22.8 mm)	56.4
Patella				
X	1.0	[0.8; 1.5]	Patellar Apex Cartilage Inf (5.1 mm)	100.0
Y	0.8	[0.7; 1.2]	Patellar Apex Tendon Sup (6.5 mm)	100.0
Z	1.2	[1.0; 1.8]	Patellar Apex Cartilage Inf (6.8 mm)	100.0

landmarks had mean errors of less than 3.0 mm on the x- and y-axes. Femoral and tibial landmarks on the z-axis exhibited the largest errors, with mean errors exceeding 3.0 mm but without median errors exceeding 10.0 mm. These results are consistent with the repeatability and reproducibility analyses.

Articular landmarks (e.g., femoral condyles and patellar landmarks) show high precision with low median errors. For example, Patellar Apex Inf (X = 1.1 mm, Y = 0.6 mm, Z = 0.4 mm) achieves near-perfect accuracy with 100 % of measurements under 4 mm, highlighting the reliability of automatic methods for articular landmarks. Shaft landmarks, particularly for the femur and tibia, exhibit higher errors. This is particularly evident in regions like the femoral shaft mid med (Z = 8.59 mm) and the tibial shaft sup sup ant (Z = 8.17 mm), which are less anatomically distinct and harder to pinpoint. The tibia presents greater overall variability compared to the femur, especially in proximal and distal tibial regions (e.g., Tibial Dist Ant, X = 6.99 mm). The smaller size and less defined contours of the tibia pose a challenge for both manual and automatic landmarking techniques, resulting in larger errors.

Table 5 presents a summary of the distances on each axis separating the automatic landmark from the bounding box. These distances are measured when the automatic landmark is outside the bounding box. On average, the distance between the automatic landmark and the bounding box was 1.4 mm for femoral landmarks, with a maximum of 8.1 mm for Femoral Sup Med Shaft in the z-axis, 1.5 mm for tibial landmarks, with a maximum of 6.9 mm for Tibial Sup Sup Ant Shaft in the z-axis and 0.7 mm for patellar landmarks. Mean distances are below 1.5 mm for each bone and axis, except for the z-axis in femoral and tibial landmarks.

Across all landmarks, the proportion within the box was 77.2 % on the x-axis, 70.3 % on the y-axis, and 77.1 % on the z-axis. On average, the proportion of landmarks within the three-dimensional bounding box

Table 5

Summary of mean distances between the automatic landmark and the bounding box per direction and overall.

	X (mm)	Y (mm)	Z (mm)	Mean distance by bone landmarks (mm)
Femoral	1.2	1.2	2.2	1.4
Tibia	1.2	1.3	2.0	1.5
Patella	0.5	0.4	1.3	0.7

was 45.3 % for femoral landmarks (with 54.2 % for femoral shaft landmarks), 39.4 % for tibial landmarks (with 47.2 % for tibial shaft landmarks), and 42.4 % for patellar landmarks. Shaft landmarks were within the bounding box for 50.7 % of cases compared to 42.4 % for overall landmarks. This indicates that although automatic landmarks positioned on the shafts exhibited the greatest measurement errors, they were proportionally more often within the range of manual landmarking. This can be explained by the fact that the size of the bounding box increases when manual landmarking is less precise. Consequently, while automatic landmarking showed more dispersion along the z-axis compared to manual landmarking, the larger bounding box for shaft landmarks increased the likelihood of correct positioning of automatic landmarks. Fig. 5 illustrates two examples of automatic landmarking: one within the bounding box and one outside it.

Finally, parametric testing (Student's *t*-test) showed significant differences ($p < 0.05$) in 81.3 % of landmarks, and non-parametric testing (Wilcoxon) revealed significant differences in 59.4 % of landmarks.

3.2.2. Representative example: Automatic landmarking errors

To illustrate the detailed process and results, findings for clinically relevant landmarks are presented in this section. Fig. 6 shows the error distribution for each landmark for all coordinate axes.

Confidence intervals for repeatability and reproducibility analyses as well as manual versus automatic landmarking for each coordinate are provided in Fig. 7. These plots allow for the qualitative comparison of the performance between automatic and manual landmarking.

Table 6 presents a summary of the distances in millimetres separating the automatic landmarks of interest from the bounding box on each axis. Further detailed results across all landmarks are available in the [supplementary materials D](#). The last column of the table presents the proportion of landmarks found within the bounding box for all patients,

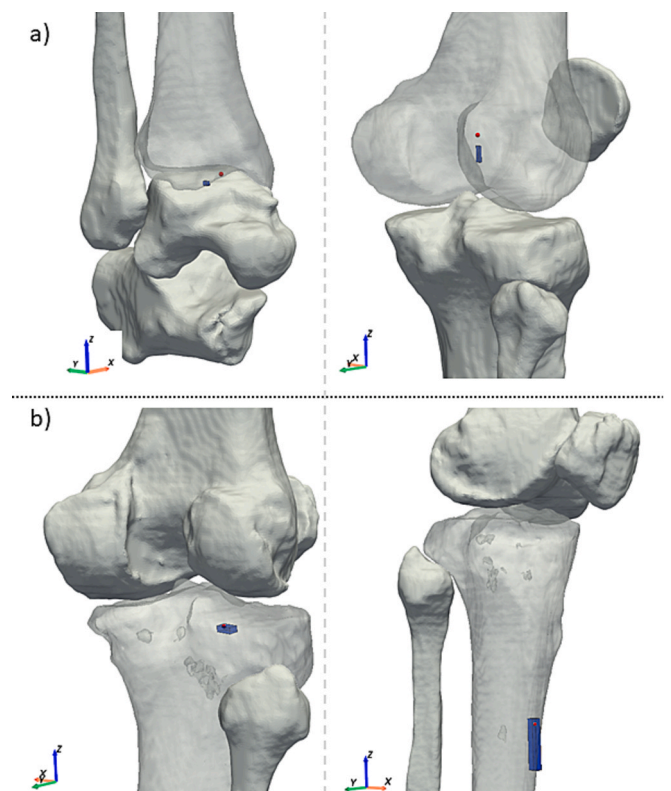


Fig. 5. Positions of automatic landmarks (red dot) compared to the manual landmarking bounding boxes (in blue) associated to each landmark of one patient (patient 1). On the top row (a), two landmarks out of the bounding boxes are shown. On the bottom row (b), two landmarks inside the bounding boxes are shown.

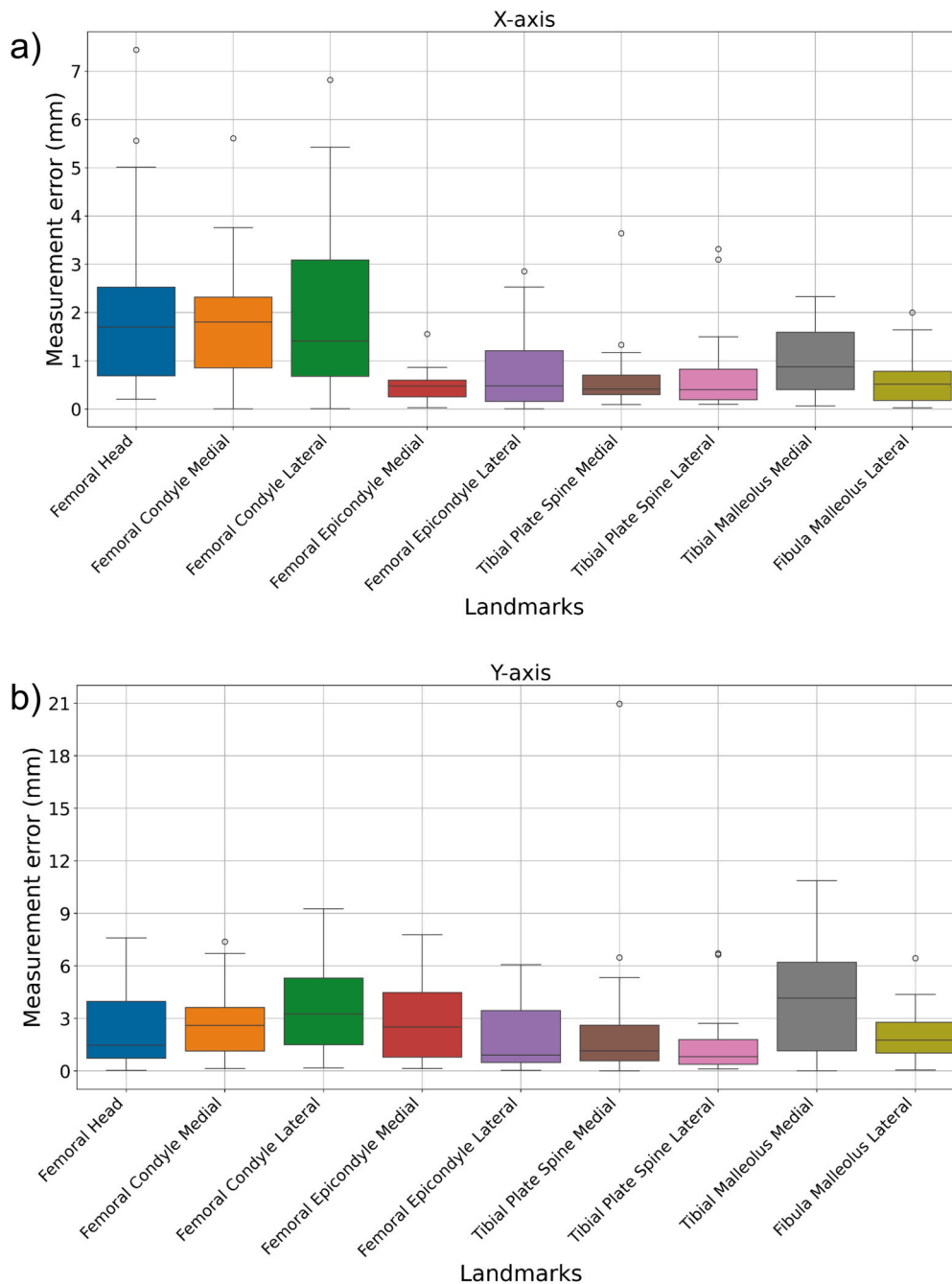


Fig. 6. Automatic versus manual landmarking error distributions for each landmark of interest in the a) x-axis, b) y-axis and c) z-axis.

ranging from 6.7 % to 50.0 %.

Table 7 shows the landmarks of interest results for automatic landmarking. Normal and homoscedastic landmark errors were considered significant except for the Femoral Epicondyle Lateral in the z-axis (85.7 % of significant error for normal and homoscedastic landmarks). Considering non-normal or non-homoscedastic landmarks, 25.0 % of errors were considered significant.

4. Discussion

This study aimed to propose an evaluation procedure for manual and automatic landmarking methods applied to osteoarthritic knee bones. Indeed, a standardized evaluation protocol seems crucial in the light of the limitations of previous studies. Nguyen et al[24]. evaluated two AI-

based techniques for 3D and 2D landmarking applied to a piglet hindlimb. The dataset used for validation was small (one piglet limb), the quality of the ground-truth data used for training and validation was not discussed, and the results were not compared to an acceptable error threshold. Phan et al[23]. evaluated their anatomical femoral landmark prediction method in a similar way on 35 patients. Subburaj et al[18]. presented an automated method to identify landmarks on 3D bone models. They compared automatic landmarks with manual landmarks obtained from palpation by three experienced clinicians. However, the process for obtaining manual landmarks was not detailed. They assessed the intra- and inter-operator variability in manual landmarking for all landmarks along the three axes independently. The average of each set of manual landmarks was used to define the gold standard. The authors provided the mean absolute error between automatic and manual

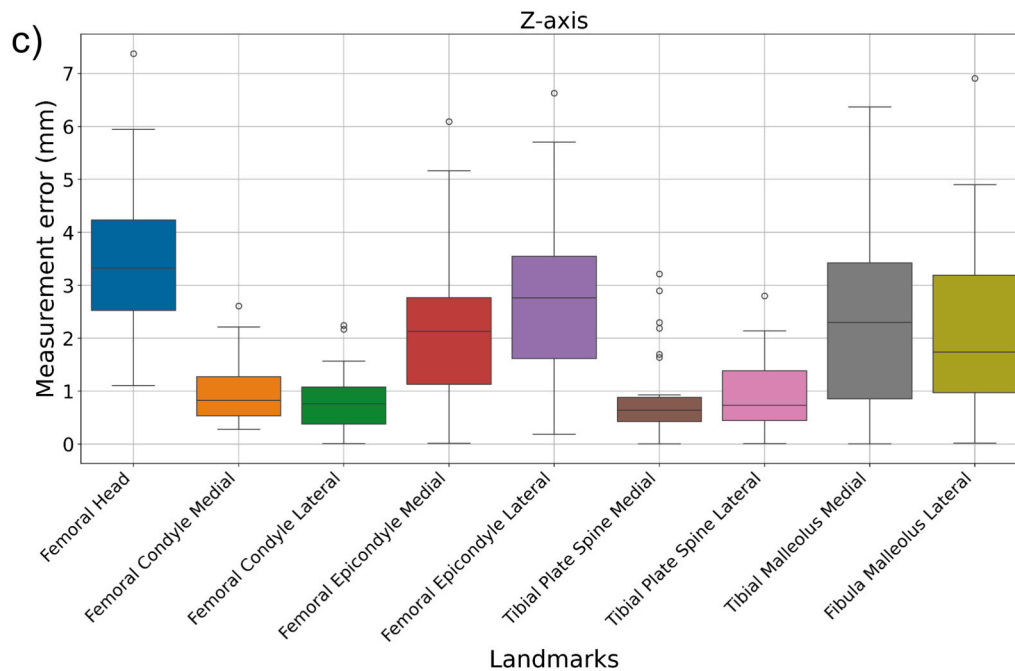


Fig. 6. (continued).

landmarking. Results showed errors of the same order of magnitude as the manual inter-operator variability. However, the method's efficiency evaluation across different patients was limited, as only one patient was used in the dataset. Consequently, they could not perform further statistical analyses of the results. Victor et al.[34] evaluated the accuracy of knee manual landmarking from CT scans. Three operators performed landmarking on images of six specimens. The study reported the intra- and inter-operator variability and the Intra-class Correlation Coefficient (ICC). The intra- and inter-operator variability were computed using the distance between the mean and the observed positions of a landmark. The ICC was computed as the ratio of the total variance of different observers' landmarking and the variance between patients. Furthermore, Fürmetz et al.[25] evaluated a protocol analysing 3D bone models of the lower limbs of three specimens. Three operators performed the landmarking, and the same evaluation as Victor et al.[34] was used. In both studies, the authors reported the mean and standard deviation of the errors but did not evaluate whether the error distributions followed a normal distribution, which is necessary to compute the ICC[35]. Furthermore, they did not assess the significance of the errors. Bori et al.[36] evaluated measurements based on automatic landmarking of amputated legs. Three experienced operators performed manual landmarking, and the researchers computed inter-operator variability and the ICC. They evaluated the agreement between the automatic and manual results using the Pearson correlation coefficient and the Bland-Altman graphs[37]. The authors evaluated the normality of the distribution using the Shapiro-Wilk test but did not report whether the error distributions were homoscedastic. In addition, they computed the errors for measurements rather than landmark positions, which may be less sensitive to landmark mispositioning. The small dataset size also limited the power of the tests. Baek et al.[38] included 43 human femurs in their study, which were all landmarked by two expert clinicians using their anatomical deformation technique. They evaluated the accuracy of their algorithm using the leave-one-out cross-validation method, which is suitable for AI-based methods but not for knowledge-based techniques. Jacinto et al.[22] complemented the evaluation of their automatic method for landmark positioning by comparing the results with the ground truth based on the manual landmarking performed by three trained operators. They also assessed the inter-operator variability of the landmarking on three knees. They compared different methods using the

one-sided Welch *t*-test to quantify the test of equal means. Although this method was applied to 17 patients, the authors did not discuss the impact of the sample size. Furthermore, they did not explain the choice of the statistical test. The Welch *t*-test is designed for unequal population variances but still requires the assumption of distribution normality to be met[39].

In our study, the automatic landmarking results were compared with manual landmarking results performed by two trained clinician operators in order to evaluate and validate their accuracy. The procedure was applied on a database of thirty patients with osteoarthritic knees. The sample size was chosen according to the classically admitted threshold used in statistical analysis to apply the student *t*-test. Yet, the distribution normality hypothesis was not true for many analyses. It is important to note that in most of the literature studies, this threshold was not respected[22,25,34]. Rather than defining a gold standard for 3D automatic landmarking, this study emphasises the need for a rigorous evaluation of such techniques, including verifying the quality of the chosen ground truth and comparing results with clinical requirements. Table 8 presents a comparative analysis of reproducibility errors across different studies[18,22,40,41].

The manual landmarking method revealed intra-operator landmark errors ranging from 0.0 to 42.1 mm, with a mean median error of 2.1 mm for all landmarks in both repeatability analyses. Specifically, the mean median error was 2.1 mm for femoral landmarks, 2.6 mm for tibial landmarks, and 1.1 mm for patellar landmarks. The likelihood of errors increases in the presence of substantial morphological variability between subjects, as was the case in this cohort, in which many patients exhibited degenerative morphological changes (bone wear and osteophytes). The intra-operator variability was compared with the results of Della Croce et al.[40]. They conducted landmarking with one operator who marked six times on two patients (8 femoral and 6 tibial landmarks). In contrast, our study involved two operators marking twice on fifteen patients (41 femoral and 39 tibial landmarks). Our study showed lower error rates, with a mean of median errors of 2.1 mm for femoral landmarks compared to Della Croce et al.'s 5.6 mm, and 2.6 mm for tibial landmarks compared to 3.8 mm. When focusing on landmarks used for axis measurements on the femur or tibia, Della Croce et al.'s errors were 9.8 mm for the medial epicondyle vs 1.8 mm in our study, 10.0 mm vs 2.3 mm for the lateral epicondyle, 8.6 mm vs 1.5 mm for the

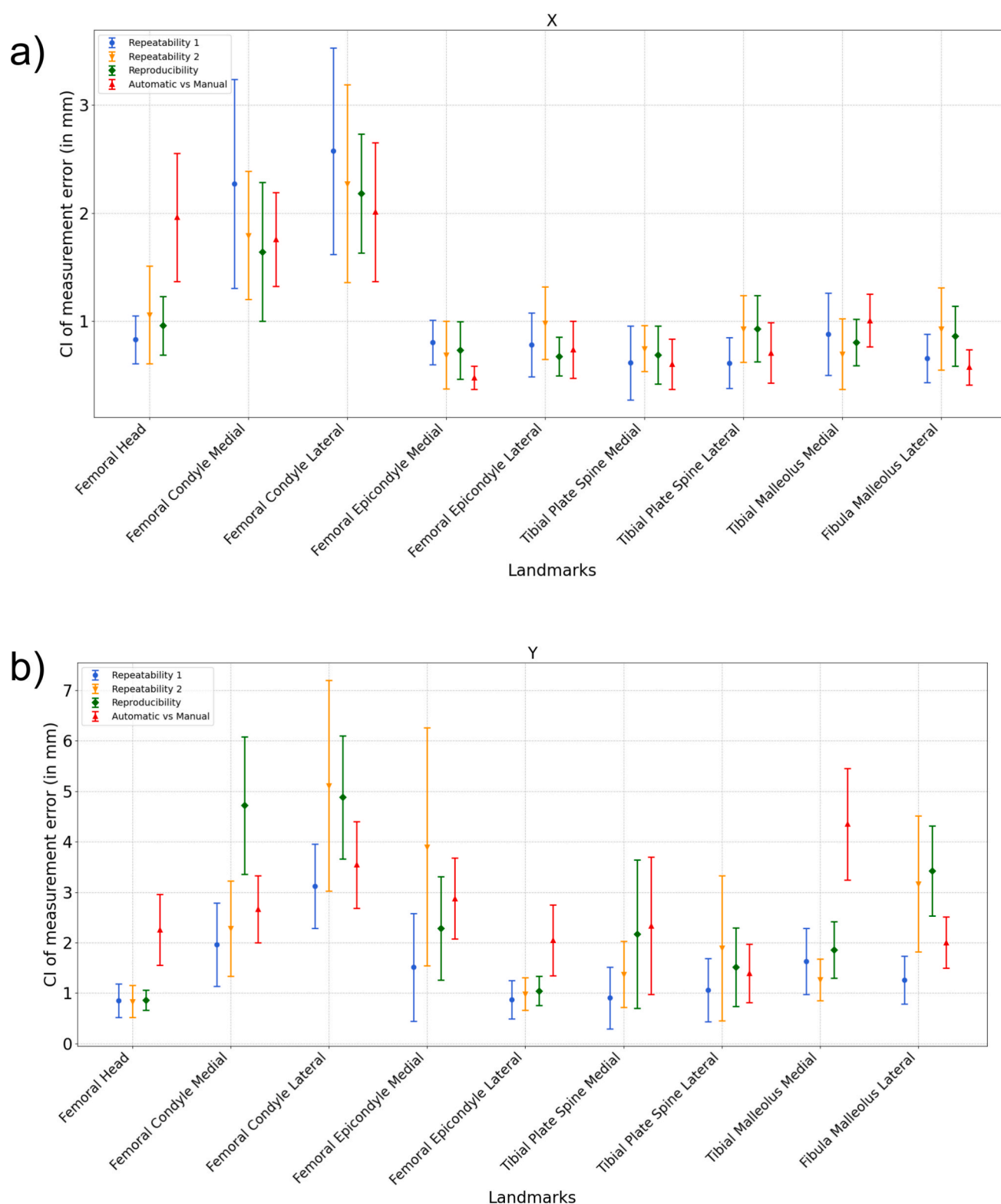


Fig. 7. Confidence intervals of manual analyses (repeatability 1 and 2 and reproducibility) and automatic evaluation for landmarks of interest in the a) x-axis, b) y-axis and c) z-axis.

medial part of the tibial plateau and 5.2 mm vs 2.2 mm for the lateral tibial plateau. However, this comparison should be interpreted with caution, as the higher number of landmarks per bone in our study may reduce overall error estimates by averaging, potentially introducing a favourable bias. This underscores the importance of considering not

only global variability metrics but also landmark-specific uncertainties. **The findings highlight the need to examine not only the overall error, but also intrinsic uncertainty about specific landmarks that may have a potential impact on clinical measurements.** For example, when assessing mechanical axes, the choice of method can

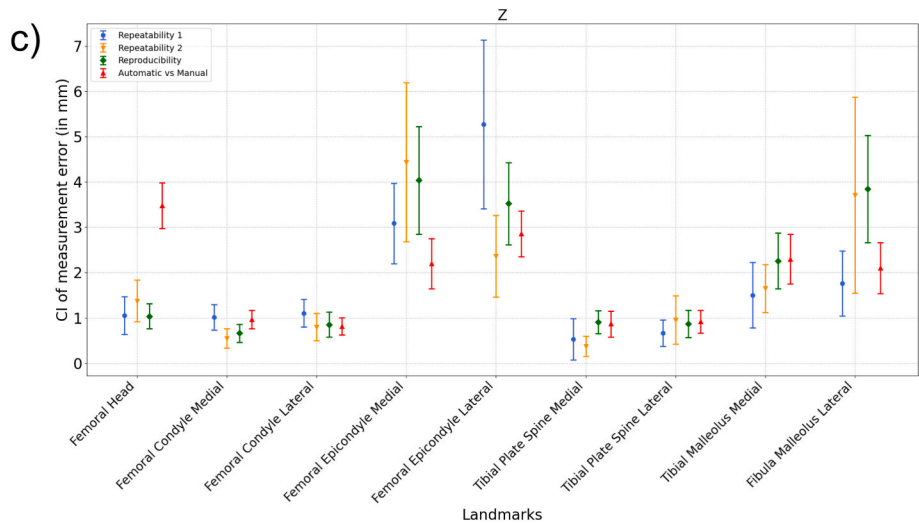


Fig. 7. (continued).

Table 6
Mean distances (in mm) between the automatic landmark and the bounding box in the x, y and z-axis and the proportion of the landmark in the bounding box.

Landmarks	X (mm)	Y (mm)	Z (mm)	Proportion (%)
Femoral Head	0.5	0.4	0.0	46.7
Femoral Condyle Medial	1.4	0.3	0.0	6.7
Femoral Condyle Lateral	0.8	0.6	1.5	13.3
Femoral Epicondyle Medial	1.0	0.7	0.8	36.7
Femoral Epicondyle Lateral	0.6	0.7	5.7	46.7
Tibial Plate Spine Medial	1.5	0.6	1.7	43.3
Tibial Plate Spine Lateral	0.0	2.2	4.3	40.0
Tibial Medial Malleolus	0.6	0.3	1.0	33.3
Fibula Lateral Malleolus	2.5	1.1	6.9	50.0

significantly influence the resulting angles[42]. Moreover, anatomical regions prone to deformation—such as the tibial spines or condylar landmarks—can introduce even greater variability.

Our study revealed inter-operator variability ranging from 0.0 to 43.5 mm, with mean median errors slightly higher than in the repeatability analysis (2.3 mm vs 2.1 mm). Our method generally yields lower mean results than other studies, except in the case of tibial landmarks compared to both Jacinto et al. studies[22,41]. Considering the lateral part of the tibial plateau, for example, we obtained an average error of 4.1 mm, compared with 3.4 mm for Jacinto et al.[22] and 4.9 mm for Subburaj et al[18]. These discrepancies may be attributed to the trained operators' higher precision with fewer landmarks to place on the tibia, leading to fewer errors in their study. Jacinto et al.[22] used five landmarks, while our study used thirty-nine tibial landmarks, potentially accounting for the observed discrepancies. In comparison, Subburaj et al.[18] measured variations in manual positioning using three knees, each marked by three different orthopaedic surgeons, while Della Croce et al.[40] employed six registered physical therapists with gait laboratory experience. Expertise and number of operators might have an impact on the reproducibility results. Given that both operators in our study have the same level of expertise, the observed higher reproducibility error values are likely attributable to differences in their technique and interpretation of landmarks. Even with equivalent training and experience, small variations in landmark approaches, such as differences in positioning, perspective or judgement, can introduce inconsistencies. These variations can lead to discrepancies in the placement of landmarks, ultimately reducing inter-operator consistency. Furthermore, it is important to note that manual landmarking is a time-consuming and tedious operation. Despite the valuable information provided by landmarking, implementing manual landmarking in clinical

practice appears challenging. Therefore, the higher reproducibility errors do not necessarily indicate a lack of skill, but rather highlight the inherent challenges and subjective nature of the landmarking process. **This also underscores the need to take into account results for landmarks of interest for applications such as angle measurement, rather than global results, which might appear superior but exhibit unsatisfactory results for measurements of clinical interest.**

Consistency across different conditions is another critical factor. Operators might have used the image slices for landmarking while others might prefer using the 3D reconstruction, which can affect the precision of landmark localisation. Our study showed that accurately positioning landmarks on the femoral and tibial shafts is particularly difficult due to their unique anatomical characteristics. Indeed, these structures are cylindrical and elongated, lacking distinct and easily identifiable anatomical features that can serve as reliable reference points. This lack of prominent landmarks makes it challenging to achieve consistent and precise placement. **Additionally, the shafts are often subject to various anatomical variations and irregularities, such as slight bends or twists, which can further complicate the landmarking process.** The results showed higher error rates in these regions, especially on the proximo-distal axis, where mean of median errors was twice as high as other coordinate axes.

We observed that the accuracy of landmarking differs notably between simpler and more complex anatomical regions. For example, patellar landmarks, which are located in relatively straightforward anatomical areas, exhibit mean median errors of less than 2.0 mm, demonstrating high precision and accuracy in both manual and automatic methods. This suggests that landmarks in regions with more distinct anatomical features are easier to place accurately. Therefore, the observed differences in accuracy demonstrate that regions with more distinct anatomical features, such as ridges and contours, facilitate more precise landmarking, while the smoother and less defined surfaces, such as shaft regions, introduce significant challenges and increase the likelihood of errors. However, regions with minimal relief—namely the diaphysis—are primarily used to define shaft centres via 3D barycentric calculations for establishing vertical axes; as a result, any landmarking errors in these areas have only a limited impact on barycentre position and on measurements in the frontal and sagittal planes.

Our results also emphasised the importance of standardised measurement techniques to reduce inter-operator variability, as significant discrepancies were found in manual landmarking. Shaft landmarks showed higher inter-operator dispersion relative to intra-operator dispersion, suggesting that their definition is particularly prone to misinterpretation. Advanced imaging techniques, such as 3D imaging

Table 7

Detailed landmarks of interest test results for automatic vs manual landmarking. P-values with an (*) indicate significant differences. SD: standard deviation; CI: confidence interval.

	Coordinate	Normal (Shapiro-Wilk)	Homoscedastic (Levene)	Mean error +/- SD error (mm)	Median (mm)	CI 95 % (mm)	P-value (Student)	P-value (Wilcoxon)	Pearson Correlation	Spearman Correlation
Femoral Head	X	x	x	–	1.7	[1.4; 2.6]	–	0.07	–	0.8
	Y	x	x	–	1.5	[1.6; 3.0]	–	0.35	–	0.9
	Z	✓	✓	3.5 +/- 1.4	–	[3.0; 4.0]	<0.05*	–	0.9	–
Femoral Condyle Medial	X	✓	x	–	1.8	[1.3; 2.2]	–	0.13	–	0.5
	Y	✓	✓	2.7 +/- 1.9	–	[2.0; 3.3]	<0.05*	–	–0.1	–
	Z	x	x	–	0.8	[0.8; 1.2]	–	<0.05*	–	1.0
Femoral Condyle Lateral	X	x	x	–	1.4	[1.4; 2.7]	–	0.97	–	0.2
	Y	✓	x	–	3.3	[2.7; 4.4]	–	0.08	–	–0.4
	Z	x	x	–	0.8	[0.6; 1.0]	–	<0.05*	–	1.0
Femoral Epicondyle Medial	X	x	x	–	0.5	[0.4; 0.6]	–	0.5	–	1.0
	Y	x	x	–	2.5	[2.1; 3.7]	–	<0.05*	–	–0.1
	Z	✓	x	–	2.1	[1.7; 2.8]	–	0.8	–	0.6
Femoral Epicondyle Lateral	X	x	x	–	0.5	[0.5; 1.0]	–	<0.05*	–	1.0
	Y	x	x	–	0.9	[1.3; 2.8]	–	0.02	–	–0.2
	Z	✓	✓	2.9 +/- 1.4	–	[2.4; 3.4]	0.2	–	0.4	–
Tibial Plate Spine Medial	X	x	x	–	0.4	[0.4; 0.8]	–	0.5	–	0.9
	Y	x	x	–	1.2	[1.0; 3.7]	–	0.8	–	0.6
	Z	x	x	–	0.6	[0.6; 1.2]	–	0.1	–	1.0
Tibial Plate Spine Lateral	X	x	x	–	0.4	[0.4; 1.0]	–	0.9	–	0.9
	Y	x	x	–	0.8	[0.8; 2.0]	–	0.1	–	0.7
	Z	✓	✓	0.9 +/- 0.7	–	[0.7; 1.2]	<0.05*	–	1.0	–
Tibial Medial Malleolus	X	✓	✓	1.0 +/- 0.7	–	[0.8; 1.3]	<0.05*	–	0.9	–
	Y	✓	✓	4.4 +/- 3.1	–	[3.3; 5.5]	<0.05*	–	0.2	–
	Z	✓	✓	2.3 +/- 1.5	–	[1.8; 2.9]	<0.05*	–	0.9	–
Fibula Lateral Malleolus	X	x	x	–	0.5	[0.4; 0.7]	–	<0.05*	–	1.0
	Y	x	x	–	1.8	[1.5; 2.5]	–	0.1	–	0.8
	Z	x	x	–	1.7	[1.5; 2.7]	–	0.8	–	0.9

and high-resolution scans, can help mitigate these limitations by providing clearer and more detailed views of complex anatomical regions. Incorporating multi-modal imaging data could enhance the robustness of landmarking algorithms, providing additional context and reducing ambiguity in landmark identification.

This study then evaluated an automatic landmarking method, which produced landmark errors ranging from 0.0 to 22.8 mm, with a mean median deviation of 2.4 mm. Kuiper et al.[43], who evaluated 32 landmarks, reported a mean difference of 2.17 ± 1.37 mm between fully automatic and manual methods. Their reported errors ranged from 0.36 ± 0.19 mm for the femoral head centre to 3.63 ± 2.21 mm for the femoral greater trochanter. Notably, they did not quantify errors on femoral and tibial shaft landmarks, as these were used solely for angle computation. In contrast, our evaluation included shaft landmarks,

which exhibited some of the highest variability among all anatomical points. Including these landmarks in our evaluation increases the overall error metrics and reflects a more comprehensive, but also more challenging, assessment of automatic landmarking performance. Among all landmarks, the femoral shaft landmarks exhibited the largest dispersion compared to tibial landmarks. Jacinto's study[22] indicated that both manual and automatic landmarking were more precise for the tibia than for the femur, a trend observed only in the automatic landmarking in our study. This discrepancy could be due to the same amount of shaft landmarks on the tibial and femoral bones but fewer tibial landmarks than femoral ones in our study, as well as the use of plane sections by the operators. Table 8 presents a comparative analysis of automatic landmarking errors across different studies. Our proposed method demonstrated lower error results than other studies and is comparable to

Table 8

Comparisons of our global and detailed tibial and femoral bone automatic landmarking results (in mm) with published results. (*) represent results for the computation of mean of means location, (**) for mean of means distances and (***) represent results for the computation of means of medians.

	Analysis	Subburaj et al.[18] *	Jacinto et al.[41] **	Della Croce et al. [40] **	Our study ***
Global Femur	Intra	NA	NA	5.6	2.1
	Inter	4.1	2.8	8.8	2.2
	Auto	3.4	5.1	NA	2.5
Global Tibia	Intra	NA	NA	3.8	2.2
	Inter	3.7	1.4	8.2	2.7
	Auto	3.3	3.7	NA	2.6
Medial Epicondyle	Intra	NA	NA	9.8	1.7
	Inter	4.8	3.3	14.9	1.4
	Auto	4.4	4.8	NA	1.7
Lateral Epicondyle	Intra	NA	NA	10.0	1.8
	Inter	3.9	1.4	19.2	1.6
	Auto	3.6	4.3	NA	1.4
Medial Distal Condyle	Intra	NA	NA	6.4	1.4
	Inter	3.9	2.8	14.1	2.1
	Auto	3.0	5.1	NA	1.7
Lateral Distal Condyle	Intra	NA	NA	6.5	2.1
	Inter	3.5	5.0	13.4	2.3
	Auto	2.3	6.6	NA	1.8
Medial tibial plate	Intra	NA	NA	8.6	1.1
	Inter	6.0	1.2	19.7	1.7
	Auto	4.9	3.8	NA	2.9
Lateral tibial plate	Intra	NA	NA	10.0	1.3
	Inter	5.4	1.3	13.8	2.0
	Auto	4.6	3.1	NA	4.1

manual landmarking, as the error magnitudes are similar in both repeatability and reproducibility analyses. Focusing on landmarks used for axis measurements on the femur or tibia, Jacinto et al. 's error for the medial distal epicondyle was 4.8 mm, while Subburaj et al.'s was 4.4 mm compared to 1.7 mm in our study. However, Jacinto et al. 's error was 4.3 mm and Subburaj et al.'s was 3.6 mm compared to the 1.4 mm that we measured for the lateral epicondyle. For the medial and lateral parts of the tibial plateau, Jacinto et al. 's error is 3.8 mm and 3.1 mm and Subburaj et al.'s is 4.9 mm and 4.6 mm compared to 2.9 mm and 4.1 mm here.

It is important to note that landmark positioning in each study was performed on different structures and with different methods. Subburaj et al.[18] localized landmark regions based on a classification system (peak, ridge, pit, ravine) and utilized *a priori* spatial information, while Jacinto's method generated spatial coordinates of anatomical points within the knee frame[22]. The landmarking of a representative point within an anatomical region is inherently less precise than that of a point located on a well-defined geometric feature due to the absence of distinct boundaries or reliefs to guide the identification. The differences highlight the diversity in landmarking strategies and their respective advantages and limitations. The irregular shapes and proximity to other anatomical features in these regions make accurate landmarking particularly difficult. The presence of large osteophytes in some patients and the degenerative deformation of the knee likely contributed to misdetections, affecting the accuracy of both manual and automatic landmarking. This complexity can lead to higher error rates and reduced measurement precision. Notably, in Jacinto's case (arthroplasty), the bones are represented only in the region of the knee (distal femur and proximal tibia) while in our case, a full lower limb was used[22]. Implementing machine learning and deep learning models could significantly enhance the accuracy and consistency of these registrations [44]. These models, trained on large datasets, could recognize complex patterns that may not be apparent to human operators, thus reducing errors and increasing precision.

Clinical applications of a reliable lower limb automatic landmarking technique would be valuable for the use of image-based navigation or

robotic systems for TKA implantation. Indeed, most of these technologies currently require manual landmarking, which appears time-consuming in daily clinical practice. Quick and precise anatomical evaluation using 3D CT automatic landmarking appears as a crucial tool when discussing surgical indication for congenital lower limb deformities correction (tibial posterior slope reduction, coronal malalignment correction) or when planning a knee arthroplasty or osteotomy. In addition, the same applications can be used for pelvic, hip, foot and ankle surgery using models including landmarks of interest on the involved bones.

Our study has several limitations. Manual landmarking was performed using TwinSight's TwinForge software, which only permits placement of landmarks on orthogonal CT slices with concurrent 3D position visualization. Automatic landmarking, by contrast, relied on meshes generated via automatic segmentation from SurgiTwin, a SaaS platform currently under FDA submission. Consequently, we could not fine-tune CT slice orientation in the landmarking software, and the anatomical planes remained defined by the patient's lower-limb positioning during image acquisition. The supine position in the CT scan satisfactorily reproduces the femoral and tibial axial plane should the knee be in full extension. Since flexum is a frequent condition in osteoarthritic knees, this might have slightly compromised landmarking precision, especially for shaft landmarks on the z-axis, which were all positioned on the same axial slice that was not always strictly orthogonal to the tibial and femoral mechanical axes. The supine position during CT acquisition does not allow for control of lower limb orientation, often resulting in spontaneous external rotation. This alters the appearance of coronal and sagittal CT slices between patients. In addition, visual positioning of the shaft landmarks, consistently with their definition (with no tools to measure precise distances on the z-axis), explains the poor results reported. In terms of statistical methodology, our analysis offers the advantage of describing repeatability, reproducibility, and comparison with the automatic technique for each landmark. This is achieved at the cost of a high number of tests performed, thus increasing the alpha risk. An alternative approach to rule out this issue would be the use of mixed-effects models to present overall generalized results. Finally, for clinical relevance, the accuracy of landmark positioning should also be assessed using lower limb metrics of interest as primary endpoints.

Future research should prioritize the evaluation and improvement of landmarking techniques, particularly in addressing the topographic challenges posed by different anatomical regions with local deformation, such as osteophytes. While understanding anatomical variability and developing specialized algorithms are important, it is equally critical to assess how accurately landmarks are placed and how errors in this process can impact key surgical metrics such as angles and distances. By focusing on the accuracy and precision of landmarking, researchers can better ensure that surgical planning is not only well-informed but also capable of accommodating the inherent variability and complexity of individual anatomical structures, ultimately leading to more successful outcomes in knee surgery.

5. Conclusion

When evaluating the performance of an automatic landmarking method, it is essential to compare it to a ground truth based on repeatability and reproducibility. Moreover, it is important to examine not only global errors, but also specific errors linked to landmarks of clinical interest. It is crucial to verify the hypotheses for performing statistical tests such as normality and homoscedasticity and to evaluate the tests' power considering the population size. If these requirements are respected, this evaluation method becomes applicable to osteoarthritic bones with and without osteophytes.

The clinical implications of accurate landmarking are significant. Precise landmarking is crucial for patient outcomes, particularly in surgical planning, diagnostics, and other clinical applications. Improved

landmarking techniques can lead to better clinical decisions and, consequently, enhanced patient care.

Interdisciplinary collaboration between engineers who develop registration methods, clinicians who possess anatomical knowledge and measurement expertise, and researchers who analyse the impact on results, is essential for developing and refining registration and landmarking techniques. This collaboration helps overcome the practical challenges inherent to landmarking, and ensures that new methods are clinically relevant, effective and beneficial to patients.

CRedit authorship contribution statement

Clément Horteur: Writing – original draft, Investigation, Data curation, Validation, Formal analysis, Conceptualization. **Nolwenn Fougeron:** Writing – original draft, Software, Formal analysis, Conceptualization, Writing – review & editing, Validation, Methodology, Data curation. **Benoît Gaulin:** Writing – original draft, Data curation, Writing – review & editing, Formal analysis. **Ysé Roch:** Writing – review & editing, Software, Formal analysis, Conceptualization, Writing – original draft, Methodology, Data curation. **Marek Bucki:** Writing – review & editing, Software. **Estelle Palluel:** Writing – review & editing, Supervision. **Yohan Payan:** Supervision, Funding acquisition, Writing – review & editing, Project administration, Conceptualization. **Antoine Perrier:** Supervision, Methodology, Funding acquisition, Conceptualization, Writing – review & editing, Project administration, Investigation, Data curation.

Declaration of competing interest

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

Acknowledgement

This work was funded by Fondation pour la Recherche Médicale under the project FRM DIC20161236448. The IRMaGe MRI facility was funded by the ANR-11-INSB-0006 grant.

This work received the financial support of the *Programme Innovation Technologique* of the Région Auvergne-Rhône-Alpes.

This project was provided with computing HPC and storage resources by GENCI at IDRIS thanks to the grant 2023-A0160313879 on the supercomputer Jean Zay's A100 partition.

Dr. Joris Gai and the CIC-IT of Grenoble Alpes University Hospital provided their methodological support during the review process.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ejrad.2025.112292>.

References

- [1] D. Saragaglia, B. Rubens-Duval, J. Gaillot, G. Lateur, R. Pailhé, Total knee arthroplasties from the origin to navigation: history, rationale, indications, *Int. Orthop.* 43 (2019) 597–604, <https://doi.org/10.1007/s00264-018-3913-z>.
- [2] E.A. Treu, J.J. Frandsen, C.D. Woodley, S.K. Loughmiller, B.E. Blackburn, C. L. Peters, Accelerometer-based Navigation in Primary Total knee Arthroplasty leads to improved Alignment but No Change in Patient-Reported Outcomes, *J. Arthroplasty* 38 (2023) S222, <https://doi.org/10.1016/j.arth.2023.02.081>.
- [3] S. Chaouche, C. Jacquet, M. Fabre-Aubrespy, A. Sharma, J.-N. Argenson, S. Parratte, et al., Patient-specific cutting guides for open-wedge high tibial osteotomy: safety and accuracy analysis of a hundred patients continuous cohort, *Int. Orthop.* 43 (2019) 2757–2765, <https://doi.org/10.1007/s00264-019-04372-4>.
- [4] J. Victor, A. Premanathan, Virtual 3D planning and patient specific surgical guides for osteotomies around the knee: a feasibility and proof-of-concept study, *Bone Joint J* 95-B:153–8 (2013), <https://doi.org/10.1302/0301-620X.95B11.32950>.
- [5] C. Rivière, J.M. Vigdorchik, P.-A. Vendittoli, Mechanical alignment: the end of an era?, *Orthop. Traumatol. Surg. Res.* 105 (2019) 1223–1226, <https://doi.org/10.1016/j.otsr.2019.07.005>.
- [6] P. Sadoghi, P.-A. Vendittoli, S. Lustig, J. Leal, H. Graichen, C. Rivière, et al., Less religion and more science in the discussion of personalized alignment in total knee arthroplasty: we need to lead the transition process!, *Knee Surg. Sports Traumatol. Arthrosc.* 30 (2022) 2883–2885, <https://doi.org/10.1007/s00167-022-07079-z>.
- [7] J. Shatrov, C. Battelier, E. Sappey-Marinié, S. Gunst, E. Servien, S. Lustig, Functional Alignment philosophy in Total knee Arthroplasty - Rationale and technique for the varus morphotype using a CT based robotic platform and individualized planning, *SICOT J* 8 (2022) 11, <https://doi.org/10.1051/sicotj/2022010>.
- [8] J. Shatrov, C. Foissey, M. Kafelov, C. Batailler, S. Gunst, E. Servien, et al., Functional Alignment philosophy in Total knee Arthroplasty-Rationale and Technique for the Valgus Morphotype using an image based Robotic Platform and Individualized Planning, *J Pers Med* 13 (2023) 212, <https://doi.org/10.3390/jpm13020212>.
- [9] P. Winnock de Grave, J. Kellens, T. Luyckx, T. Tampere, F. Lacaze, K. Claeys, Inverse Kinematic Alignment for Total knee Arthroplasty, *Orthop. Traumatol. Surg. Res.* 108 (2022) 103305, <https://doi.org/10.1016/j.otsr.2022.103305>.
- [10] P.-A. Vendittoli, S. Martinov, W.G. Blakeney, Restricted Kinematic Alignment, the Fundamentals, and Clinical applications, *Front. Surg.* 8 (2021) 697020, <https://doi.org/10.3389/fsurg.2021.697020>.
- [11] S.M. Howell, M. Gill, T.J. Shelton, A.J. Nedopil, Reoperations are few and confined to the most valgus phenotypes 4 years after unrestricted calipered kinematically aligned TKA, *Knee Surg. Sports Traumatol. Arthrosc.* 30 (2022) 948–957, <https://doi.org/10.1007/s00167-021-06473-3>.
- [12] M.-D. Ahrend, H. Baumgartner, C. Ihle, T. Histing, S. Schröter, F. Finger, Influence of axial limb rotation on radiographic lower limb alignment: a systematic review, *Arch. Orthop. Trauma Surg.* 142 (2022) 3349–3366, <https://doi.org/10.1007/s00402-021-04163-w>.
- [13] L. Jud, T. Roth, P. Fürnstahl, L. Vlachopoulos, R. Sutter, S.F. Fucentese, The impact of limb loading and the measurement modality (2D versus 3D) on the measurement of the limb loading dependent lower extremity parameters, *BMC Musculoskelet. Disord.* 21 (2020) 418, <https://doi.org/10.1186/s12891-020-03449-1>.
- [14] N.A. Segal, J. Bergin, A. Kern, C. Findlay, D.D. Anderson, Test-retest reliability of tibiofemoral joint space width measurements made using a low-dose standing CT scanner, *Skeletal Radiol.* 46 (2017) 217–222, <https://doi.org/10.1007/s00256-016-2539-8>.
- [15] M. Tanzer, A.M. Makhdum, Preoperative Planning in Primary Total knee Arthroplasty, *J. Am. Acad. Orthop. Surg.* 24 (2016) 220–230, <https://doi.org/10.5435/JAAOS-D-14-00332>.
- [16] R.S. Hinman, R.L. May, K.M. Crossley, Is there an alternative to the full-leg radiograph for determining knee joint alignment in osteoarthritis? *Arthritis Rheum.* 55 (2006) 306–313, <https://doi.org/10.1002/art.21836>.
- [17] C.-W. Wang, C.-T. Huang, M.-C. Hsieh, C.-H. Li, S.-W. Chang, W.-C. Li, et al., Evaluation and Comparison of Anatomical Landmark Detection Methods for Cephalometric X-Ray Images: a Grand Challenge, *IEEE Trans. Med. Imaging* 34 (2015) 1890–1900, <https://doi.org/10.1109/TMI.2015.2412951>.
- [18] K. Subburaj, B. Ravi, M. Agarwal, Automated identification of anatomical landmarks on 3D bone models reconstructed from CT scan images, *Comput. Med. Imaging Graph.* 33 (2009) 359–368, <https://doi.org/10.1016/j.compmedimag.2009.03.001>.
- [19] D.L. Miranda, M.J. Rainbow, E.L. Leventhal, J.J. Crisco, B.C. Fleming, Automatic determination of anatomical coordinate systems for three-dimensional bone models of the isolated human knee, *J. Biomech.* 43 (2010) 1623–1626, <https://doi.org/10.1016/j.jbiomech.2010.01.036>.
- [20] J.-B. Renault, G. Aüllo-Rasser, M. Donnez, S. Parratte, P. Chabrand, Articular-surface-based automatic anatomical coordinate systems for the knee bones, *J. Biomech.* 80 (2018) 171–178, <https://doi.org/10.1016/j.jbiomech.2018.08.028>.
- [21] M. Juneja, P. Garg, R. Kaur, P. Manocha, B.S. Prateek, et al., A review on cephalometric landmark detection techniques, *Biomed. Signal Process. Control* 66 (2021) 102486, <https://doi.org/10.1016/j.bspc.2021.102486>.
- [22] H. Jacinto, R. Kéchichian, S. Valette, R. Prost, Positioning of anatomical landmarks in orthopedics by MESH registration, *IEEE International Conference on Image Processing (ICIP)* 2014 (2014) 3572–3576, <https://doi.org/10.1109/ICIP.2014.7025725>.
- [23] C.-B. Phan, S. Koo, Predicting anatomical landmarks and bone morphology of the femur using local region matching, *Int. J. Comput. Assist. Radiol. Surg.* 10 (2015) 1711–1719, <https://doi.org/10.1007/s11548-015-1155-8>.
- [24] V. Nguyen, L.F. Alves Pereira, Z. Liang, F. Mielke, J. Van Houtte, J. Sijbers, et al., Automatic landmark detection and mapping for 2D/3D registration with BoneNet, *Front. Vet. Sci.* 9 (2022) 923449, <https://doi.org/10.3389/fvets.2022.923449>.
- [25] J. Fürmetz, J. Sass, T. Ferreira, J. Jalali, L. Kovacs, F. Mück, et al., Three-dimensional assessment of lower limb alignment: Accuracy and reliability, *Knee* 26 (2019) 185–193, <https://doi.org/10.1016/j.knee.2018.10.011>.
- [26] S. Ahlback, Osteoarthritis of the knee. a radiographic investigation, *Acta Radiol Diagn (stockh) Suppl* 277 (1968) 7–72.
- [27] T. Iwano, H. Kurosawa, H. Tokuyama, Y. Hoshikawa, Roentgenographic and clinical findings of patellofemoral osteoarthritis. with special reference to its relationship to femorotibial osteoarthritis and etiologic factors, *Clin. Orthop. Relat. Res.* (1990) 190–197.
- [28] Y. Takakura, Y. Tanaka, T. Kumai, S. Tamai, Low tibial osteotomy for osteoarthritis of the ankle. results of a new operation in 18 patients, *J. Bone Joint Surg. Br.* 77 (1995) 50–54.
- [29] D.J. Hunter, A. Guermazi, G.H. Lo, A.J. Grainger, P.G. Conaghan, R.M. Boudreau, et al., Evolution of semi-quantitative whole joint assessment of knee OA: MOAKS (MRI Osteoarthritis knee score), *Osteoarthritis Cartilage* 19 (2011) 990–1002, <https://doi.org/10.1016/j.joca.2011.05.004>.

- [30] A.A. Khalifa, A.B. Mullaji, A.M. Mostafa, O.A. Farouk, A Protocol to Systematic Radiographic Assessment of Primary Total knee Arthroplasty, *Orthop. Res. Rev.* 13 (2021) 95–106, <https://doi.org/10.2147/ORR.S320372>.
- [31] P.J. Besl, N.D. McKay, A method for registration of 3-D shapes, *IEEE Trans. Pattern Anal. Mach. Intell.* 14 (1992) 239–256, <https://doi.org/10.1109/34.121791>.
- [32] K.S. Arun, T.S. Huang, S.D. Blostein, Least-Squares Fitting of two 3-D Point Sets, *IEEE Trans. Pattern Anal. Mach. Intell.* PAMI-9 (1987) 698, <https://doi.org/10.1109/TPAMI.1987.4767965>.
- [33] J.R. Hurley, R.B. Cattell, The procrustes program: Producing direct rotation to test a hypothesized factor structure, *Behav. Sci.* 7 (1962) 258–262, <https://doi.org/10.1002/bs.3830070216>.
- [34] J. Victor, D. Van Doninck, L. Labey, B. Innocenti, P.M. Parizel, J. Bellemans, How precise can bony landmarks be determined on a CT scan of the knee? *Knee* 16 (2009) 358–365, <https://doi.org/10.1016/j.knee.2009.01.001>.
- [35] D. Liljequist, B. Elfving, R.K. Skavberg, Intraclass correlation - a discussion and demonstration of basic features, *PLoS One* 14 (2019) e0219854, <https://doi.org/10.1371/journal.pone.0219854>.
- [36] E. Bori, S. Pancani, S. Vigliotta, B. Innocenti, Validation and accuracy evaluation of automatic segmentation for knee joint pre-planning, *Knee* 33 (2021) 275–281, <https://doi.org/10.1016/j.knee.2021.10.016>.
- [37] J.M. Bland, D.G. Altman, Statistical methods for assessing agreement between two methods of clinical measurement, *Lancet* 1 (1986) 307–310.
- [38] S.-Y. Baek, J.-H. Wang, I. Song, K. Lee, J. Lee, S. Koo, Automated bone landmarks prediction on the femur using anatomical deformation technique, *Comput. Aided Des.* 45 (2013) 505–510, <https://doi.org/10.1016/j.cad.2012.10.033>.
- [39] B. Derrick, D. Toher, P. White, Why Welch's test is Type 1 error robust, *The Quantitative Methods for Psychology* 12 (2016) 30–38.
- [40] della Croce U, Cappozzo A, Kerrigan DC., Pelvis and lower limb anatomical landmark calibration precision and its propagation to bone geometry and joint angles, *Med. Biol. Eng. Comput.* 37 (1999) 155–161, <https://doi.org/10.1007/BF02513282>.
- [41] H. Jacinto, S. Valette, R. Prost, Multi-atlas automatic positioning of anatomical landmarks, *J. Vis. Commun. Image Represent.* 50 (2018) 167–177.
- [42] S.J. Jang, K.N. Kunze, J.C. Casey, J.R. Steele, D.J. Mayman, S.A. Jerabek, et al., Variability of the femoral mechanical-anatomical axis angle and its implications in primary and revision total knee arthroplasty, *Bone Jt Open* 5 (2024) 101–108, <https://doi.org/10.1302/2633-1462.52.BJO-2023-0056.R1>.
- [43] R.J.A. Kuiper, P.R. Seevinck, M.A. Viergever, H. Weinans, R.J.B. Sakkers, Automatic Assessment of Lower-Limb Alignment from Computed Tomography, *J. Bone Joint Surg. Am.* 105 (2023) 700–712, <https://doi.org/10.2106/JBJS.22.00890>.
- [44] R. Rostamian, M. Shariat Panahi, M. Karimpour, H.G. Kashani, A. Abi, A deep learning-based multi-view approach to automatic 3D landmarking and deformity assessment of lower limb, *Sci. Rep.* 15 (2025) 534, <https://doi.org/10.1038/s41598-024-84387-z>.