Evaluating the Reliability, Accuracy, and Agreement of Artificial Intelligence-Generated Angular Measurements in Preoperative and Postoperative Radiographs for Total Knee Arthroplasty
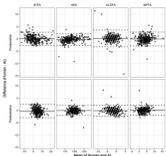
Emma N Horton, David Forrest Scott

INTRODUCTION: Lower limb angular measurements are essential in orthopaedic research and preoperative planning for total knee arthroplasty (TKA) patients. Traditional methods of measuring angles on long-standing radiographs are time-consuming and subject to variability. Artificial intelligence (AI) offers a promising alternative, with emerging tools capable of automated, reproducible measurements. There is limited literature on AI's viability across preoperative and postoperative TKA radiographs within a single patient cohort. The present study evaluates the reliability, accuracy, and agreement of AI-derived angular measurements for hip-knee-ankle (HKA) angle, anatomical tibiofemoral angle (ATFA), mechanical lateral distal femoral angle (mLDFA), and mechanical medial proximal tibial angle (MPTA), compared to manual measurements.
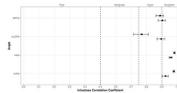
METHODS: A prospective registry was queried for patients who underwent unrestricted kinematic alignment TKA using medial-pivot implants. Preoperative (n = 270) and one-year postoperative (n = 241) long-standing weight-bearing radiographs were independently analyzed by a single highly trained human reader and by a deep-learning AI software. To evaluate the reliability, accuracy, and agreement of AI-generated angular measurements compared to manual measurements, we used intraclass correlation coefficients (ICC), mean absolute error (MAE), root mean square error, and Pearson and Spearman correlation coefficients.

RESULTS: Preoperative ICC values ranged from 0.769 (mLDFA) to 0.983 (ATFA); postoperative ICC values ranged from 0.897 (mLDFA) to 0.959 (HKA). Correlation coefficients were moderate to very strong, with Pearson values of 0.776–0.985 preoperatively and 0.903–0.959 postoperatively. MAE ranged from 0.801° to 1.065° preoperatively and 0.742° to 1.120° postoperatively. AI landmarking errors occurred in 4.07% of preoperative and 1.24% of postoperative radiographs.

DISCUSSION AND CONCLUSION: AI-generated lower limb angular measurements showed good reliability, accuracy, and agreement with manual measurements for mLDFA, HKA, ATFA, and MPTA in preoperative and postoperative primary TKA patients. Our findings support the integration of this AI technology with human oversight for lower limb measurements, particularly in high-volume clinical and research settings.