Automated Ankle Range-of-Motion Measurement for the Hemophilia Joint Health Score: Marker-Based and Markerless Pose Estimation on iPhone

Imad Azizi, Christoph Mauel, Nicolás Saravia, and Martina Sepúlveda

University of Bonn, Bonn, Germany

Abstract. Ankle range of motion (ROM) is a core item of the Hemophilia Joint Health Score (HJHS), the standard instrument for tracking joint damage in people with hemophilia, and the ankle is one of the joints most affected by hemophilic arthropathy in adults. An existing iPhone application, Digital HJHS 2.1 (DHJHS), already digitizes the score, but did not provide sufficient accuracy for ankle ROM measurements. Building on that application, we developed and evaluated two approaches against manual goniometer measurements: a marker-based approach that tracks reflective markers under the camera flashlight, and a markerless approach that estimates foot keypoints directly from an image using deep pose-estimation models. Each approach was assessed in two stages, a setting exploration to select the best capture configuration and a main study across four participants and two different raters. Due to well-known issues with differences in goniometer usage, our initial ground truth systematically overestimated plantarflexion by roughly 21°, so we repeated the reference measurements under qualified guidance and re-ran the entire evaluation. Against the corrected ground truth the marker-based approach reached a mean absolute error (MAE) of 2.63° with no measurement deviating from the goniometer by more than 10°, whereas the best markerless model, RTMPose, reached 9.02° MAE and was accurate everywhere except in plantarflexion. The complete pipeline runs on the device. Our results indicate that a phone can measure ankle ROM at a clinically useful accuracy, and that the reproducibility of the reference measurement is at least as important as the measurement method itself.

Keywords: Hemophilia · Range of Motion · Pose Estimation · Goniometry · Mobile Health

1 Introduction

Hemophilia is an inherited bleeding disorder in which recurrent bleeding into the joints progressively destroys them. Each bleed can provoke synovitis and, over time, irreversible cartilage and bone damage, a process known as hemophilic

arthropathy. Clinicians monitor this damage with the Hemophilia Joint Health Score (HJHS), a structured physical examination that was originally developed and validated in children and has since demonstrated strong construct validity for detecting arthropathy in adults. [4,17] Range of motion is scored for every joint, and the ankle deserves particular attention, because it is the joint most commonly affected by hemophilic arthropathy and the most frequent source of joint pain in several countries [7].

In clinical practice the reference instrument for range of motion is the universal goniometer, a manual protractor whose two arms are aligned with the leg and the foot. Goniometry is fast and inexpensive, but its reliability is limited and depends heavily on the examiner. Systematic reviews report that agreement within one examiner is generally acceptable, whereas agreement between examiners is weaker, and for the ankle the evidence for reliable between-examiner plantarflexion measurement is especially scarce. [12,5] At the same time, a systematic review of clinically accessible smartphone applications found that many reach adequate reliability and validity for range-of-motion measurement, provided this is verified for the specific joint of interest [11]. Taken together, these findings motivate augmenting, or eventually replacing, the manual instrument with an automated measurement on a device that clinicians already carry.

This project extends the Digital Hemophilia Joint Health Score (DHJHS), a native iOS application that already digitizes the complete score but is not yet in clinical use [6]. The DHJHS offers a camera-based angle measurement that provided sufficient accuracy for the knee and the elbow, but struggled with the ankle. The ankle is intrinsically difficult to capture: the foot is small, its plane is easily misjudged from a single viewpoint, and curling of the toes (forefoot flexion) is readily mistaken for true ankle motion. Our objective was therefore to make ankle range-of-motion measurement on the phone accurate, robust to forefoot flexion, reproducible across patients, and able to run entirely on the device (Section 2).

We addressed the problem from two complementary directions and evaluated them under one shared pipeline. The first is a marker-based approach that attaches a small number of reflective markers to well defined anatomical landmarks and reads them under the camera flashlight, deriving the joint angle from their geometry. The second is a markerless approach that uses a deep pose-estimation network to locate the relevant keypoints directly from the image, with no markers at all. For the markerless approach we compared four publicly available frameworks: RTMPose Wholebody [8,9], RTMPose Body-with-Feet [8], MediaPipe [1], and MeTRAbs [16].

This paper makes three contributions. First, we present a working on-device ankle ROM measurement integrated into DHJHS, in both a marker-based and a markerless variant. Second, we report a controlled evaluation of both approaches against a goniometer, first to select the best capture setting and then to quantify accuracy across several participants and raters. Third, we highlight the limitations of the goniometer as a measurement tool, as inter-rater variability introduced a systematic bias into our ground truth, affecting our results.

Concretely, we established four core objectives for the automated ankle measurement. The system must:

(i) **Achieve clinical accuracy:** Measure ankle range of motion with a maximum deviation of 5°, which is the widely accepted clinical threshold [5].
(ii) **Ensure robustness to forefoot flexion:** Isolate true ankle motion from toe curling, a common confounder when capturing the joint from a single viewpoint.
(iii) **Maximize reproducibility:** Deliver consistent results across different patients and examiners to minimize user-dependent variability.
(iv) **Operate strictly on-device:** Execute the complete measurement pipeline locally on the iPhone without requiring network access, fully integrated into the existing DHJHS application rather than acting as a standalone prototype.

We revisit these objectives in Section 4.

2 Methods

This section describes the two measurement approaches, the application into which they were integrated, the design of the two-stage study, and the metrics and statistical tests used for evaluation. We first set the context and the existing application on which the work builds, then present each approach in turn, and finally the evaluation methodology.

2.1 Overview and starting point

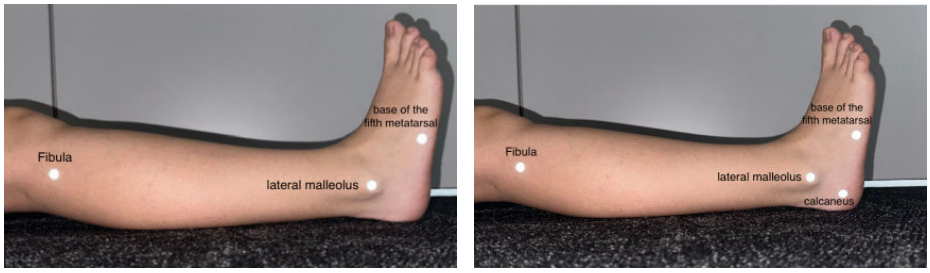
The project consists of two parts that share a single definition of the ankle angle. The iOS application performs the measurement live on the phone, and an offline evaluation pipeline, written in Python, applies the same marker detection and the same angle definitions to recorded images and videos so that accuracy can be quantified against the goniometer. Using one angle definition on both sides was a deliberate decision, because it ensures that the numbers reported in this paper reflect the behaviour that a clinician observes in the application.

The application we build on, DHJHS, is a native iOS app that digitizes HJHS 2.1 [6]. It covers all eight assessment categories, replaces the paper score sheets with guided input screens, and computes the total score automatically from the normative tables of the HJHS manual, which removes the manual summation and transcription steps of the paper workflow. For range of motion, DHJHS estimates joint angles from the camera using the MediaPipe pose framework, which at the time was the only one of the compared frameworks that ships directly for iOS. It pairs a lightweight model for the live preview with a heavier model for a captured photo or short video, then removes outliers and averages the remaining per-frame angles to produce a single value. In the previous group's feasibility testing against a manual goniometer, the elbow and the knee matched the reference closely, whereas the ankle was the least consistent

joint [6]. Closing that gap is the starting point of our work. We kept the existing documentation, scoring, and overall application structure and focused on the ankle; the marker-based approach and the on-device markerless pipeline described below are our new contributions, added on top of this existing and functioning system.

2.2 Marker-based measurement

The marker-based approach attaches small, round, reflective markers to defined anatomical landmarks of the leg and foot. Under the camera flashlight the markers separate well from the background because of their reflective coating, appearing as bright, uniform white circles. The application offers two sets of marker placement, selectable before the measurement, illustrated in Fig. 1. The three-marker set (Fig. 1a) uses the head of the fibula, the lateral malleolus, and the base of the fifth metatarsal. The four-marker set (Fig. 1b) adds the lateral tuberosity of the calcaneus (heel).



(a) Three-marker set: the angle θ_3 is taken at the malleolus vertex.

(b) Four-marker set: the angle θ_4 lies between the leg line and the foot line.

Fig. 1: Marker placement and angle construction on the lateral side of the leg and foot. Both sets share the fibula, the lateral malleolus and the base of the fifth metatarsal; the four-marker set adds the calcaneus (heel). The coloured lines and arcs show how the ankle angle is constructed for each set.

We support both sets because they serve different purposes. The four-marker set reproduces the classical goniometer construction, with one line along the leg and one along the lateral border of the foot, which keeps both lines anatomically meaningful. The three-marker set was explored during development while looking for possible placements with fewer markers. Removing the heel marker turned out to work well in testing, removes one landmark, and saves a small amount of setup time. All chosen landmarks, the fibula head, the lateral malleolus, the base of the fifth metatarsal and the calcaneus, are bony prominences that can be located reliably by palpation. We selected them deliberately for this reason: landmarks that different examiners find in the same place make the marker placement more

reproducible and reduce the subjective influence of the individual rater on the measurement.

Angle computation and calibration. Both sets express the measurement relative to the patient's own neutral position, which is measured in the first recording of the assessment. We call this step the calibration, and every later value is reported as the deviation from this recorded neutral, so that differences in resting ankle position between patients do not affect the result. For the three-marker set (Fig. 1a) the angle is taken directly at the lateral malleolus and reads about 130° at neutral. For the four-marker set (Fig. 1b) the angle is taken between the leg line (fibula to malleolus) and the foot line (calcaneus to metatarsal), reproducing the goniometer and reading close to 90° at neutral. Because both are expressed relative to the recorded neutral, these different resting values are irrelevant. In the offline evaluation, we deliberately chose not to use calibration because we wanted to test the system's basic ability to detect specific angles. Therefore, a fixed neutral of about 130° is used instead for the three-marker set. In the case of video recordings, the system collects the per-frame angles over a capture window of about 1.5 seconds, discards outliers, and averages the rest.

Let M denote the two-dimensional image position of the lateral malleolus, and let F, T and C denote the positions of the fibula, the fifth metatarsal and the calcaneus. The raw three-marker angle, taken at the malleolus, is given by Eq. (1), and the four-marker angle, taken between the leg line and the foot line, by Eq. (2):

$$\theta_3 = \arccos\left(\frac{(F-M)\cdot(T-M)}{\|F-M\|\,\|T-M\|}\right), \tag{1}$$

$$\theta_4 = \arccos\left(\frac{(F-M)\cdot(T-C)}{\|F-M\|\,\|T-C\|}\right). \tag{2}$$

The value presented to the clinician is the deviation from the recorded neutral angle θ_0 (Eq. (3)),

$$\Delta\theta = \theta - \theta_0, \tag{3}$$

where positive values denote plantarflexion and negative values dorsiflexion.

Marker detection. Detection is identical for both marker sets. We use a blob detector rather than a neural network, to ensure that it runs at the live camera frame rate. Each frame is converted to grayscale and thresholded adaptively against the brightest region, so that the reflective markers stand out under the flashlight.

The bright regions are then grouped into clusters, each representing a candidate for marker positions, which is then filtered by its area, its fill ratio (the fraction of its bounding box that is actually filled), its aspect ratio, and its circular symmetry. A candidate set of markers is accepted only when four independent criteria agree: the markers are bright under the flashlight, they are

round, they are all of a similar size, and their spatial arrangement corresponds to the expected relative positions of leg and foot landmarks. If any criterion fails, the application keeps searching instead of reporting a value.

We found that the similar-size criterion is the most useful in our specific setting, because the physical markers are manufactured identically, so a group of equally sized circles forming a leg is almost always the real set, while isolated reflections, for example on the floor, differ in size and are rejected.

2.3 Markerless measurement

The markerless approach locates the relevant landmarks with a deep pose-estimation network. To choose which networks to test, we followed the comparative benchmark of monocular pose-estimation frameworks by Kahl et al. [10], which evaluated sixteen frameworks against a Vicon motion-capture reference and found MeTRAbs to be the best overall. Our decisive requirement was that a framework must provide a keypoint on the foot, since without a foot landmark the ankle angle cannot be formed. This requirement excluded many otherwise strong frameworks. Among those that do provide foot keypoints, we selected four that performed well in the benchmark and could realistically be run on iOS: RTMPose Wholebody, which we abbreviate as RTMPose [8,9], RTMPose Body-with-Feet, abbreviated as BodyFeet [8], MediaPipe [1], and MeTRAbs [16].

For every framework the ankle angle is built from three keypoints: the knee, the heel, and the small toe. The angle is then computed exactly as in the marker-based approach, as the angle at the heel between the direction to the knee and the direction to the small toe. Three of the frameworks (RTMPose, BodyFeet and MediaPipe) return two-dimensional keypoints, so the angle is computed in the image plane, whereas MeTRAbs returns metric three-dimensional keypoints, to which the identical formula is applied in three dimensions. This lets the same evaluation code assess a three-dimensional method alongside the two-dimensional ones without any change to the pipeline.

On the device we run the selected model through ONNX Runtime [15], a cross-platform inference engine that lets us integrate models which are not available as a native iOS package. Inference runs entirely on the phone, so no image or patient data leaves the device, which matters for medical data, and the application executes the same model file that the offline evaluation used, so its behaviour is consistent with the accuracy we report. In the application, the markerless measurement takes a single photo rather than a video.

2.4 Study design and data

Each approach was evaluated in two stages. For the marker-based approach, the three-marker setup was used throughout the entire evaluation. The setting exploration is a broad sweep over the capture parameters, recorded from a single participant, whose purpose is to find the best way to capture the data. The investigated parameters (summarized in Table 1) include the distance to the camera, the presence or absence of a sock, the lighting conditions, the camera height, and

the recording modality. The main study then fixes the single best configuration and records four participants with two different goniometer operators (rater).

Table 1: Evaluation parameters investigated for the marker-based and markerless approaches.

parameter	marker-based approach	markerless approach
distance to camera	0.5m, 1m	0.5m, 1m
sock presence	sock, no sock	sock, no sock
lighting conditions	bright, dim	N/A
camera height	0cm, 30cm	0cm, 30cm
recording modality	photos, videos	photos, videos
frameworks	N/A	RTMPose, BodyFeet, Mediapipe, MeTRAbs

Because the two approaches favoured different media, each main-study dataset contains only the corresponding type: videos for the marker-based approach and photos for the markerless approach. All participants were healthy volunteers. For every participant we recorded three defined positions in the same order: the neutral position at about 90°, the maximum dorsiflexion, and the maximum plantarflexion. During each recording the participant lay supine on the floor with the leg flat and fully extended, and the foot was filmed from the lateral side. All images and videos were captured with an iPhone 13 Pro Max (Apple A15 Bionic, 12 MP wide-angle camera), the same class of device on which the application runs. Ground truth was read from a goniometer and encoded in each filename together with the capture parameters, so that the pipeline parses it automatically and calibrates every participant against their own neutral recordings. Table 2 summarizes the resulting datasets.

Table 2: Dataset sizes. The setting exploration sweeps the capture parameters on a single participant; the main study fixes the best configuration and adds participants and a second rater.

approach	stage	measurements	participants
marker-based	setting exploration	71 images, 24 videos	1
marker-based	main study	36 videos	4 (2 raters)
markerless	setting exploration	62 images, 61 videos (×4 frameworks)	1
markerless	main study	57 images (RTMPose)	4 (2 raters)

2.5 Metrics and statistics

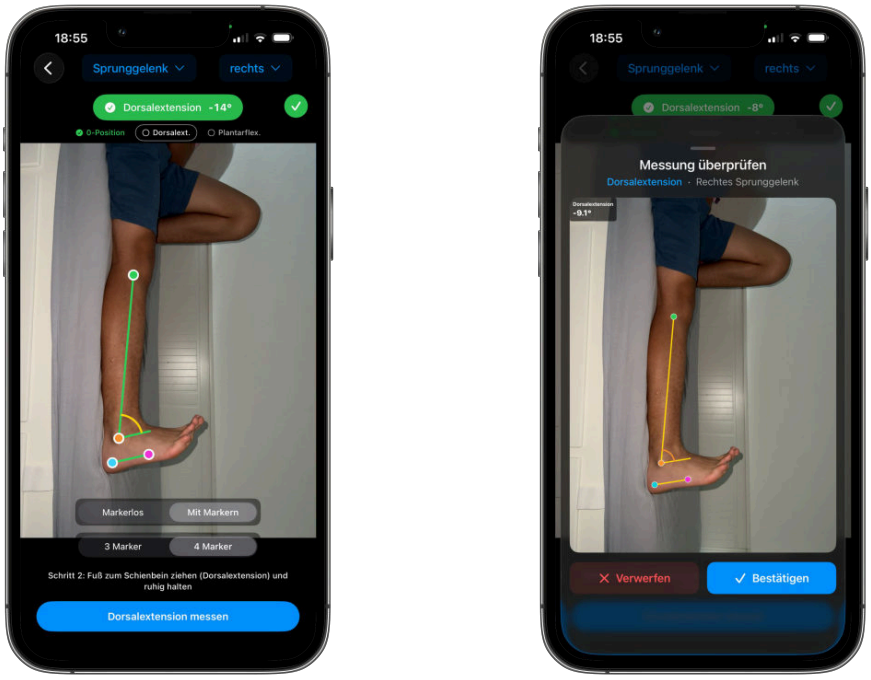
We evaluate accuracy with two main metrics. The first is the mean absolute error (MAE) in degrees against the goniometer. The second is a set of accuracy buckets, adopted from Gawenda et al. [6], that classify each measurement by its deviation from the reference and relate it to the typical 3° to 5° error of manual goniometry: up to 3° is a hit, 3° to 6° is tolerated, 6° to 10° is an inaccuracy, and more than 10° is a miss. We additionally report the mean signed error (the bias, where a negative value means the method reads lower than the goniometer), the standard deviation of the error as a measure of precision, and the range-of-motion (ROM) slope, which is the slope of predicted against true angle: a value of 1.0 is ideal, below 1.0 means the measured range is compressed, and above 1.0 means it is exaggerated.

For the statistical comparisons we used non-parametric tests, which do not assume normally distributed errors. The Kruskal-Wallis test checks whether three or more groups (for example the four frameworks, or the three movement classes) differ overall, and the Mann-Whitney U test checks whether two groups (for example images against videos, or sock against no sock) differ. Where several such comparisons were made together, we applied the Holm correction, which lowers the chance of a false positive across the whole set of tests. All tests used a significance level of 0.05.

2.6 The iOS application

The additions above were integrated into the existing DHJHS codebase, which is written in SwiftUI, persists its data with SwiftData, and drives the camera through AVFoundation. Our marker detector is implemented in Swift and runs on the live camera stream, while the markerless models run through ONNX Runtime as described in Section 2.3. In marker mode the flashlight is switched on automatically, since the reflective markers depend on it, and the live angle overlay is drawn directly on the leg while the app warns when the foot rotates out of the measurement plane.

Guided measurement workflow. The application guides the user through the measurement in a fixed order, shown in Fig. 3. The clinician first selects the joint and the side, then chooses the approach and, for the marker-based approach, the marker count, as seen in Fig. 2a. The neutral 0-position is recorded first, which fixes the patient's calibration reference, followed by dorsiflexion and plantarflexion. During each movement the app shows the live angle overlay in marker mode, or captures a single annotated photo in markerless mode. Each measurement can then be confirmed or discarded by the clinician on a review sheet (Fig. 2b) before it is stored for that joint and side, overwriting any previous value. The interface is available in German and English, and the application needs no network connection, so it works in an ordinary examination room.



(a) Camera view with method and marker-count selection.

(b) Result review with the annotated angle.

Fig. 2: The DHJHS measurement flow on the phone. The placeholders are to be replaced with screenshots of the running application.

3 Results

We first describe the ground-truth problem that we uncovered during the project and how we resolved it, because it frames every number that follows. We then report the corrected results for the marker-based and the markerless approach in turn, presenting the setting exploration before the main study in each case.

3.1 A systematic bias in the ground truth

Our earliest evaluation runs, carried out with the markerless approach against the initial goniometer measurements, looked encouraging at first. The best framework, RTMPose, reached an MAE of 5.24° on images in the setting exploration and 5.59° in the main study. However, a closer analysis of the error distribution revealed a systematic negative bias of about −5° overall, largest in plantarflexion, where the model measured on average 6.5° below the reference (Fig. 4). An

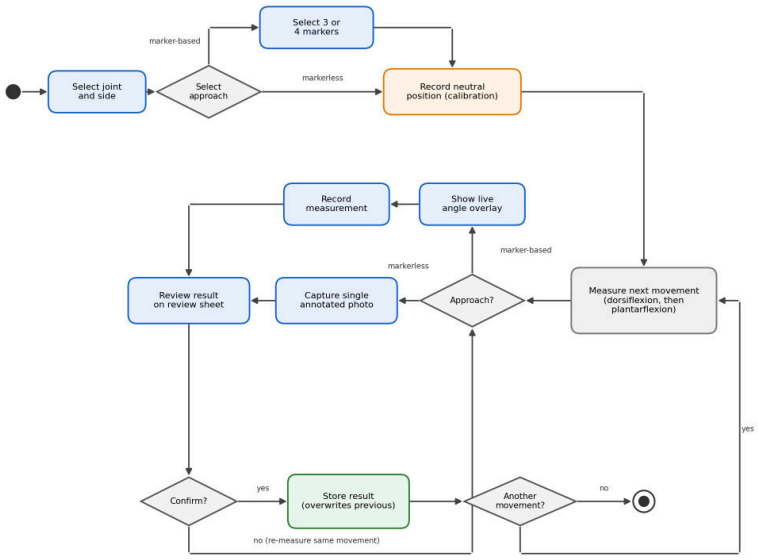


Fig. 3: Guided measurement workflow in the application.

error that is offset by an almost constant amount in one direction usually points to a problem in the reference rather than in the model (Table 3).

Table 3: Markerless RTMPose against the original (uncorrected) goniometer ground truth.

stage	mode	MAE (°)	bias (°)	plantarflexion bias (°)
setting exploration	images	5.24	−1.87	−5.92
setting exploration	videos	5.77	−1.97	–
main study	images	5.59	−5.11	−6.48

We therefore repeated the reference measurements, this time under qualified guidance. The corrected values were markedly lower than the originals. Averaged over the four participants, the maximum plantarflexion recorded by the original goniometry was 55.5°, against 34.2° for the corrected measurement, a difference of 21.2° or about 38 % (Fig. 5a). As displayed in Figure 5b, the first measurements overestimated the plantarflexion for every participant.

These findings highlight the well-documented challenges of manual goniometry. Because the measurement is highly sensitive to subtle adjustments in technique and positioning, inter-rater reliability is frequently reported as a limiting factor in the literature [12,5]. There are numerous valid approaches to applying

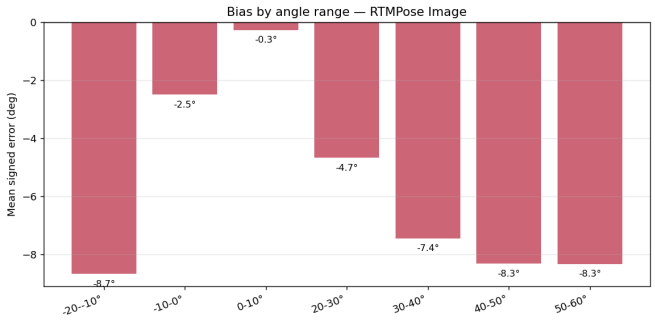
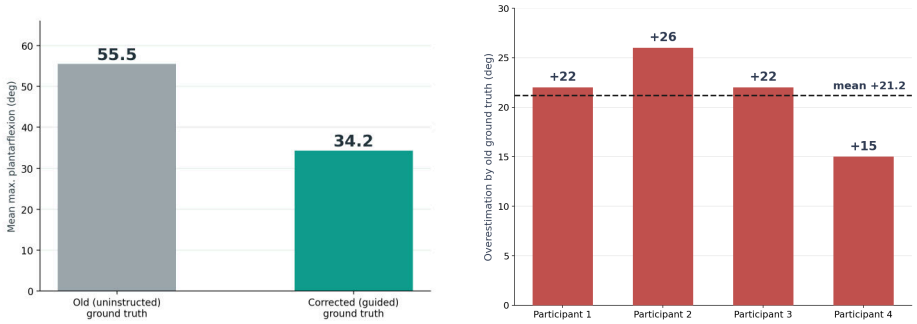


Fig. 4: Signed error of RTMPose across the angle range in the markerless main study, evaluated against the original goniometer ground truth.



(a) Mean maximum plantarflexion.

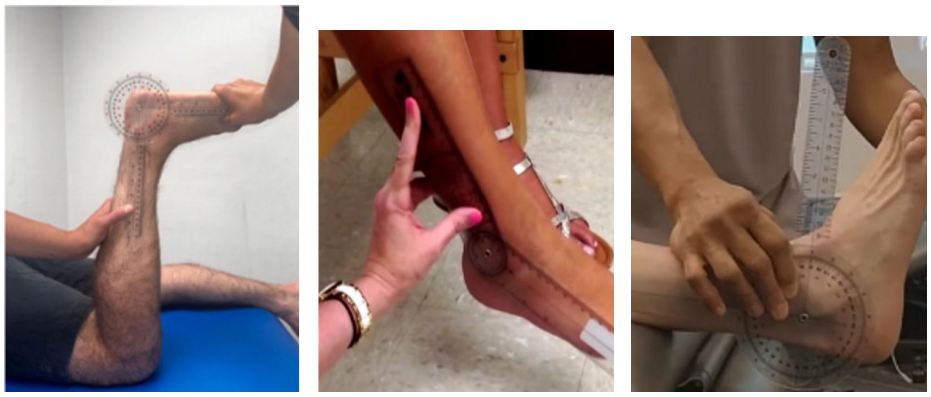
(b) Overestimation per participant.

Fig. 5: The original goniometer reference overestimated plantarflexion consistently. Correcting it lowered the mean maximum plantarflexion from 55.5° to 34.2°, and the overestimation was present for every participant.

a goniometer, and Figure 6 illustrates a selection of these different measurement methods, showing how easily variations can occur in practice. All results in the remainder of this section use the corrected ground truth.

3.2 Marker-based approach

Setting exploration. The marker-based setting exploration dataset was recorded from a single participant across all capture parameters, as photos and as videos. Accuracy was high in both media, with an MAE of 1.50° on images and 1.07° on videos and a ROM slope close to 1.0, which means the range of motion was captured without compression (Table 4). Of the capture parameters, only the sock had a significant effect: when the participant was wearing no sock the error on images was 0.90°, rising to 2.08° when wearing a sock ($p < 0.001$), whereas distance ($p = 0.74$) and room lighting ($p = 0.25$) had no significant effect. The



(a) Measurement in prone position, knee at 90°. [2]

(b) Measurement with the patient sitting, legs hanging off the table. [13]

(c) Measurement in supine position, knee extended. [3]

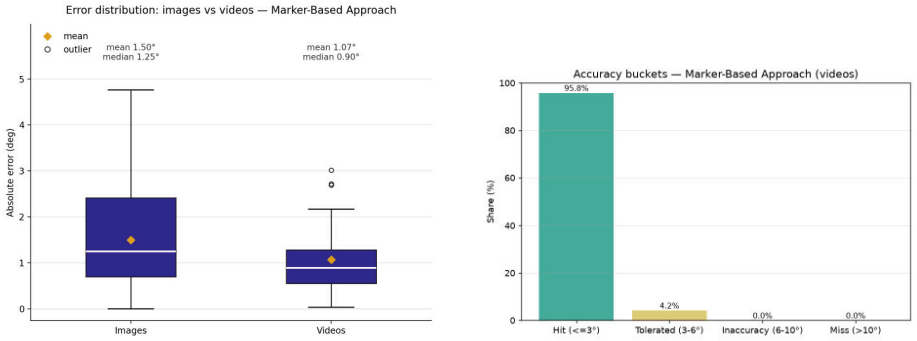
Fig. 6: Selection of different valid measurement methods using a manual goniometer.

best configuration is therefore a bare ankle recorded on video, with the distance and the lighting free to vary. Since the sock was the only parameter with a significant effect, we recorded the main study with a bare ankle on video and fixed the remaining, non-significant parameters at a camera distance of 0.5 m in a dimly lit room.

Table 4: Marker-based setting exploration: accuracy per media type (single participant).

mode	n	MAE (°)	RMSE (°)	bias (°)	slope	time (ms/file)
images	71	1.50	1.83	0.16	1.065	62
videos	24	1.07	1.34	0.37	1.042	2178

Videos were slightly more accurate than images (1.07° against 1.50°), although the two did not differ significantly ($p = 0.10$), and the video error distribution exhibited lower variance (1.74° compared to 3.39° for images), with fewer extreme cases (Fig. 7a). Processing a video took about 2178 ms per file against 62 ms for an image in the offline pipeline (Fig. 10), which is still well within acceptable limits, so we selected videos for the marker-based main study. Figure 7b displays the accuracy buckets. The marker-based approach achieved 95.8 % hits on video measurements, with no misses in either medium.



(a) Error distribution (b) Accuracy buckets

Fig. 7: Marker-based setting exploration: (a) error distribution for images against videos, and (b) accuracy buckets for video measurements.

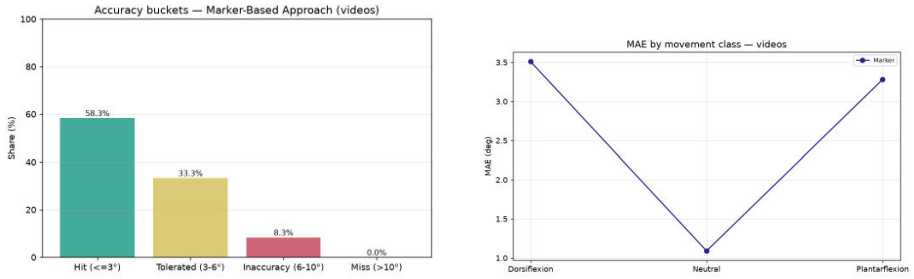
Main study. In the main study the best configuration, a bare ankle recorded on video, was captured from four participants with two raters and evaluated against the corrected goniometer. The results showed an MAE of 2.63° (SD = 2.09°) with a 95 % confidence interval of [1.92°, 3.34°], and a ROM slope of 1.035 (Table 5). There was no systematic bias found, with a mean signed error of 0.18° in the data. The largest single error over all 36 measurements was 6.96°, so no prediction fell into the miss bucket, while 58.3 % of predictions were hits, 33.3 % tolerated inaccuracies, and 8.3 % inaccuracies (Fig. 8a). Notably, 88.9 % (32 of 36) of all measurements fell within a 5° error margin. By movement class, neutral was the most accurate at 1.09° MAE, while the MAE for dorsiflexion (3.51°) and plantarflexion (3.28°) were higher. The difference between movement classes was significant (Kruskal-Wallis $p = 0.008$; Fig. 8b). Moreover, the predicted angles showed a near-perfect correlation to their respective ground truth measurement (Pearson r = 0.987).

Table 5: Main study, best configuration per approach, evaluated against the corrected goniometer.

approach	mode	n	MAE (°)	bias (°)	max$\|e\|$ (°)	slope	miss
marker-based	videos	36	2.63	0.18	6.96	1.035	0 %
markerless (RTMPose)	images	57	9.02	6.80	30.02	1.423	40.4 %

3.3 Markerless approach

Setting exploration. The markerless setting exploration compared the four frameworks over images and videos, again from a single participant. RTMPose



(a) Accuracy buckets.

(b) MAE by movement class.

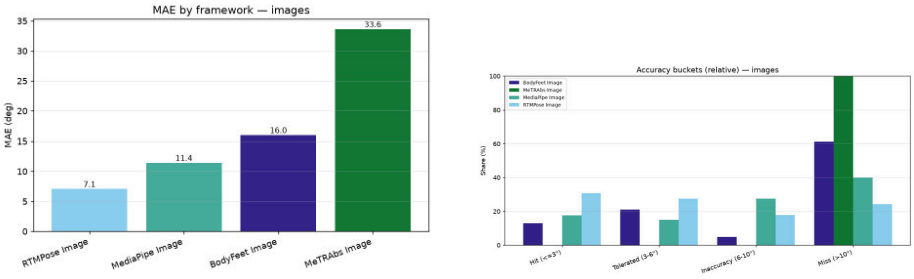
Fig. 8: Marker-based main study against the corrected ground truth. Every measurement was within 10° of the goniometer, and the neutral position was measured more accurately than dorsiflexion and plantarflexion.

was the most accurate framework in both media, with an MAE of 7.06° on images and 7.41° on videos, followed by MediaPipe, while BodyFeet and MeTRAbs achieved lower accuracies (Table 6, Fig. 9). The difference between frameworks was highly significant (Kruskal-Wallis $p < 10^{-23}$), and RTMPose was significantly more accurate than MediaPipe on images ($p = 0.03$), although the two did not differ significantly on videos ($p = 0.10$). MediaPipe also failed to detect the pose in a large fraction of images (detection rate 64.5 %), which is a practical disadvantage.

Table 6: Markerless setting exploration: accuracy per framework on images (single participant). Processing times are compared across all configurations in Fig. 10.

framework	n det.	%	MAE (°)	RMSE (°)	time (ms/file)
RTMPose	62	100.0	7.06	9.55	510
MediaPipe	40	64.5	11.44	16.25	406
BodyFeet	62	100.0	15.98	21.11	451
MeTRAbs	62	100.0	33.63	34.75	1692

The choice of medium was decided by processing time rather than by accuracy. On images RTMPose took about 510 ms per file, but on videos about 44500 ms, roughly ninety times longer, because the network has to run on every frame. The other frameworks behaved similarly. Figure 10 compares the per-file processing time of the main configurations on a logarithmic scale. Markerless video (about 44500 ms for RTMPose) is by far the slowest, roughly two orders of magnitude above the markerless image configurations and the marker-based detector. Since utilizing video recordings did not improve markerless accuracy and made processing impractical, we selected images for the markerless main



(a) Mean absolute error per framework. (b) Accuracy buckets per framework.

Fig. 9: Markerless setting exploration on images. RTMPose is the most accurate framework, both by mean error and across the accuracy buckets.

study. The best image configuration for RTMPose was a distance of about 1 m, the camera near floor height, and a bare ankle, although for RTMPose none of these parameters was significant on its own.

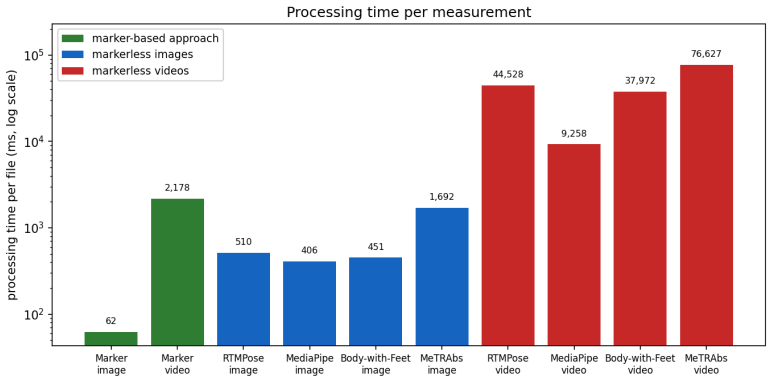
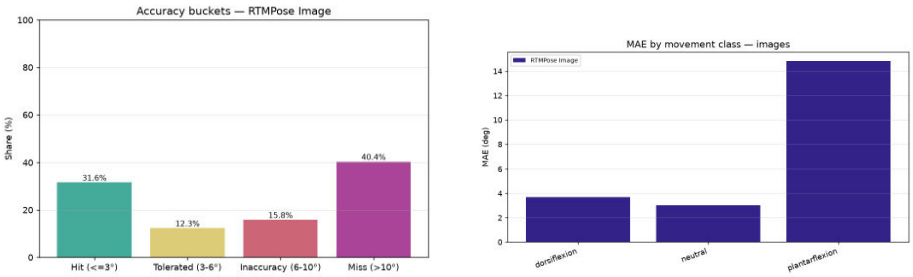


Fig. 10: Processing time per file for the main configurations, on a logarithmic scale.

Main study. In the main study, RTMPose was evaluated on images of the four participants against the corrected goniometer values. The results showed an MAE of 9.02° (SD = 7.59°) with a 95 % confidence interval of [7.00°, 11.03°], and a ROM slope of 1.423 (Table 5). In contrast to the marker-based approach, the markerless approach also carried a large positive mean signed error (bias) of 6.80°, meaning it systematically overestimated the ankle angle. Fig. 11a shows the accuracy buckets. The largest fraction of predictions were classified as

misses with 40.4 %, while only 31.6 % were hits, 12.3 % tolerated inaccuracies, and 15.8 % inaccuracies. Overall, only 38.6 % (22 of 57 measurements) remained within a 5° error margin.

The errors were not spread evenly across the movements. Neutral (3.04°) and dorsiflexion (3.71°) were quite accurate and comparable to the marker-based approach, while the deviation in plantarflexion was far larger at 14.82°. This difference was highly significant (Kruskal-Wallis $p < 10^{-7}$; Fig. 11b and Fig. 12). Furthermore, plantarflexion was responsible for the vast majority of misses, with 95.65 % of predictions in that category coming from plantarflexion recordings.



(a) Accuracy buckets.

(b) MAE by movement class.

Fig. 11: Markerless (RTMPose) main study against the corrected ground truth. Plantarflexion is the dominant source of both the error and the misses.

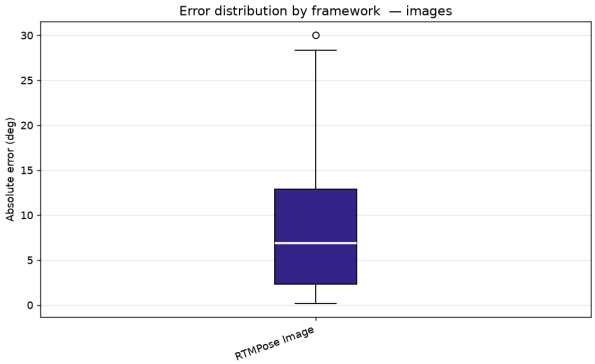


Fig. 12: Distribution of the absolute error of RTMPose in the markerless main study. Most measurements are accurate, but a long tail of large errors, concentrated in plantarflexion, raises the mean.

4 Discussion

The two approaches reached very different levels of accuracy. The marker-based approach measured ankle ROM with an MAE of 2.63°, with 88.9 % of predictions falling into the clinically acceptable range of 5° [5]. Furthermore, not a single measurement deviated from the goniometer by more than 10°. This highlights the potential of the marker-based approach and is in line with the reliability that smartphone-based ROM tools have shown for other joints [11]. In contrast, the markerless approach yielded an MAE of 9.02° and a miss rate of over 40 %, rendering it unsuitable for clinical use in its current state. However, it is striking that over 95 % of misses stem from plantarflexion recordings, while predictions for neutral position (MAE 3.04°) and dorsiflexion (MAE 3.71°) fall well within the clinically acceptable range of 5°. This isolates the plantarflexion as the main issue in the markerless approach.

We attribute this to the positioning of the keypoints used for angle computation. In the markerless approach, the choice of keypoints was strictly limited to the landmarks that are detected by the used pose estimation frameworks. Most of the considered frameworks only provided a very limited selection of foot keypoints, all of which were located in the toe area. During active plantarflexion, the foot undergoes a triplanar motion: not only does the ankle extend downward, but the midfoot (tarsometatarsal) and metatarsophalangeal joints also flex downward, causing the forefoot to naturally curl and roll inward. Simultaneously, the hindfoot joints tilt inward (inversion), creating a cohesive locking mechanism for the entire foot. [18] Consequently, the results are negatively affected by forefoot flexion. Unlike a goniometer, which aims to isolate the ROM of the ankle joint, the angle computed between the small toe, heel, and knee captures the combined ROM of all aforementioned joints. This also explains the large positive bias and the ROM slope exceeding 1.0.

A further contributing factor is that the used frameworks were trained for general human pose estimation rather than for precise clinical diagnostics. This observation is consistent with reports indicating that the accuracy of such networks at a given joint is limited by the anatomical labelling and quality of their training data, rather than by the pose estimation method itself [14].

The two approaches also differ in their practical application. While the marker-based approach excelled in terms of accuracy, it requires the time-consuming application of reflective markers by the clinician. This manual step introduces a slight subjective component to the assessment, which can theoretically limit broad standardization. However, since the chosen keypoints correspond to reproducible anatomical landmarks, the introduced error is expected to be minimal. Ultimately, the subjectivity involved in marker application is significantly less pronounced than the inherent user-dependency and misalignment risks seen in traditional goniometric measurements. In comparison, the markerless approach offers a major clinical advantage, as it does not require physical attachments, reducing rater subjectivity to a minimum, while saving time and material in the process.

The project also offered an important insight that was independent of our proposed solutions by highlighting the hidden unreliability of the underlying reference measurement. The original goniometry overestimated plantarflexion by about 21° for every participant, and because the error was systematic it appeared as a plausible constant bias rather than as obvious noise. This is in line with the goniometry literature, which identifies between-examiner agreement, and plantarflexion in particular, as the weak points of the manual measurement [12,5], and it is the reason we re-measured the reference under qualified guidance. This finding ultimately underscores the value of automated approaches, as their consistent execution effectively eliminates the inter-rater variability that typically compromises manual goniometry.

Beyond the measurement methods themselves, two choices in our implementation were important for the accuracy and the reproducibility reported above. By utilizing ONNX Runtime, we allow the markerless model to run directly on the device from the same model file used in the offline evaluation, which means that the accuracy we measured in Python is a realistic estimate of what the application actually produces, and it also keeps all patient images on the phone. The second concerns the reference against which each angle is measured. Every measurement is expressed relative to the patient's own recorded neutral position rather than a fixed textbook angle. This calibration matters because many patients cannot bring the ankle into the standard 90° neutral, whether because of hemophilic arthropathy or other conditions, so a fixed reference would be inappropriate for them. Measuring against each patient's individual resting position instead allows the same procedure to work reliably regardless of how far the ankle can move.

Returning to the objectives from Section 1, the degree of accomplishment depends on the approach. The marker-based approach achieved an MAE of 2.63° and stays robust across all movement classes, including plantarflexion, thereby satisfying objectives i and ii. The markerless approach fell short of the required 5° deviation with an MAE of 9.02°. Due to suboptimal keypoints, the markerless approach is negatively affected by forefoot flexion falling short of objective ii, as indicated by the poor performance on plantarflexion recordings. These issues also cause the markerless approach to miss out on the 5° deviation demanded by objective i. Both approaches run entirely on the device and are integrated into the existing DHJHS application (objective iv). Reproducibility (objective iii) is addressed by the palpable marker placement and the deterministic, automated angle computation, although a formal inter-rater study on patients remains future work.

4.1 Limitations

This work has clear limitations. The sample is not representative with only four participants, none of whom were actual hemophilia patients, so the numbers should be read as a feasibility result and not as a clinical validation. The corrected goniometer, although measured carefully and under qualified guidance, is still a manual instrument, whose measurement spread affects the obtained

results. Furthermore, all reported statistics and charts stem from the offline evaluation pipeline. While the application uses the same angle definitions, and we rebuilt the evaluation code as closely as possible, the transition from evaluation to application can still cause the application to produce slightly different results. Finally, the markerless keypoint set is fixed and not trained for this task, which is the root cause of the plantarflexion problem discussed above.

5 Conclusion

We have shown that ankle range of motion for the Hemophilia Joint Health Score can be measured on an ordinary iPhone, entirely on the device. The marker-based approach reached an MAE of 2.63° against a carefully measured goniometer, with 58 % of measurements within 3° and none deviating by more than 10°, which is accurate enough to be useful in practice. The markerless approach yielded an error of 9.02° and lacks overall reliability. Yet, by pinpointing its weakness specifically to plantarflexion instead of widespread failure, the results clearly show how the method can be optimized. Across both approaches the study showed that the reproducibility of the reference measurement is central, and that an automated, consistent measurement resolves the variability inherent to manual goniometry.

6 Outlook

The most promising next step follows directly from the movement-class results. While RTMPose achieved the highest overall accuracy and excelled in dorsiflexion and neutral positions, the generally weaker MediaPipe framework demonstrated markedly lower errors during plantarflexion with 3.63° compared to RTMPose's 17.78° (Table 7). A practical system could combine the two, using RTMPose in general and switching to MediaPipe for plantarflexion.

Table 7: Per-movement mean absolute error (°) on images in the setting exploration.

framework	dorsiflexion	neutral	plantarflexion
RTMPose	4.54	4.43	17.78
MediaPipe	11.98	14.15	3.63

Beyond such a hybrid, a model trained specifically for this clinical task, rather than a general-purpose keypoint detector, would be expected to close the plantarflexion gap that limits the markerless approach today, since that gap stems from the keypoint set rather than from the method. Finally, the complete system should be validated on a larger and more representative dataset that includes actual patients with hemophilia.

References

1. Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., Grundmann, M.: BlazePose: On-device real-time body pose tracking. arXiv preprint arXiv:2006.10204 (2020). https://doi.org/10.48550/arXiv.2006.10204
2. Bhatia, M., Dhiran, A., Jones, A.: Prone-position modified silfverskiöld test: Feasibility and reliability. Foot & Ankle Orthopaedics **11**(1), 24730114261417693 (2026). https://doi.org/10.1177/24730114261417693
3. Dr. Dennis Valdez, Certified Athleic Therapist: How do you measure ankle range of motion with a goniometer? Video (2020), youTube: https://www.youtube.com/watch?v=uOzcFzZ2V4U. Last accessed: 19.07.2026
4. Feldman, B.M., Funk, S.M., Bergstrom, B.M., Zourikian, N., Hilliard, P., van der Net, J., Engelbert, R.H.H., Petrini, P., van den Berg, M., Manco-Johnson, M.J., Rivard, G.E., Abad, A., Blanchette, V.S.: Validation of a new pediatric joint scoring system from the international hemophilia prophylaxis study group: Validity of the hemophilia joint health score. Arthritis Care & Research **63**(2), 223–230 (2011). https://doi.org/10.1002/acr.20353
5. Gajdosik, R.L., Bohannon, R.W.: Clinical measurement of range of motion: review of goniometry emphasizing reliability and validity. Physical Therapy **67**(12), 1867–1872 (1987). https://doi.org/10.1093/ptj/67.12.1867
6. Gawenda, L., Kreß, M., Werner, S., Schwandt, R.: Digital hemophilia joint health score (2025), iOS Lab project group report, Rheinische Friedrich-Wilhelms-Universität Bonn
7. Hmida, J., Hilberg, T., Ransmann, P., Tomschi, F., Klein, C., Koob, S., Kraft, C.N., Wirtz, D.C., Strauss, A.C.: Most subjectively affected joints in patients with haemophilia: what has changed after 20 years in Germany? Haemophilia **28**(4), 663–670 (2022). https://doi.org/10.1111/hae.14564
8. Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., Chen, K.: RTMPose: Real-time multi-person pose estimation based on MMPose. arXiv preprint arXiv:2303.07399 (2023). https://doi.org/10.48550/arXiv.2303.07399
9. Jiang, T., Xie, X., Li, Y.: RTMW: Real-time multi-person 2D and 3D whole-body pose estimation. arXiv preprint arXiv:2407.08634 (2024). https://doi.org/10.48550/arXiv.2407.08634
10. Kahl, F., Wegner, P., Kapsecker, M., Nissen, L., Faber, J., Jonas, S.M., Reimer, L.M.: A comparative assessment of accuracy in video-based monocular human pose estimation frameworks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2026), frameworks compared against a Vicon reference; MeTRAbs performed best overall
11. Keogh, J.W.L., Cox, A., Anderson, S., Liew, B., Olsen, A., Schram, B., Furness, J.: Reliability and validity of clinically accessible smartphone applications to measure joint range of motion: A systematic review. PLOS ONE **14**(5), e0215806 (2019). https://doi.org/10.1371/journal.pone.0215806
12. Martin, R.L., McPoil, T.G.: Reliability of ankle goniometric measurements: a literature review. Journal of the American Podiatric Medical Association **95**(6), 564–572 (2005). https://doi.org/10.7547/0950564
13. MCCCPTAP: Ankle ROM using a goniometer. Video (2014), youTube: https://www.youtube.com/watch?v=KMUHxsLiUBw. Last accessed: 19.07.2026
14. Needham, L., Evans, M., Cosker, D.P., Wade, L., McGuigan, P.M., Bilzon, J.L., Colyer, S.L.: The accuracy of several pose estimation methods for 3D joint centre localisation. Scientific Reports **11**, 20673 (2021). https://doi.org/10.1038/s41598-021-00212-x

15. ONNX Runtime developers: ONNX Runtime. https://onnxruntime.ai (2021), cross-platform inference engine for on-device model execution
16. Sárándi, I., Linder, T., Arras, K.O., Leibe, B.: MeTRAbs: Metric-scale truncation-robust heatmaps for absolute 3D human pose estimation. IEEE Transactions on Biometrics, Behavior, and Identity Science **3**(1), 16–30 (2021). https://doi.org/10.1109/TBIOM.2020.3037257
17. St-Louis, J., Abad, A., Funk, S., Tilak, M., Classey, S., Zourikian, N., McLaughlin, P., Lobet, S., Hermans, C., Nagel, K., Blanchette, V.S., Feldman, B.M.: The Hemophilia Joint Health Score version 2.1 validation in adult patients study: A multicenter international study. Research and Practice in Thrombosis and Haemostasis **6**(2), e12690 (2022). https://doi.org/10.1002/rth2.12690
18. Zhang, X., Sun, R.: Joint synergy and muscle activity in the motion of the ankle–foot complex. Journal of Experimental Biology **227**(16) (2024). https://doi.org/10.1242/jeb.248069