Radiology: Artificial Intelligence

AI IN BRIEF

Deep Learning for Radiographic Measurement of Femoral Component Subsidence Following Total Hip Arthroplasty

Pouria Rouzrokh, MD, MPH, MHPE • Cody C. Wyles, MD* • Shyam J. Kurian, MD • Taghi Ramazanian, MD • Jason C. Cai, MBBS • Qiao Huang, PhD • Kuan Zhang, PhD • Michael J. Taunton, MD • Hilal Maradit Kremers, MD • Bradley J. Erickson, MD, PhD*

From the Department of Radiology, Radiology Informatics Laboratory (P.R., J.C.C., Q.H., K.Z., B.J.E.), Department of Health Sciences Research (C.C.W., T.R., M.J.T., H.M.K.), Department of Orthopedic Surgery (C.C.W., T.R., M.J.T., H.M.K.), Department of Clinical Anatomy (C.C.W.), and Mayo Clinic Alix School of Medicine (S.J.K.), Mayo Clinic, 200 First St SW, Rochester, MN 55905. Received July 25, 2021; revision requested August 30; revision received March 30, 2022; accepted April 15. **Address correspondence to** H.M.K. (email: *maradit@mayo.edu*).

* P.R. and C.C.W. contributed equally to this work.

Supported by the Mayo Foundation Presidential Fund and the National Institutes of Health (grants R01AR73147 and P30AR76312).

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2022; 4(3):e210206 • https://doi.org/10.1148/ryai.210206 • Content codes: AI MK

Femoral component subsidence following total hip arthroplasty (THA) is a worrisome radiographic finding. This study developed and evaluated a deep learning tool to automatically quantify femoral component subsidence between two serial anteroposterior (AP) hip radiographs. The authors' institutional arthroplasty registry was used to retrospectively identify patients who underwent primary THA from 2000 to 2020. A deep learning dynamic U-Net model was trained to automatically segment femur, implant, and magnification markers on a dataset of 500 randomly selected AP hip radiographs from 386 patients with polished tapered cemented femoral stems. An image processing algorithm was then developed to measure subsidence by automatically annotating reference points on the femur and implant, calibrating that with respect to magnification markers. Algorithm and manual subsidence measurements by two independent orthopedic surgeon reviewers in 135 randomly selected patients were compared. The mean, median, and SD of measurement discrepancy between the automatic and manual measurements were 0.6, 0.3, and 0.7 mm, respectively, and did not demonstrate a systematic tendency between human and machine. Automatic and manual measurements were strongly correlated and showed no evidence of significant differences. In contrast to the manual approach, the deep learning tool needs no user input to perform subsidence measurements.

Supplemental material is available for this article.

Total hip arthroplasty (THA) is a common procedure for patients with end-stage hip diseases (1). Although successful in the vast majority of patients, complications do occur and can result in THA failure. One of the most important roles for serial postoperative radiographs after THA is to detect complications such as component subsidence and other signs of loosening. Prosthesis loosening occurs in approximately 2% of patients and is routinely among the top three reasons for revision surgery in registries across the world (2).

Femoral component subsidence is defined as implant migration in reference to a constant femoral landmark (3). Although subsidence greater than 5 mm is generally considered a sign of loosening and prosthesis failure (4,5), more subtle subsidence is challenging to detect manually (6).

Femoral component subsidence is assessed on two or more consecutive anteroposterior (AP) pelvic or hip radiographs by using various measurement techniques (7). Most techniques measure the distance between a reference point on the femur (eg, greater trochanter) and a reference point on the stem (eg, distal tip), adjust for magnification based on magnification markers, and finally, report subsidence as the difference between the measured distances. Controversies exist about optimal technique, and error levels greater than 1 mm are common (8). Computer applications that interface with digital radiography software are available to facilitate subsidence measurement (8,9). These applications require manual reference point annotations by the user before estimating subsidence, rendering them time-consuming and prone to error.

In line with previous studies that automated complication-related measurements on imaging data for patients with THA (10,11), our study aimed to develop a fully automated deep learning tool, composed of both a U-Net segmentation model and subsequent image processing pipeline, to measure femoral component subsidence on serial AP hip radiographs without any user input.

Materials and Methods

Patient Selection

Following institutional review board approval of this Health Insurance Portability and Accountability Act–compliant study, with the waiver of informed consent, we retrospectively evaluated our institution's arthroplasty registry to identify 386 patients who underwent consecutive primary THA with polished tapered cemented femoral stems (Exeter; Stryker). The clinical and radiologic success of these stems is widely recognized, and their frequency of use in different institutes could be as high as 70%–90% (12). However, we intentionally selected polished tapered cemented stems as a proof-of-concept for our study, as they are uniquely designed to undergo

Abbreviations

AP = anteroposterior, GUI = graphical user interface, THA = total hip arthroplasty

Summary

Authors developed a fully automatic deep learning tool to measure subtle femoral component subsidence on anteroposterior radiographs without any user input.

Key Points

- By using deep learning U-Net models and image processing, a fully automatic tool was developed to measure subtle femoral component subsidence on anteroposterior hip radiographs without any user input.
- The algorithm achieved a mean absolute error of 0.6 mm (SD = 0.7) and median error of 0.3 mm (IQR = 0.9) compared with rater-averaged manual annotations, with no evidence of a difference in measurements between groups (P = .89).
- The tool has both clinical and research applications and may improve total hip arthroplasty postoperative surveillance and result in earlier detection of impending component failure.

Keywords

Total Hip Arthroplasty, Femoral Component Subsidence, Artificial Intelligence, Deep Learning, Semantic Segmentation, Hip, Joints

subtle (1–2 mm) controlled subsidence during the first 1–2 years after surgery. Selected patients underwent surgery between 2000 and 2020 and included 114 men (mean age, 77 years; range, 53–98 years) and 272 women (mean age, 76 years; range, 38–93 years).

Development of Deep Learning Segmentation Model

We leveraged a dynamic U-Net model to automatically segment the femur, implant, and magnification markers on 2048 × 2048-pixel input AP hip radiographs (Fig 1A, 1B). The model had an EfficientNetB0 architecture as its encoder, with initial weights pooled from a similar model pretrained on the ImageNet dataset (13–15). The decoder used squeeze-and-excitation attention and had its weights started using the Hu distribution (16,17). The source code for the model is available at *https://github.com/qubvel/segmentation_models.pytorch*.

From the 386 patients, 700 radiographs were obtained and split into 500 (71.4%) training, 100 (14.3%) validation, and 100 (14.3%) test images on the patient level. Training details are provided in Appendix E1 (supplement).

Image Processing Algorithm

An image processing algorithm was then developed to receive two input masks generated from subsequent AP hip radiographs from a patient and calculate the "standard femur-stem distance" on each of the two radiographs by *(a)* finding the pixel coordinates for the tip of the stem (stem point) and the coordinates for the most superior point on the greater trochanter (femur point); *(b)* measuring the pixel distance between the stem point and the femur point; and *(c)* standardizing that distance based on the angle of femoral axis and the distance between the magnification markers. Stem subsidence between the two radiographs was finally calculated by subtracting the measured standard femur-stem distance between images (Fig 1C, 1D).

Model Evaluation and Statistical Analysis

To evaluate model performance, automatic measurements were compared with the average manual measurements of two orthopedic surgeons in 135 patients; data from these patients had not been used to develop the segmentation or image processing algorithms. Comparisons were assessed using descriptive statistics, the Spearman correlation test, the Lin concordance correlation coefficient, the Bland-Altman plot, and the two-tailed t test. To showcase performance and increase clinical or research user applicability, we also deployed the tool through a stand-alone graphical user interface (GUI). All tests were done using Python Statistics (version 3.9) and SciPy (version 1.7.3) packages. Significance level of .05 was used.

Please refer to Appendix E1 (supplement) for detailed methods.

Results

Model Performance

Our deep learning algorithm achieved the following Dice similarity coefficients on the test set: 0.97 for detecting femur, 0.98 for detecting implant, and 0.94 for detecting magnification markers (Fig 2).

Comparison between Automatic and Manual Measurements

The Table summarizes descriptive statistics for the manual measurements, automatic measurements, and measurement errors of subsidence in 135 patients with polished tapered cemented femoral stems. Among these patients, 96 had manually documented subsidence of at least 0.1 mm (range, 0.1–14.0 mm). Rater-averaged manual subsidence measurements had an average of 2.7 mm (95% CI: 2.4, 3.3), and automatic measurements had an average of 2.8 mm (95% CI: 2.3, 3.3). Automated measurements were greater than 0 mm in seven of nine patients who had rater-averaged manual measurements of greater than 0 mm and less than 1 mm. The algorithm achieved a mean absolute error of 0.6 mm (SD = 0.7) and median absolute error of 0.3 mm (IQR = 0.9) compared with rater-averaged manual annotations. The mean absolute and median absolute errors were 21% and 19% of the mean and median of the rater-averaged manual measurements, respectively. Automatic measurements and rater-averaged manual annotations had a concordance correlation coefficient of 0.95 and were strongly correlated (Spearman ρ = 0.96, $P < .001$), with no evidence of differences between measurement groups (P = .89). The 95% limits of agreement between the automated and rater-averaged manual measurements were less than 2 mm in the Bland-Altman plot. Measurement discrepancies larger than 2 mm occurred in three (2.2%) patients, all of whom had severe heterotopic ossification around the proximal end of the femur that developed between radiographic time points, leading to failure of the deep learning model to accurately segment the femur.

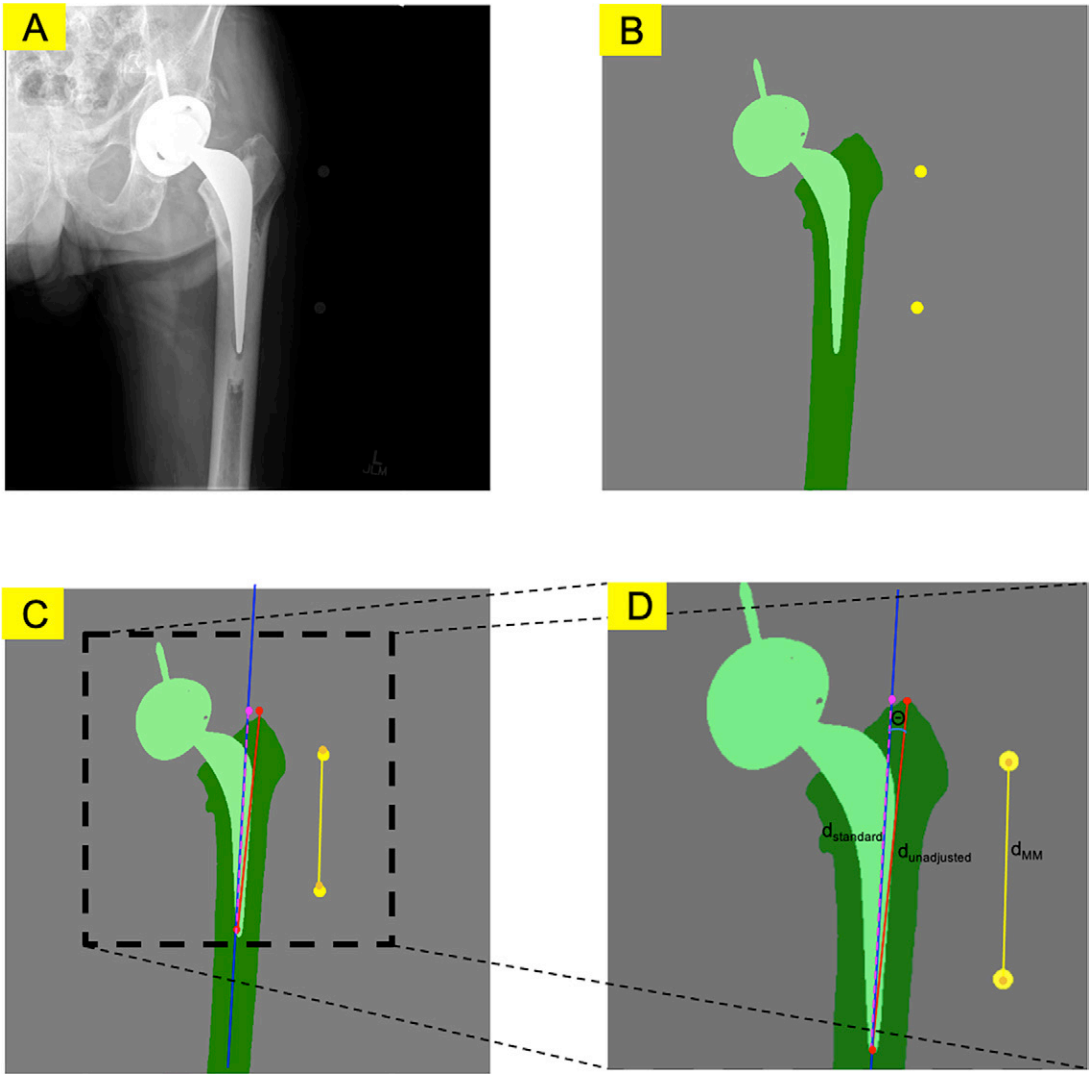

Figure 1: Schematic workflow of the deep learning subsidence measurement tool. **(A)** Standard anteroposterior (AP) hip image fed to the tool. **(B)** Automatic segmented mask generated by the deep learning segmentation model. **(C)** Image processing algorithm used to measure the intermarker line (yellow line), femur-stem line (red line), and femoral axis (blue line). **(D)** Zoomed-in version of **C**. Suppose $d_{unadjusted}$ denotes the length of the femur-stem line, Θ denotes the angle between the femur-stem line and the femoral axis, and d_{MM} denotes the length of the intermarker line. The standardized femur-stem distance (the length of the dashed purple line) will be calculated as follows: $d_{standard} = d_{unadjusted} \times \cos \Theta \times 100/d_{MM}$. The difference between the $d_{standard}$ values measured on two successive AP hip radiographs will give the stem subsidence value.

Graphical User Interface

Figure 3 demonstrates the GUI we developed that can run without special hardware or software needed for deep learning. A demo video of the GUI application is provided (Movie).

Please refer to Appendix E1 (supplement) for detailed results, including an investigation of challenging and failure cases.

Discussion

Measuring subsidence on serial radiographs is tedious, time-consuming, and prone to error. This study developed a deep learning tool that can automatically measure subsidence given two input radiographs without any requirement for user annotations.

The workflow of our study consisted of three successive steps. First, we developed a deep learning segmentation model that can segment the femur, prosthesis, and magnification markers from two input postoperative radiographs. Second, an image processing algorithm was developed that could identify specific reference points on both radiographs and rely on those points to measure the level of stem subsidence between radiographs. Finally, we implemented both segmentation and image processing algorithms as a stand-alone GUI. This end software is user-friendly and fast, demonstrating a proof-of-concept for the capabilities that a deep learning tool can have in measuring postoperative complications in patients who underwent THA. To our knowledge, there are currently no available tools to automatically quantify implant subsidence. All current methods rely on cumbersome built-in features present in nearly every radiography software, such as that used for the reference standard annotations in this study.

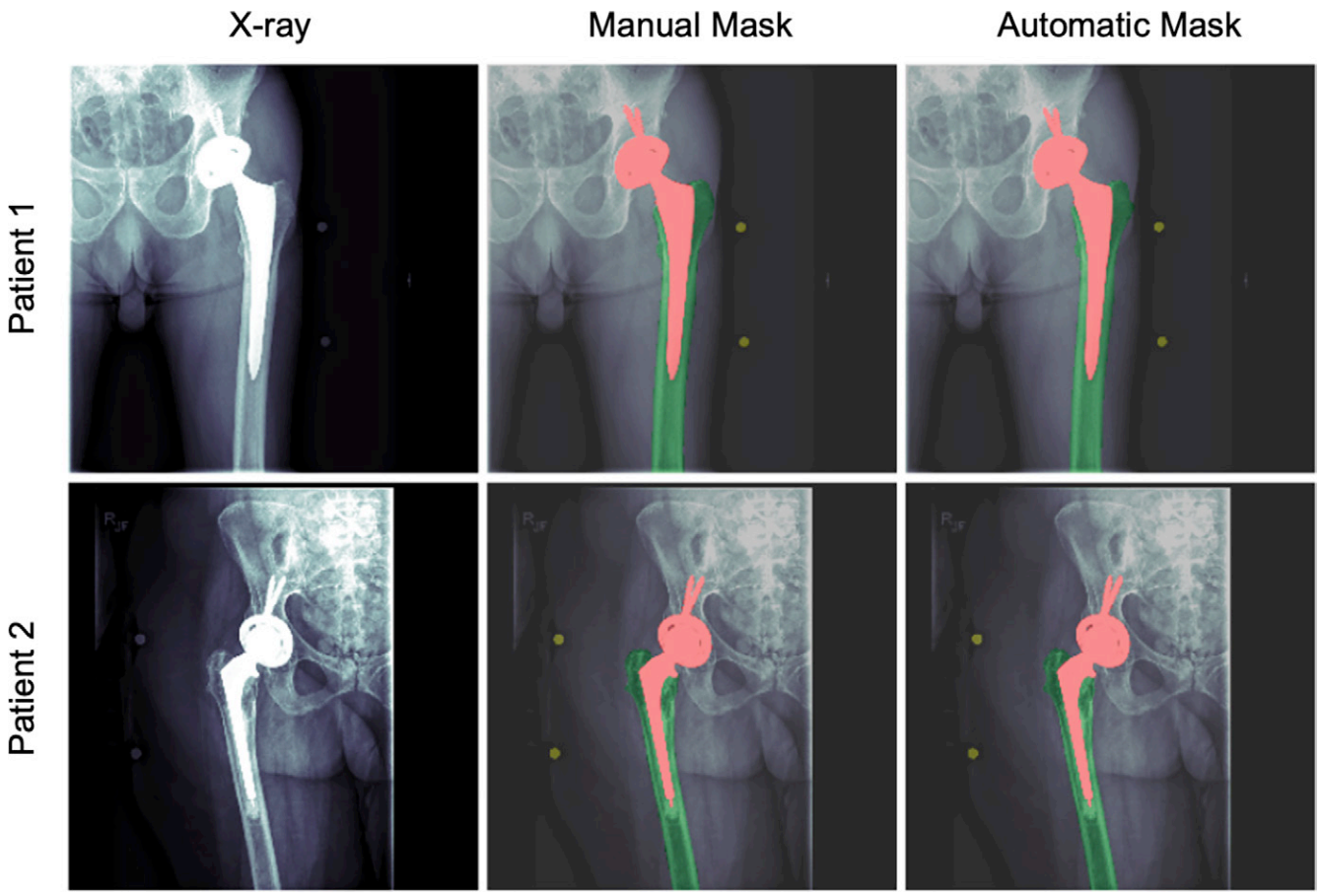

Figure 2: Segmentation masks generated from the semantic segmentation model for two anteroposterior hip radiographs present in the test set. The middle column shows the manually segmented masks (ground truths) for comparison purposes. Top row: Radiographs in a 79-year-old woman who underwent total hip arthroplasty and was followed up over a period of 9 years. Bottom row: Radiographs in an 80-year-old woman who underwent the same operation and was followed up over a period of 12 years.

Descriptive Statistics for the Rater-averaged Manual Measurements, Automatic Measurements, and Measurement Errors of the Reported Subsidence in 135 Patients with Cemented Polished Tapered Stems

Statistic	Manual Annotation	Automatic Annotation	Absolute Measurement Error
Minimum (mm)	0.0	0.0	0.0
Maximum (mm)	14.0	13.4	5.9
Zero counts	39	42	38
Mean ± SD (mm)	2.7 ± 3.1	2.8 ± 3.1	0.6 ± 0.7
Median (mm)*	1.8 (4.3)	2.2 (4.6)	0.3 (0.9)

Note.—Zero counts denote the number of cases with subsidence level = 0 mm.
* Data in parentheses are IQRs.

This study is not without limitations. First, although the mean absolute error was less than 1 mm, errors can still result in false-positive or false-negative findings. Second, we only verified performance on polished tapered cemented femoral stems, and further evaluation of other stem types is necessary to ensure generalizability. Although we provided examples of how our tool can still work properly when applied to other stems, formal validation is mandatory. Third, nonstandard AP hip radiographs may result in segmentation model failure and subsequent measurement errors. Examples of challenging radiographs for the tool include those with lower resolution, missing magnification markers or markers with a substantially different shape than what was routinely used in our institute, unusual hardware, fractures, and development of severe heterotopic ossification (Appendix E1 [supplement]).

In conclusion, we developed a fully automatic deep learning tool to measure femoral component subsidence on AP hip radiographs. Performance metrics indicated highly comparable

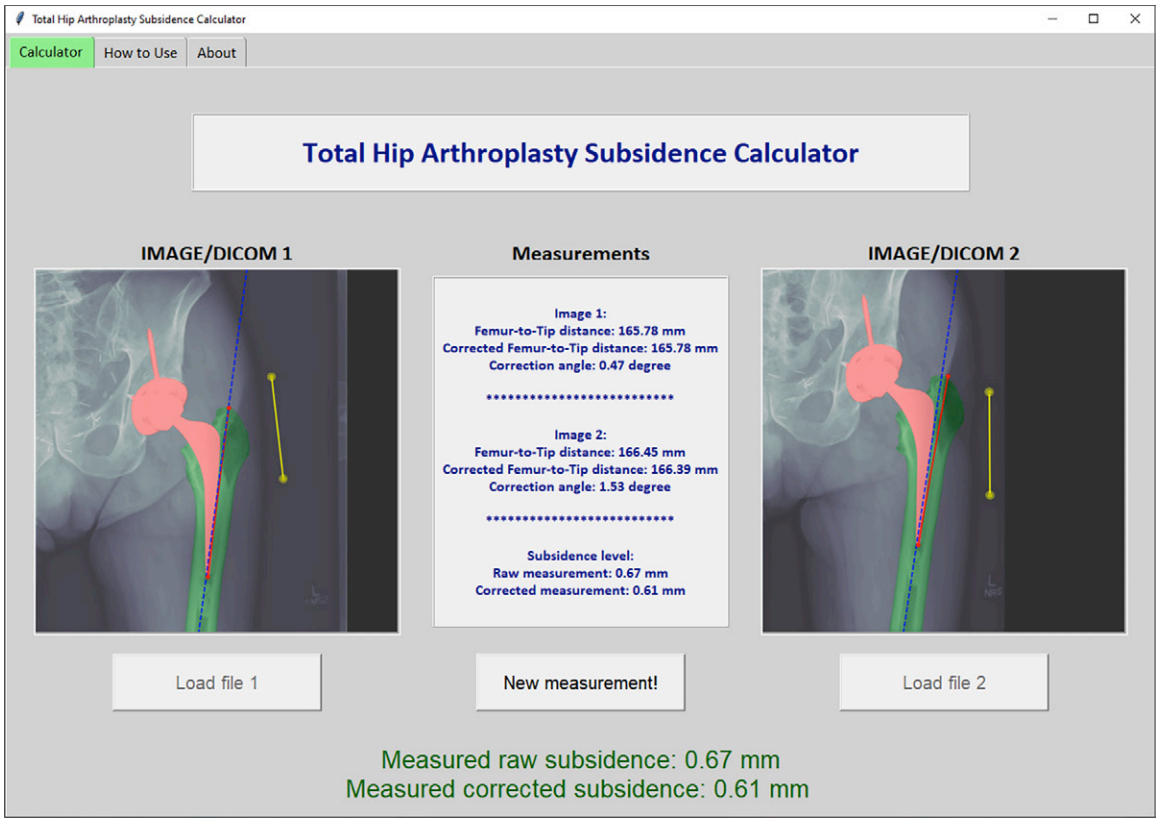

Figure 3: A screenshot of the stand-alone graphical user interface (GUI) we developed for our deep learning subsidence measurement tool. Upon selecting the input radiographs, the user can press the measurement button, and the tool will automatically measure the subsidence and show annotated images in seconds. See a short demo clip (Movie) of this GUI. DICOM = Digital Imaging and Communications in Medicine.

performance to human annotation, with very infrequent discrepancies of greater than 2 mm. If trained and validated on larger datasets with more heterogeneous stem types, our tool may hold potential as a screening tool in clinical and research settings to identify patients at risk for complications due to stem subsidence. We are currently building a carefully crafted dataset of all stem types from multiple institutes to evaluate such potential in our tool and further improve its generalizability across all stem types.

Author contributions: Guarantors of integrity of entire study, **P.R., C.C.W., H.M.K., B.J.E.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **P.R., C.C.W., S.J.K., T.R., J.C.C., Q.H., K.Z., H.M.K.**; clinical studies, **C.C.W., T.R., M.J.T., H.M.K.**; experimental studies, **P.R., J.C.C., K.Z., H.M.K.**; statistical analysis, **P.R., C.C.W., J.C.C., Q.H., H.M.K.**; and manuscript editing, all authors

Disclosures of conflicts of interest: P.R. No relevant relationships. **C.C.W.** No relevant relationships. **S.J.K.** No relevant relationships. **T.R.** No relevant relationships. **J.C.C.** No relevant relationships. **Q.H.** No relevant relationships. **K.Z.** No relevant relationships. **M.J.T.** Royalties/licenses from DJO Surgical; consulting fees from DJO Surgical; leadership role with *Journal of Arthroplasty*. **H.M.K.** Mayo Foundation Presidential Fund NIH grants R01AR73147 and P30AR76312. **B.J.E.** Chair of SIIM Research Committee; consultant to the editor for *Radiology: Artificial Intelligence*.

References

1. Karachalios T, Komnos G, Koutalos A. Total hip arthroplasty: Survival and modes of failure. EFORT Open Rev 2018;3(5):232–239.
2. Springer BD, Fehring TK, Griffin WL, Odum SM, Masonis JL. Why revision total hip arthroplasty fails. Clin Orthop Relat Res 2009;467(1):166–173.
3. Ries C, Boese CK, Dietrich F, Miehlke W, Heisel C. Femoral stem subsidence in cementless total hip arthroplasty: a retrospective single-centre study. Int Orthop 2019;43(2):307–314.
4. Kärrholm J, Borssén B, Löwenhielm G, Snorrason F. Does early micromotion of femoral stem prostheses matter? 4-7-year stereoradiographic follow-up of 84 cemented prostheses. J Bone Joint Surg Br 1994;76(6):912–917.
5. Sudhahar TA, Morapudi S, Branes K. Evaluation Of Subsidence Between Collarless And Collared Corail Femoral Cement Less Total Hip Replacement. J Orthop 2009;6(2):e3. http://www.jortho.org/2009/6/2/e3/index.htm.
6. Al-Najjim M, Khattak U, Sim J, Chambers I. Differences in subsidence rate between alternative designs of a commonly used uncemented femoral stem. J Orthop 2016;13(4):322–326.
7. Ilchmann T, Eingartner C, Heger K, Weise K. Femoral subsidence assessment after hip replacement: an experimental study. Ups J Med Sci 2006;111(3):361–369.
8. Schütz U, Decking J, Decking R, Puhl W. Assessment of femoral component migration in total hip arthroplasty: digital measurements compared to RSA. Acta Orthop Belg 2005;71(1):65–75.
9. Krismer M, Biedermann R, Stöckl B, Fischer M, Bauer R, Haid C. The prediction of failure of the stem in THR by measurement of early migration using EBRA-FCA. J Bone Joint Surg Br 1999;81(2):273–280.
10. Borjali A, Chen AF, Muratoglu OK, Morid MA, Varadarajan KM. Detecting mechanical loosening of total hip replacement implant from plain radiograph using deep convolutional neural network. arXiv preprint arXiv:1912.00943. https://arxiv.org/abs/1912.00943. Posted December 2, 2019. Accessed June 2021.
11. Rouzrokh P, Wyles CC, Philbrick KA, et al. A Deep Learning Tool for Automated Radiographic Measurement of Acetabular Component Inclination and Version After Total Hip Arthroplasty. J Arthroplasty 2021;36(7):2510–2517.e6.
12. Yates PJ, Burston BJ, Whitley E, Bannister GC. Collarless polished tapered stem: clinical and radiological results at a minimum of ten years' follow-up. J Bone Joint Surg Br 2008;90(1):16–22.
13. Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv preprint arXiv:1905.11946. https://arxiv.org/abs/1905.11946. Posted May 28, 2019. Accessed June 2021.

14. Yakubovskiy P. Segmentation Models Pytorch. GitHub Repos. GitHub; 2020. https://github.com/qubvel/segmentation_models.pytorch. Accessed June 2021.
15. Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: a Large-Scale Hierarchical Image Database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Conference Location, June 20–25, 2009. Piscataway, NJ: IEEE, 2009; 248–255.
16. He K, Zhang X, Ren S, Sun J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. arXiv preprint arXiv:1502.01852. https://arxiv.org/abs/1502.01852. Posted February 6, 2015. Accessed June 2021.
17. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-Excitation Networks. 2019. arXiv preprint arXiv:1709.01507. https://arxiv.org/abs/1709.01507. Posted September 5, 2017. Accessed June 2021.