THA Software System Comparing Outputs from AI to Human Operators

Andrew Hudson [1], Brittany Marshall[1], Scott Leestma[1], Rodrigo Diaz, MD[1]

[1]DePuy Synthes/Johnson & Johnson, Warsaw, IN

ahudson3@its.jnj.com

Disclosures: Andrew Hudson (3A,4-DePuy Synthes/Johnson & Johnson), Brittany Marshall (3A,4-DePuy Synthes/Johnson & Johnson), Scott Leestma (3A,4-DePuy Synthes/Johnson & Johnson), Rodrigo Diaz, MD (3A, 4-DePuy Synthes/Johnson & Johnson)

INTRODUCTION: The use of technology in total hip arthroplasty (THA) has been shown to lead to improved outcomes (Xu et al). Achieving a targeted leg length, offset, cup inclination, and cup anteversion can help to restore the femoral mechanics and reduce the occurrence of dislocations (Lewinnek et al). One technology that provides this information to the surgeon is a non-invasive, image-processing software VELYS Hip Navigation (VHN) (DePuy Synthes). This technology inputs x-ray or fluoroscopy images, guides the user through a manual landmarking and templating process to register the pelvis and hip, and outputs the leg length, offset, inclination, and anteversion angles. Due to the manual process of landmarking to identify bone structures of interest on radiology images, a machine learning (ML) model has been developed to infer and place the default position of the landmark annotation tool onto the radiology images. The Human-in-the-Loop machine learning model, as the type of AI used, expects the user to confirm all landmark positions and modify, as necessary. The purpose of this study is to evaluate if the machine learning model, or AI, inferred landmark points are equivalent to human operators.

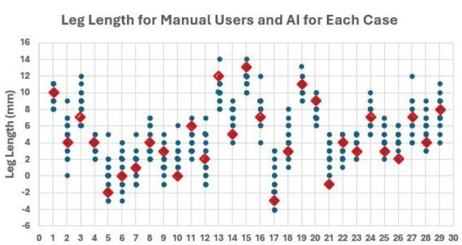
METHODS: The software with the AI landmarks were compared against humans operating the software. Ten operators (engineers and marketers with anatomical and product knowledge) worked through 29 cases, recording the five outputs (inclination, anteversion, leg length, femoral offset, and total offset) across two trials, with the time to execute the workflows recorded in Trial 2. The sample size of 29 cases is derived from a zero-defect acceptance sampling plan in which the sample size of n=29 would fail with 95% probability if the true rate of the undesirable event (AI does not fall within the range of human operator outputs) is 10% or greater. The operator and trial sample size is derived from the principles of a typical gage R&R study. The images used for this study are a random representative sample of deidentified production images that are independent of the training data set; through the deidentification process, no sex data was saved. The 29 cases were then run through with the AI model placing the landmarks, with one human-in-the-loop adjustment allowed. The acceptance criteria for this study is that the AI generated data points for the 5 outputs (leg length, femoral offset, total offset, cup inclination, and cup anteversion) had to be within the range of the 20 human operator data points (10 operators with 2 trials) on all 29 images to pass.

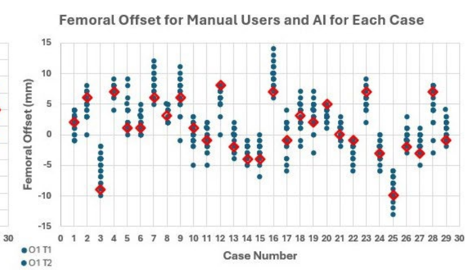
RESULTS: All AI points fell within the range of human operator values, as seen in Figure 1 for the Leg Length, Femoral Offset, and Total Offset outputs, and Figure 2 for Inclination and Anteversion. The time to complete each case, per workflow, was averaged for each user across all cases, then each user's average time was averaged. This time is compared to the average time across all cases using AI. Results of software operation time are shown in Table 1.

DISCUSSION: The acceptance criteria for this study passed with two deviations. Both deviations occurred on AI cases in which the software worked as intended and alerted the user to a difference in landmarking between two of the input images, forcing the user to move a landmark on that landmarking page (the second landmark to be moved) to resolve the error and proceed through the case. The biggest limitation of this study is that the operation of the software took place in a non-clinical environment, instead of the primary environment of the operating room. The operators were also engineers and marketing staff instead of sales representatives and surgeons.

SIGNIFICANCE/CLINICAL RELEVANCE: As AI technology enters the field of orthopaedics, this study highlights that AI can produce outputs that are equivalent to that of human operators, while also saving time, allowing for operators to perform other tasks in the operating room.

REFERNCES: Xu K et al. Computer navigation in total hip arthroplasty: a meta-analysis of randomized controlled trials. Int J Surg. 2014;12(5):528-533. Lewinnek GE et al. Dislocations after total hip-replacement arthroplasties. J Bone Joint Surg Am. 1978;60(2):217-220.





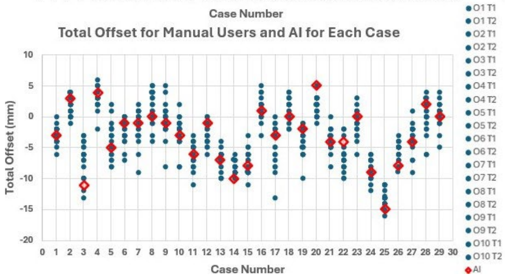


Figure 1: The 20 human outputs (10 Operators, 2 Trials) for Leg Length, Total Offset, and Femoral Offset values in blue dots across all 29 cases (x-axis), and the single AI point as a red transparent diamond for each case, falling inside the range of human operators. Summary table to highlight results.

Output	Result
Leg Length	PASS
Total Offset	PASS
Femoral Offset	PASS
Inclination	PASS
Anteversion	PASS

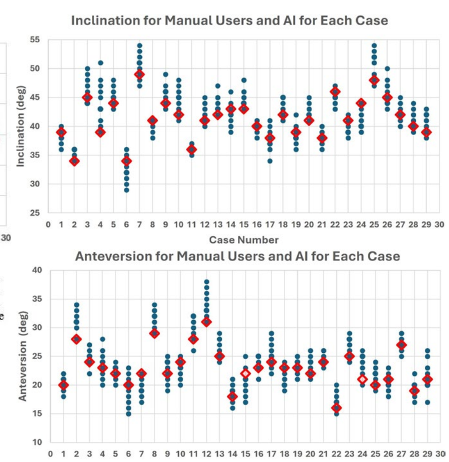


Figure 2: The 20 human outputs (10 Operators, 2 Trials) Inclination and Anteversion in blue dots across all 29 cases (x-axis), and the single AI point as a red transparent diamond for each case, falling inside the range of human operators.

Table 1: The average time for users to complete the manual workflow, averaged across all users and all cases, compared to the average time for the AI workflow, averaged across all cases.

Workflow to Output:	Average Manual Time [across all users and all cases]	Average AI Time [across all cases]
Inclination & Anteversion	1 minute 39 seconds	36 seconds
Leg Length & Offset	3 minutes 41 seconds	1 minute 40 seconds