Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Routine perioperative radiographs for early risk stratification of femoral head osteonecrosis after femoral neck fracture fixation: a retrospective cohort study

Xiaoyun Guo
Tianjin University
Zhe Han
Tianjin Hospital
Zonghan Zhang
Tianjin University
Xiukun Zhong
Tianjin University
Qiang Dong
Tianjin Hospital
Kai Song
ksong@tju.edu.cn
Tianjin University
Xinlong Ma
Tianjin Hospital

Research Article

Keywords: Femoral neck fracture, femoral head osteonecrosis, perioperative radiographs, deep learning, clinical prediction model, risk stratification.

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11316293/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Routine perioperative radiographs for early risk stratification of femoral head osteonecrosis after femoral neck fracture fixation: a retrospective cohort study

Xiao-yun Guo[1]†, Zhe Han[2]†, Zong-han Zhang[1], Xiu-kun Zhong[1], Qiang Dong[2]*, Kai Song[3],[4],[5]*, Xin-long Ma[6]

[1] School of Chemical Engineering and Technology, Tianjin University, Tianjin 300072, China. [2] Department of Hip Trauma, Tianjin Hospital, Tianjin 300211, China.

[3] School of Synthetic Biology and Biomanufacturing, Tianjin University, Tianjin 300072, China.

[4] Frontiers Science Center for Synthetic Biology, Tianjin University, Tianjin 300072, China.

[5] Key Laboratory of Systems Bioengineering (Ministry of Education), Tianjin University, Tianjin 300072, China.

[6] Department of Orthopedics, Tianjin Hospital, Tianjin 300211, China.

† Xiao-yun Guo and Zhe Han contributed equally. * Correspondence: Qiang Dong (dongqiangtj@yeah.net); Kai Song (ksong@tju.edu.cn).

Abstract

Background: Femoral head osteonecrosis (ONFH) after internal fixation of femoral neck fracture is difficult to predict before definitive radiographic changes appear. Preoperative and immediate postoperative radiographs may contain prognostic information not fully captured by routine patient characteristics and visual radiographic grades. We compared a four-view radiographic model with a six-variable conventional baseline and explored the value of combining clinical and imaging information.

Methods: This single-centre retrospective cohort included 877 adults treated with closed reduction and cannulated compression screw fixation (CCS), of whom 222 developed ONFH. The development set comprised 727 patients, and 150 formed a fixed internal test set. The conventional baseline was a logistic regression model using age, sex, body mass index, time from admission to surgery, fracture displacement and reduction quality. The four-view model used preoperative and immediate postoperative anteroposterior and lateral radiographs. The primary measure was the area under the receiver operating characteristic curve (AUROC); the area under the precision–recall curve (AUPRC), Brier score and calibration were also assessed. Uncertainty and paired differences were estimated by patient-level bootstrap resampling; the preoperative–postoperative biplanar comparison used DeLong's test.

Results: In the test set, the four-view model had an AUROC of 0.864 versus 0.704 for the conventional baseline, a paired difference of 0.160 (95% confidence interval [CI] 0.068–0.256). The corresponding AUPRCs were 0.759 and 0.432, a paired difference of 0.327 (95% CI 0.153–0.461). Brier scores were 0.148 and 0.171, respectively. The uncalibrated four-view probabilities showed calibration error. An exploratory model combining the six variables with the four-view output had an AUROC of 0.863 and showed no additional discrimination.

Conclusions: A model using four routine radiographic views obtained before and immediately after fixation discriminated subsequent ONFH better than the six-variable conventional baseline, whereas fusion did not improve discrimination. External validation and calibration assessment are needed before using predicted probabilities to guide individual follow-up decisions. Unlike approaches based on later postoperative images, this study assesses prognostic information available before and immediately after fixation, before radiographic signs of ONFH emerge.

Keywords

Femoral neck fracture; femoral head osteonecrosis; perioperative radiographs; deep learning; clinical prediction model; risk stratification.

Background

Femoral neck fracture (FNF) is a common hip fracture and remains an important cause of disability and mortality worldwide [1]. Internal fixation is an important head-preserving treatment, particularly for younger patients and selected patients with nondisplaced fractures [2]. Nevertheless, femoral head osteonecrosis (ONFH) is a serious complication after fixation. A meta-analysis reported pooled ONFH incidences of 20.7% after displaced fractures and 4.7% after nondisplaced fractures [3]. ONFH may cause pain, femoral head collapse and impaired hip function, ultimately requiring arthroplasty [2, 4]. Earlier identification of patients at high risk could therefore inform risk-stratified follow-up.

Definitive radiographic signs of ONFH often appear late, making perioperative identification of high-risk patients difficult [4, 5]. Preoperative and immediate postoperative radiographs are routinely acquired in the care of FNF. Although ONFH is not yet visible, these images document fracture morphology, postoperative osseous alignment and implant position. Previous machine-learning studies based on structured clinical data and visually assessed radiographic variables have identified fracture displacement or Garden classification and reduction quality as important predictors of postoperative ONFH [6, 7]. However, summarising a complex fracture and its reduction in a small number of categories may not capture details of displacement, the location of cortical defects or the relationship between implants and bone. A meta-analysis of Garden I–II fractures found that preoperative posterior tilt of at least 20° was associated with an increased

risk of fixation failure [8], illustrating clinically relevant morphological variation within the same classification. Visual grading may also be affected by interobserver variability [9].

Deep learning enables direct analysis of the complete radiograph [10, 11]. Previous radiographic deep-learning studies of ONFH have mainly focused on detecting or staging established disease [10, 11]. One study combined patient data with postoperative radiographs to predict ONFH after FNF fixation, but used images obtained approximately six months after surgery [12]. The prognostic performance of radiographs available before and immediately after surgery, relative to conventional patient characteristics and visual radiographic assessments, therefore remains to be clarified. It is also uncertain whether multiple radiographic views improve early risk stratification.

We developed a four-view deep-learning model using preoperative and immediate postoperative anteroposterior (AP) and lateral radiographs from patients treated with CCS. We compared its discrimination and calibration for subsequent ONFH with a model based on routine patient characteristics and visual radiographic assessments. We also explored clinical–imaging fusion, alternative image combinations and image-based interpretability analyses.

Methods

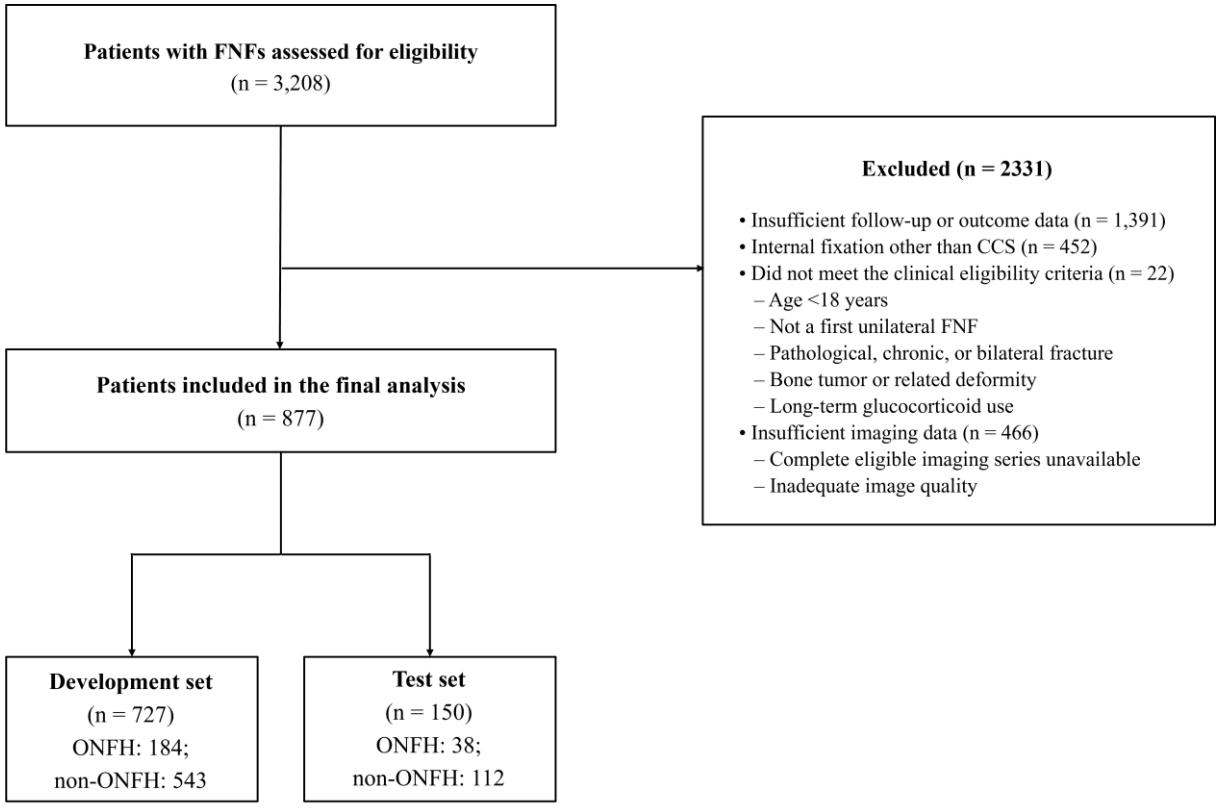
Study population and outcome definition

This single-centre retrospective cohort screened 3,208 patients who underwent closed reduction and internal fixation for FNF at Tianjin Hospital between 2014 and 2024. Adults were eligible if they had a first unilateral FNF, underwent CCS and had preoperative and immediate postoperative AP and lateral radiographs. Patients with ONFH were required to have follow-up until diagnosis; those without ONFH required at least 24 months of follow-up. We excluded pathological, old or bilateral FNFs; bone tumours or relevant deformities; and long-term

glucocorticoid use. The final cohort comprised 877 patients, including 222 with and 655 without ONFH (Fig. 1; Table 1). The study was approved by the Ethics Committee of Tianjin Hospital (approval no. 2024-081).

Fig. 1. Patient selection and dataset allocation

Of 3,208 patients assessed for eligibility, 2,331 were excluded and 877 were analysed: 727 in the development set (184 with and 543 without ONFH) and 150 in the fixed internal test set (38 with and 112 without ONFH). CCS, cannulated compression screw fixation; FNF, femoral neck fracture; ONFH, femoral head osteonecrosis.



The ONFH outcome was ascertained by the treating orthopaedic surgeon from routine follow-up records. A positive outcome was defined as radiographic ONFH of stage II or higher according to the 2019 revised Association Research Circulation Osseous (ARCO) system [13], or hip

arthroplasty performed for ONFH. Patients classified as not having ONFH had at least 24 months of follow-up, no ONFH-related abnormality on their last radiograph and no ONFH diagnosis in the medical record.

Conventional clinical and visually assessed radiographic variables

The conventional baseline included six variables: age, sex, body mass index (BMI), time from admission to surgery, preoperative fracture displacement and postoperative reduction quality. Garden types I–II were classified as nondisplaced and types III–IV as displaced. Reduction quality was retrieved from the clinical database as a binary assessment ("good" or "poor") based on routine postoperative radiographic evaluation and was reviewed by a senior orthopaedic surgeon. The original database category was retained without retrospective reclassification using a new quantitative threshold. Age, BMI and time to surgery were treated as continuous variables; sex, displacement and reduction quality were binary variables.

Radiograph acquisition and preprocessing

Four radiographs were analysed for each patient: preoperative AP, preoperative lateral, immediate postoperative AP and immediate postoperative lateral views. Images were acquired with MOBILETT XP Digital (Siemens Healthineers), Discovery XR656 (GE HealthCare) or DRX-Evolution (Carestream Health) systems. On each image, the femoral head centre and radius, r, were determined from its contour, and the proximal femoral shaft defined the shaft axis and distal direction. A square region of interest (ROI) with side length $5r$ was centred at a point one radius distal to the femoral head centre along the shaft axis, with one pair of sides parallel to that axis. The square was extracted after rotation alignment. Left-hip images were horizontally flipped to standardise orientation.

The cropped images underwent automatic grayscale curve adjustment followed by dark-channel-prior enhancement [14], and were then resized to 224 × 224 pixels. Representative images before and after enhancement are shown in Supplementary Fig. S1. After scaling pixel intensities to [0, 1], images were standardised using a fixed mean of 0.266 and standard deviation of 0.219, both calculated from original postoperative and virtual implant removal (VRI) images of the 727 development-set patients. Random augmentation was applied only during training; validation and testing used deterministic preprocessing. Details of the curve adjustment, dark-channel-prior procedure and augmentation parameters are given in Additional file 1.

Data partitioning and model development

Data were randomly partitioned at the patient level, stratified by ONFH outcome and balanced for fracture displacement. The development set comprised 727 patients (184 with and 543 without ONFH); 150 patients formed the fixed internal test set (38 with and 112 without ONFH).

Conventional baseline. Using the same five development-set folds and the same six variables, we compared unpenalised logistic regression, ridge-penalised logistic regression, the least absolute shrinkage and selection operator (LASSO), elastic net, random forest and histogram-based gradient boosting. Within each outer fold, continuous variables were standardised using the means and standard deviations estimated from the training portion. Algorithms requiring hyperparameter tuning used a further three-fold cross-validation within the training data. Pooled out-of-fold predictions were compared primarily by the area under the receiver operating characteristic curve (AUROC), with the area under the precision–recall curve (AUPRC), Brier score and calibration also considered. Unpenalised logistic regression was selected as the principal conventional baseline because its development-set performance was similar to that of the other methods and the model was simple and interpretable. It was refitted on all 727

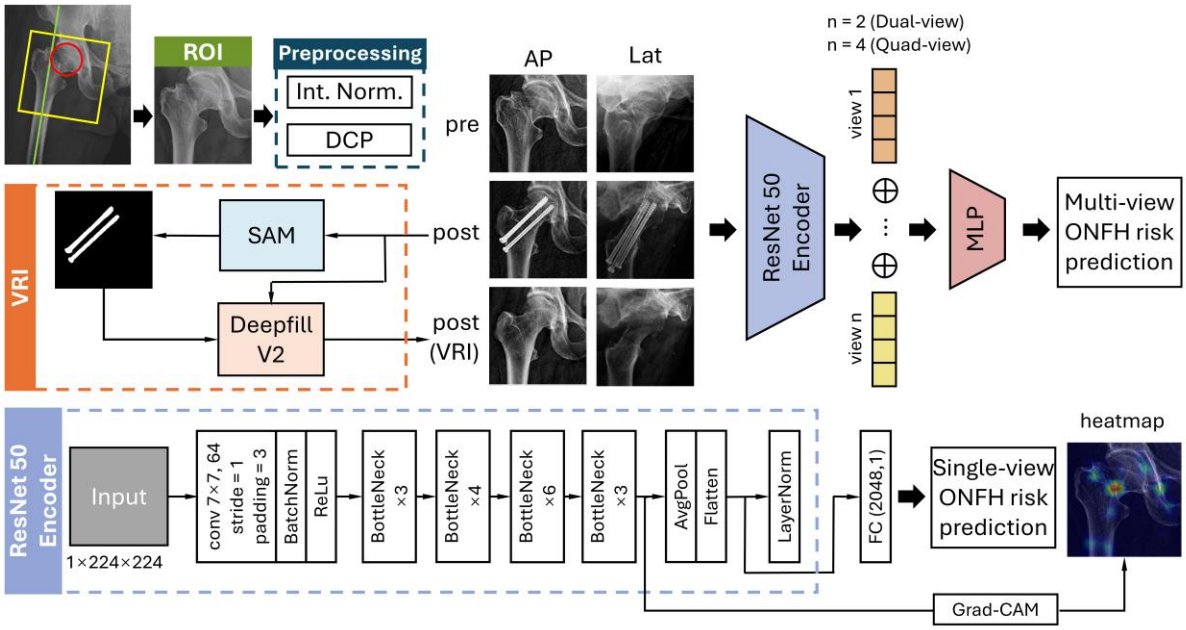
development-set patients; the fixed internal test set was transformed using only the standardisation parameters estimated from the complete development set before subsequent evaluation and comparison.

Radiographic models. A modified ResNet-50 initialised with ImageNet-1K weights extracted image features [15]. The first convolution was adapted for single-channel input, its initial stride was reduced from 2 to 1, and the initial max-pooling layer was removed to limit early loss of fine osseous detail. Multi-view models used shared-weight branches to extract features from each image; concatenated features were passed to a two-layer classification head to output an ONFH risk probability. We trained single-view models, biplanar models at the same time point, preoperative–postoperative pairs in the same projection and a four-view model. Figure 2 illustrates preprocessing, shared feature extraction and multi-view fusion. Each model was trained in five development-set folds, with the checkpoint selected by validation loss within each fold. In the test set, the five fold-specific predicted probabilities were averaged for each patient.

Fig. 2. Image preprocessing and multi-view prediction workflow

The diagram shows cropping of the femoral head–neck region, image preprocessing, feature extraction with a shared-weight ResNet-50 and multi-view feature fusion. At upper left, the red circle outlines the femoral head, the green line indicates the femoral shaft axis, and the yellow square indicates the initial ROI aligned with that axis. Image preprocessing included grayscale curve adjustment, dark-channel-prior (DCP) enhancement, resizing to 224 × 224 pixels, intensity scaling to [0, 1] and standardisation, as detailed in the Methods. Preoperative and immediate postoperative anteroposterior (AP) and lateral (Lat) images are passed through the feature extractor; their features are concatenated and passed to a classification head to predict femoral head osteonecrosis (ONFH) risk. The symbol *n* denotes the number of input views, not patients.

Virtual implant removal (VRI) and single-view Grad-CAM are secondary analyses; VRI images were used to train corresponding models separately. The example heatmap illustrates the method and is not a test-set result.



Interpretability and virtual implant removal analyses

The four-view model processes each image through a shared-weight feature extractor and fuses the resulting features. Attribution of the final prediction to one branch may depend on information from the other views and is therefore difficult to interpret. To examine the regions used when each image predicts the outcome on its own, we generated gradient-weighted class activation mapping (Grad-CAM) heatmaps from separately trained single-view models [16]. These heatmaps were not intended to explain the fused four-view decision directly. Each patient-level heatmap was normalised to [0, 1], and group-level maps were generated by taking the pixel-wise mean separately for patients with and without ONFH within each input configuration.

Virtual implant removal was a secondary analysis of the contribution of implant-associated image regions to model performance. Interactive prompts guided the Segment Anything Model to generate implant masks [17], which were converted to binary masks defining the regions to be reconstructed. DeepFillv2 was then used to reconstruct these regions [18]. During DeepFillv2 training, real implant masks from a segmentation library were used to simulate missing regions resembling the implant shapes. All reconstructed images were reviewed by a senior orthopaedic surgeon. Original postoperative images and VRI images were used to train their respective ONFH prediction models separately. Performance was compared in the same fixed test set using patient-level paired analysis. VRI images did not replace the original postoperative images in the primary analysis; the VRI input configurations are listed in Table 2.

Clinical–imaging fusion model

To explore whether combining routine variables with four-view image information improved prediction, we fitted an unpenalised logistic regression model with seven inputs: the six clinical variables and the four-view model's predicted ONFH probability. The image probability was logit-transformed before inclusion as a continuous predictor. In the development set, each patient's image probability came from the four-view model for the validation fold to which that patient belonged; the fusion model was then fitted in all 727 patients. For the fixed internal test set, the image probability was the average of the five fold-specific model outputs and was combined with the clinical variables to generate the fusion prediction. AUROC, AUPRC, Brier score and calibration were compared among the fusion model, the clinical baseline alone and the four-view model alone in the same 150 patients.

Model evaluation and statistical analysis

AUROC was the primary performance measure; AUPRC, Brier score and calibration were secondary measures. Because no established clinical risk threshold guides follow-up, threshold-dependent classification measures were not reported. Except for the paired DeLong comparison of the postoperative and preoperative biplanar models, 95% confidence intervals (CIs) and between-model comparisons were estimated using 10,000 patient-level paired bootstrap resamples. Overall analyses used outcome-stratified resampling, and between-model comparisons were interpreted descriptively. For grouped calibration, the test set was divided into five equal-sized groups according to predicted risk. Calibration was evaluated using Wilson CIs for observed outcome frequencies, Brier scores, and recalibration intercepts and slopes.

Use of a large language model

OpenAI Codex was used to assist with translation and language editing. It was not used to generate or analyse study data. All scientific content, numerical results, interpretations and references were reviewed and verified by the authors, who take full responsibility for the final manuscript.

Results

Study population and baseline characteristics

Of 3,208 patients screened between 2014 and 2024, 877 were included; 222 (25.3%) developed ONFH and 655 (74.7%) did not. The development set comprised 727 patients (184 with ONFH) and the fixed internal test set 150 patients (38 with ONFH; Fig. 1). Displaced fractures occurred in 410 patients, 158 of whom developed ONFH; 64 of 467 patients with nondisplaced fractures developed ONFH. Reduction quality was rated poor in 92 patients, including 52 who developed

ONFH, and good in 785 patients, including 170 who developed ONFH. Other characteristics are shown in Table 1.

Table 1. Characteristics of the study population

Characteristic	All patients (n = 877)	ONFH (n = 222)	No ONFH (n = 655)	*p* value
Age, years	53.830 ± 12.390	53.590 ± 10.430	53.920 ± 12.990	0.730
Age group				0.003
<50 years	244 (27.8%)	56 (25.2%)	188 (28.7%)	
50–59 years	338 (38.5%)	108 (48.6%)	230 (35.1%)	
60–69 years	231 (26.3%)	48 (21.6%)	183 (27.9%)	
≥70 years	64 (7.3%)	10 (4.5%)	54 (8.2%)	
Sex				0.390
Female	500 (57.0%)	132 (59.5%)	368 (56.2%)	
Male	377 (43.0%)	90 (40.5%)	287 (43.8%)	
BMI, kg/m²	23.240 ± 3.000	22.990 ± 3.070	23.320 ± 2.970	0.160
Time from admission to surgery, days	3.750 ± 2.330	3.780 ± 2.470	3.740 ± 2.280	0.800
Fracture side				0.790
Left	491 (56.0%)	126 (56.8%)	365 (55.7%)	
Right	386 (44.0%)	96 (43.2%)	290 (44.3%)	
Fracture displacement				<0.001
Nondisplaced	467 (53.2%)	64 (28.8%)	403 (61.5%)	
Displaced	410 (46.8%)	158 (71.2%)	252 (38.5%)	
Reduction quality				<0.001
Good	785 (89.5%)	170 (76.6%)	615 (93.9%)	
Poor	92 (10.5%)	52 (23.4%)	40 (6.1%)	

Continuous variables are presented as mean ± standard deviation and were compared using an independent-samples *t* test; categorical variables are presented as number (percentage) and were compared using a chi-square test. BMI, body mass index; ONFH, femoral head osteonecrosis.

Performance across radiographic configurations

Results for the nine configurations using original radiographs in the fixed internal test set are shown in Table 2. Single-view AUROCs ranged from 0.770 to 0.829, with the immediate postoperative AP view performing best. The postoperative AP and lateral single-view models had AUROCs of 0.829 and 0.806, respectively, numerically higher than the corresponding preoperative models (0.770 and 0.791). The same pattern appeared in the biplanar models: postoperative and preoperative biplanar AUROCs were 0.861 and 0.807, respectively (paired difference 0.054, 95% CI −0.004 to 0.112; DeLong $p = 0.070$). Across four dual-view configurations, AUROCs ranged from 0.807 to 0.861. Combining AP and lateral views from the same time point increased the AUROC by 0.016–0.032 relative to the better corresponding single-view model; pairing preoperative and postoperative images of the same projection increased it by 0.010–0.024. These comparisons across configurations were exploratory.

The four-view model had the highest point estimates for both AUROC and AUPRC among the nine original-image configurations (0.864 and 0.759, respectively; Table 2). It was therefore used as the principal imaging model for comparison with the conventional baseline.

Table 2. Test-set performance across image configurations

Configuration	Input	AUROC (95% CI)	AUPRC (95% CI)
Single view	pre-AP	0.770 (0.678–0.855)	0.592 (0.468–0.730)
	pre-Lat	0.791 (0.692–0.876)	0.661 (0.533–0.789)
	post-AP	0.829 (0.746–0.902)	0.649 (0.522–0.790)
	post-Lat	0.806 (0.707–0.892)	0.680 (0.552–0.816)
	post-AP (VRI)	0.802 (0.709–0.881)	0.599 (0.475–0.752)
	post-Lat (VRI)	0.804 (0.708–0.888)	0.655 (0.529–0.788)
Temporal pair	AP-pair	0.839 (0.756–0.911)	0.716 (0.597–0.835)
	AP-pair (VRI)	0.831 (0.747–0.905)	0.676 (0.546–0.821)
	Lat-pair	0.830 (0.735–0.913)	0.741 (0.622–0.856)

Configuration	Input	AUROC (95% CI)	AUPRC (95% CI)
	Lat-pair (VRI)	0.821 (0.723–0.905)	0.739 (0.624–0.845)
Biplanar pair	pre-pair	0.807 (0.715–0.889)	0.670 (0.544–0.798)
	post-pair	0.861 (0.782–0.929)	0.696 (0.568–0.849)
	post-pair (VRI)	0.853 (0.774–0.921)	0.699 (0.575–0.826)
Four views	quad	0.864 (0.780–0.935)	0.759 (0.636–0.880)
	quad (VRI)	0.845 (0.756–0.921)	0.711 (0.583–0.842)

Values are point estimates and 95% CIs from the fixed internal test set of 150 patients. “pre” and “post” denote preoperative and immediate postoperative images; AP and Lat denote anteroposterior and lateral projections. AP-pair and Lat-pair combine preoperative and postoperative images in the same projection; pre-pair and post-pair combine AP and lateral images at the same time point; quad combines all four images. VRI-labelled configurations were trained separately on reconstructed images. AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision–recall curve; CI, confidence interval; VRI, virtual implant removal.

Selection of the six-variable conventional baseline

In pooled five-fold out-of-fold predictions from the 727-patient development set, the six conventional models performed similarly: AUROC ranged from 0.693 to 0.713, AUPRC from 0.456 to 0.459, and Brier score from 0.167 to 0.169. Unpenalised logistic regression had an AUROC of 0.713, AUPRC of 0.458 and Brier score of 0.167; the corresponding values for ridge-penalised logistic regression were 0.713, 0.456 and 0.167. Because the more complex methods showed no consistent advantage, unpenalised logistic regression was selected to keep the baseline and subsequent fusion model simple.

Four-view model versus the conventional baseline

In the same 150-patient fixed internal test set, the six-variable baseline had an AUROC of 0.704 (95% CI 0.604–0.796) and AUPRC of 0.432 (95% CI 0.339–0.591). The corresponding values for the four-view model were 0.864 (95% CI 0.780–0.935) and 0.759 (95% CI 0.636–0.880; Table 3). In patient-level paired comparisons, AUROC was higher by 0.160 (95% CI 0.068–0.256) and AUPRC by 0.327 (95% CI 0.153–0.461) for the four-view model.

The Brier scores were 0.171 (95% CI 0.150–0.192) for the conventional baseline and 0.148 (95% CI 0.123–0.174) for the four-view model; the paired difference (four-view minus baseline) was −0.023 (95% CI −0.050 to 0.005). The baseline had a calibration intercept of −0.073 (95% CI −0.552 to 0.538) and slope of 0.900 (95% CI 0.458–1.480); the four-view model had an intercept of −1.027 (95% CI −1.488 to −0.758) and slope of 1.629 (95% CI 1.136–2.494). Grouped calibration is shown in Fig. 3. Thus, the four-view model ranked risk more effectively in this test set, but its raw probabilities showed calibration error; better discrimination alone does not establish more accurate individual absolute-risk estimates.

Table 3. Performance of principal and exploratory models in the test set

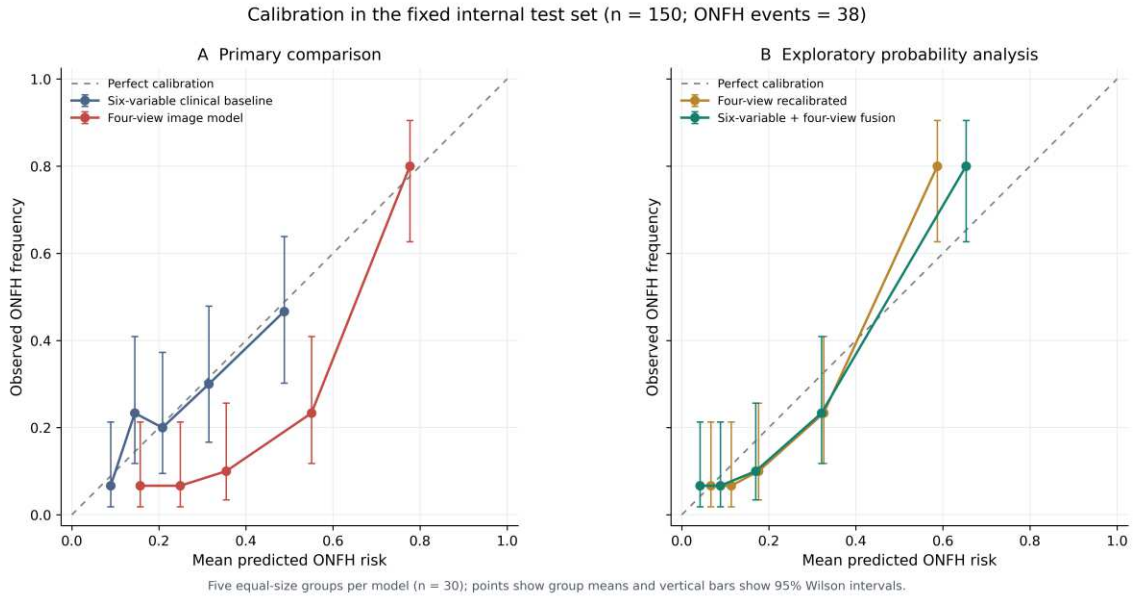
Model	AUROC (95% CI)	AUPRC (95% CI)	Brier score (95% CI)	Calibration intercept (95% CI)	Calibration slope (95% CI)
Six-variable unpenalised logistic baseline	0.704 (0.604–0.796)	0.432 (0.339–0.591)	0.171 (0.150–0.192)	−0.073 (−0.552 to 0.538)	0.900 (0.458–1.480)
Four-view raw probability	0.864 (0.780–0.935)	0.759 (0.636–0.880)	0.148 (0.123–0.174)	−1.027 (−1.488 to −0.758)	1.629 (1.136–2.494)
Six-variable + four-view fusion (exploratory)	0.863 (0.780–0.937)	0.725 (0.595–0.873)	0.115 (0.090–0.141)	0.202 (−0.183 to 0.775)	1.276 (0.854–1.998)
Recalibrated four-view probability (exploratory)	0.864 (0.781–0.934)	0.759 (0.636–0.876)	0.119 (0.098–0.142)	0.490 (0.095–1.118)	1.610 (1.131–2.435)

All values are point estimates and 95% CIs from the fixed internal test set of 150 patients. Calibration intercept and slope were estimated jointly by regressing outcome on the logit of predicted probability; ideal values are 0 and 1, respectively. Fusion and recalibration were

exploratory. AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision–recall curve; CI, confidence interval.

Fig. 3. Calibration of the clinical, four-view and exploratory models

(A) Principal comparison of the six-variable unpenalised logistic baseline with the raw four-view probability. (B) Exploratory comparison of the recalibrated four-view probability with the six-variable-plus-four-view fusion model. For each model, the same 150 test patients were divided into five equal-sized groups (30 per group) by predicted risk. Points show the mean predicted risk and observed ONFH proportion; error bars are 95% Wilson CIs, and the dashed line indicates ideal calibration. The five groups for different models need not comprise identical patients. ONFH, femoral head osteonecrosis.



Exploratory clinical–imaging fusion and probability recalibration

In the test set, the seven-input logistic fusion model had an AUROC of 0.863 (95% CI 0.780–0.937), AUPRC of 0.725 (95% CI 0.595–0.873) and Brier score of 0.115 (95% CI 0.090–0.141). Relative to the four-view model alone, its AUROC was essentially unchanged (0.863 versus

0.864), its AUPRC was numerically lower (0.725 versus 0.759), and its Brier score was numerically lower (0.115 versus 0.148). Its calibration intercept was 0.202 (95% CI −0.183 to 0.775) and slope was 1.276 (95% CI 0.854–1.998). The fusion model showed no additional discrimination.

To examine how much of the lower Brier score might be explained by changing the probability scale, we fitted a logistic recalibration model with an intercept and slope using the four-view validation-fold probabilities from the development set. In the test set, AUROC and AUPRC remained 0.864 and 0.759; the Brier score decreased to 0.119 (95% CI 0.098–0.142), a paired difference of −0.029 (95% CI −0.041 to −0.017) versus the original four-view model. The recalibrated test-set intercept was 0.490 (95% CI 0.095–1.118) and the slope was 1.610 (95% CI 1.131–2.435), indicating residual miscalibration. The similar Brier scores of the fusion and recalibration models suggest that the improvement relative to the original four-view model should not be attributed entirely to adding clinical variables. Moreover, outcomes in the respective validation folds informed image-checkpoint selection; consequently, the probabilities used for second-stage fitting were not strictly out-of-fold with respect to the entire image-training process. Fusion and recalibration were therefore interpreted as exploratory.

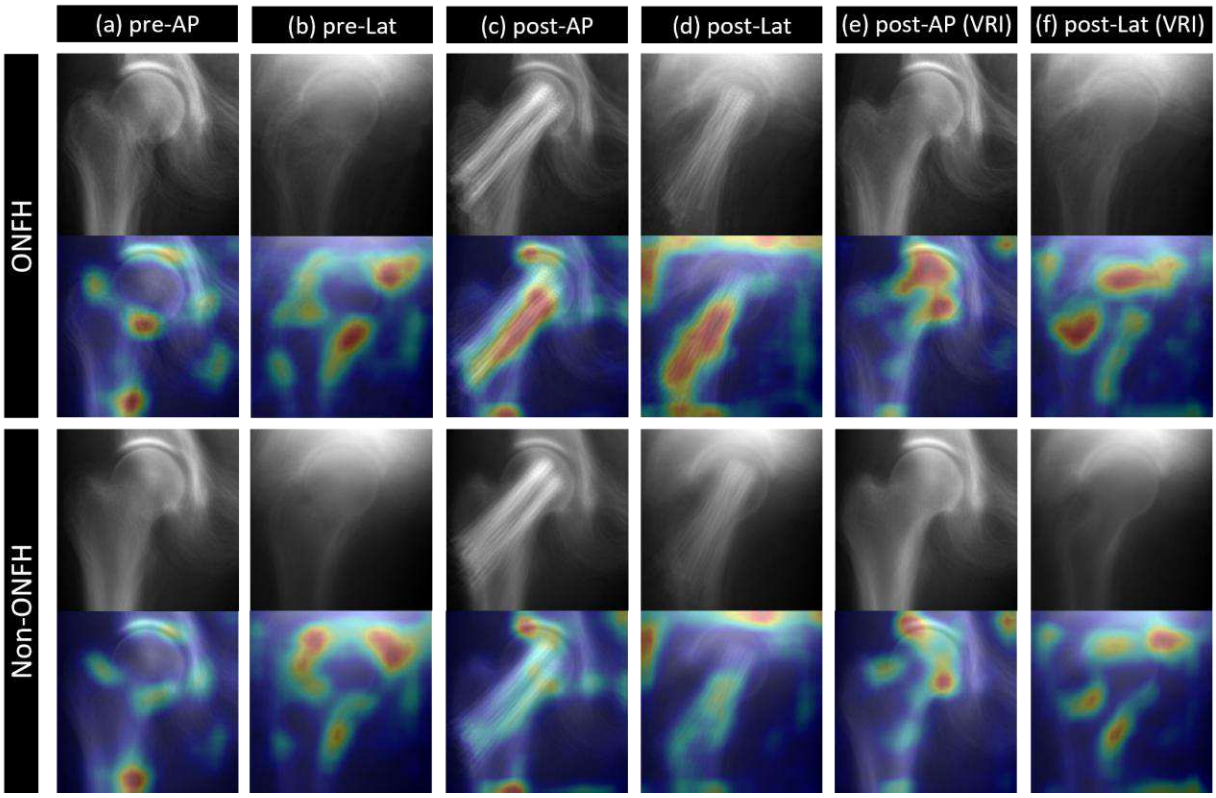
Grad-CAM visualisation and virtual implant removal

In the fixed test set, Grad-CAM heatmaps from separately trained single-view models were summarised for patients with and without ONFH (Fig. 4). Group-level activation patterns differed visibly. Compared with the group without ONFH, activation in the ONFH group was more concentrated near the head–neck junction, medial and posterior femoral neck cortices, calcar femorale and screw locations. Heatmaps from original postoperative AP and lateral models showed bands of activation along the screw trajectories. These bands were less apparent

in models trained separately on VRI images, whereas activation around the femoral head, head–neck junction and cortex remained visible.

Fig. 4. Group-level Grad-CAM summaries for single-view models

Images and corresponding Grad-CAM heatmaps from the fixed internal test set were summarised separately for patients with ONFH (upper row) and without ONFH (lower row): (a) pre-AP, (b) pre-Lat, (c) post-AP, (d) post-Lat, (e) post-AP (VRI) and (f) post-Lat (VRI). For each input configuration, patient-level heatmaps were normalised to [0, 1] and averaged pixel by pixel within each outcome group. “pre” and “post” denote preoperative and immediate postoperative images. Warmer colours indicate relatively stronger mean activation. Heatmaps were generated from independently trained single-view models, not the fused multi-view model. AP, anteroposterior; Lat, lateral; ONFH, femoral head osteonecrosis; VRI, virtual implant removal.



A reconstruction example is shown in Supplementary Fig. S2. Each VRI configuration was retrained with reconstructed images and compared with its original-image counterpart in the same 150 test patients. AUROC was numerically lower in all six configurations involving postoperative images, by 0.002–0.027 (Table 2). The four-view AUROC decreased from 0.864 with original images to 0.845 with VRI images (paired original-minus-VRI difference 0.019, 95% CI −0.008 to 0.047). The corresponding values for the postoperative biplanar model were 0.861 and 0.853 (paired difference 0.008, 95% CI −0.038 to 0.051). Both CIs included zero.

Discussion

In the same fixed internal test set, the four-view model had higher AUROC and AUPRC than the six-variable conventional baseline, with paired differences of 0.160 and 0.327, respectively. In contrast, although its Brier score was numerically lower, the CI for the difference included zero, and the raw probabilities showed calibration error. The Brier-score difference should be interpreted in terms of its magnitude and interval estimate rather than reduced to a binary distinction based on statistical significance [19]. These internal findings indicate that direct analysis of perioperative radiographs provides stronger discrimination of subsequent ONFH than the six-variable baseline used here. They do not establish that the uncalibrated image-model output is an accurate estimate of an individual's absolute risk.

Fracture displacement and reduction quality are recurring predictors in studies of ONFH after FNF fixation [6, 7, 20, 21]. These visually assessed measures summarise important aspects of the fracture and postoperative state, but structural variation remains within a given category. For example, differences in preoperative posterior tilt among Garden I–II fractures are associated with fixation failure [22]. Earlier radiographic deep-learning studies have mainly addressed detection or staging of established ONFH [10, 11]. All four inputs in our study were routine

preoperative or immediate postoperative radiographs, bringing risk assessment to the perioperative period before conspicuous signs of osteonecrosis appear.

The comparisons among image configurations also help contextualise the four-view model. AUROCs for corresponding postoperative single-view and biplanar models were numerically higher than those for preoperative configurations, suggesting that reduction and implant-related information may be relevant to the subsequent outcome. Combining AP and lateral views at the same time point, or pairing preoperative and postoperative images in the same projection, yielded numerical AUROC gains relative to the corresponding single views. AP and lateral radiographs depict the femoral neck, cortices and reduced structure from different directions, whereas preoperative and postoperative images record different stages of morphology; these are plausible sources of complementary information [20, 22]. The four-view model had the highest AUROC and AUPRC point estimates. Although postoperative biplanar and four-view AUROCs were similar, the former had a lower AUPRC, suggesting potential differences in identifying positive cases.

Adding the six conventional variables to the four-view probability did not further improve AUROC or AUPRC. Its lower Brier score should be interpreted alongside the recalibration analysis. Much of the apparent gain over the original four-view model may reflect adjustment of the probability scale rather than new discrimination from clinical variables. Equally, the lack of additional discrimination does not prove that clinical and image information are redundant. In particular, the outcomes of the development-set validation folds informed image-checkpoint selection, while test-set probabilities were averaged across five fold-specific models. The different ways in which the second-stage image inputs were generated limit the fusion and recalibration analyses to exploratory evidence.

The four-view model's calibration intercept and slope departed from their ideal values, reinforcing the need to assess discrimination separately from absolute-risk estimation. For a model intended to influence follow-up, calibration determines whether predicted probabilities can support individual risk interpretation and threshold-based decisions [23]. Recalibration reduced overall prediction error but did not bring the test-set calibration slope close to 1. A simple probability transformation is therefore insufficient evidence that the model is ready to report an individual's probability of ONFH. Independent cohorts should be used to reassess calibration, specify possible follow-up actions and associated risk thresholds, and evaluate whether those actions produce net benefit.

Grad-CAM heatmaps from the single-view models overlapped with regions clinicians examine when assessing fracture morphology, cortical continuity and implant position, providing qualitative anatomical clues to the model output. VRI models showed only small numerical AUROC decreases relative to corresponding original-image models, and CIs for the four-view and postoperative biplanar paired differences included zero. Because original and reconstructed images were used to train separate models, and reconstruction may alter osseous texture around the screws, these observations cannot isolate the predictive contributions of the implant and surrounding bone or establish a biological mechanism.

A strength of this study is the assessment of multiple configurations of routine preoperative and immediate postoperative radiographs from the same patients, together with patient-level paired comparisons between an image model and a six-variable conventional baseline in a fixed internal test set. We assessed discrimination, Brier score and calibration rather than relying on AUROC alone. Several limitations remain. First, this was a single-centre retrospective study, and only 877 of 3,208 screened patients met the image and follow-up criteria, creating potential selection bias.

The outcome was determined from routine follow-up records; patients classified as not having ONFH were required to have at least 24 months of follow-up, so occult or later ONFH may have been missed. Second, the fixed test set contained only 150 patients, including 38 ONFH cases, resulting in relatively wide interval estimates. Without external validation, performance across other institutions, imaging systems and patient populations is unknown. Independent multicentre validation, calibration and clinical-utility assessment will help clarify the four-view model's potential role in perioperative risk stratification. The fusion approach requires further development using image predictions that are strictly out-of-fold across the complete training process.

Conclusions

We developed and compared models using routine preoperative and immediate postoperative radiographs after FNF fixation, with a six-variable conventional model as the reference for assessing the added value of four-view imaging. In this single-centre fixed internal test set, the four-view model discriminated subsequent ONFH more effectively, suggesting that full radiographs contain prognostic information not fully captured by conventional variables. Exploratory clinical–imaging fusion did not improve discrimination, and the raw four-view probabilities showed calibration error. External validation, calibration and clinical-utility assessment will help determine whether this approach can inform early risk stratification and follow-up decisions.

List of abbreviations

AP, anteroposterior; AUPRC, area under the precision–recall curve; AUROC, area under the receiver operating characteristic curve; BMI, body mass index; CCS, cannulated compression screw fixation; CI, confidence interval; FNF, femoral neck fracture; Grad-CAM, gradient-

weighted class activation mapping; ONFH, femoral head osteonecrosis; ROI, region of interest; VRI, virtual implant removal.

Declarations

Ethics approval and consent to participate

This single-centre retrospective study was approved by the Ethics Committee of Tianjin Hospital (approval no. 2024-081; approval date: 12 May 2024) and was conducted in accordance with the Declaration of Helsinki. The requirement for informed consent to participate was waived by the Ethics Committee of Tianjin Hospital.

Consent for publication

Informed consent for publication of the de-identified radiographs was obtained from the participants whose images are presented in Figs. 2 and 4 and Supplementary Figs. S1 and S2. The images contain no names, examination numbers or other direct identifiers.

Availability of data and materials

Summary results are reported in the article. Patient-level clinical data and original radiographs are not publicly available because of patient privacy. Access to de-identified data may be requested from the corresponding author, subject to Tianjin Hospital's data-governance and ethics requirements.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Tianjin Health Science and Technology Project (TJWJ2023MS021) and the Tianjin Natural Science Foundation Diversified Investment General

Project (21JCYBJC01840). The funders had no role in the study design; data collection, analysis or interpretation; preparation of the manuscript; or the decision to submit the manuscript for publication.

Authors' contributions

XG contributed to formal analysis, investigation, methodology, software, validation, visualization, and writing the original draft. ZH contributed to conceptualization, data curation, formal analysis, methodology, supervision, validation, and reviewing and editing the manuscript. ZZ contributed to data curation, investigation, software, and visualization. XZ contributed to data curation, investigation, and visualization. QD contributed to conceptualization, project administration, and resources. KS contributed to conceptualization, funding acquisition, project administration, resources, supervision, and reviewing and editing the manuscript. XM contributed to funding acquisition and resources. XG and ZH contributed equally to this work. QD and KS are co-corresponding authors. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Additional file

Additional file 1 (.docx). Supplementary methods and figures. Detailed image-enhancement and intensity-normalisation procedure, network architecture, training settings and data augmentation; Supplementary Fig. S1 showing a radiograph before and after dark-channel-prior enhancement; and Supplementary Fig. S2 showing an example of virtual implant removal.

References

1. Sing CW, Lin TC, Bartholomew S, et al. Global epidemiology of hip fractures: secular trends in incidence rate, post-fracture treatment, and all-cause mortality. J Bone Miner Res. 2023;38(8):1064-1075. doi:10.1002/jbmr.4821. 2. Rogmark C, Viberg B, Wolf O, Mukka S, Costa ML, Gjertsen JE. Femoral neck fractures in adults with emphasis on surgical treatment. Acta Orthop. 2025;96:632-639. doi:10.2340/17453674.2025.44354. 3. Konarski W, Poboży T, Kotela A, et al. The risk of avascular necrosis following the stabilization of femoral neck fractures: a systematic review and meta-analysis. Int J Environ Res Public Health. 2022;19(16):10050. doi:10.3390/ijerph191610050. 4. Hu Y, Yang Q, Zhang J, Peng Y, Guang Q, Li K. Methods to predict osteonecrosis of femoral head after femoral neck fracture: a systematic review of the literature. J Orthop Surg Res. 2023;18(1):377. doi:10.1186/s13018-023-03858-7. 5. Kim CH, Shin M, Lee D, Choi SJ, Moon DH. Hidden osteonecrosis of the femoral head after healed femoral neck fractures: magnetic resonance imaging study of 58 consecutive patients. Arch Orthop Trauma Surg. 2022;142(7):1443-1450. doi:10.1007/s00402-021-03802-6. 6. Wang H, Wu W, Han C, et al. Prediction model of osteonecrosis of the femoral head after femoral neck fracture: machine learning-based development and validation study. JMIR Med Inform. 2021;9(11):e30079. doi:10.2196/30079. 7. Liu B, Hou G, Yang Z, Zhang Z, Zhou F, Tian Y. Machine learning models to predict osteonecrosis in patients with femoral neck fractures undergoing internal fixation. Injury. 2024;55(11):111830. doi:10.1016/j.injury.2024.111830. 8. Wang W, Huang Z, Peng J, Fan J, Long X. Preoperative posterior tilt can be a risk factor of fixation failure in nondisplaced femoral neck fracture: a systematic review and meta-analysis. Eur J Orthop Surg Traumatol. 2023;33(7):3197-3205. doi:10.1007/s00590-023-03518-w. 9. Cazzato G, Oliva MS, Masci G, et al. Femoral neck fracture: the reliability of radiologic

classifications. BMC Musculoskelet Disord. 2022;22(Suppl 2):1063. doi:10.1186/s12891-022-05007-3. 10. Chee CG, Kim Y, Kang Y, et al. Performance of a deep learning algorithm in detecting osteonecrosis of the femoral head on digital radiography: a comparison with assessments by radiologists. AJR Am J Roentgenol. 2019;213(1):155-162. doi:10.2214/AJR.18.20817. 11. Li Y, Li Y, Tian H. Deep learning-based end-to-end diagnosis system for avascular necrosis of femoral head. IEEE J Biomed Health Inform. 2021;25(6):2093-2102. doi:10.1109/JBHI.2020.3037079. 12. Zhu W, Zhang X, Fang S, Wang B, Zhu C. Deep learning improves osteonecrosis prediction of femoral head after internal fixation using hybrid patient and radiograph variables. Front Med (Lausanne). 2020;7:573522. doi:10.3389/fmed.2020.573522. 13. Yoon BH, Mont MA, Koo KH, et al. The 2019 revised version of Association Research Circulation Osseous staging system of osteonecrosis of the femoral head. J Arthroplasty. 2020;35(4):933-940. doi:10.1016/j.arth.2019.11.029. 14. He K, Sun J, Tang X. Single image haze removal using dark channel prior. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL: IEEE; 2009:1956-1963. doi:10.1109/CVPR.2009.5206515. 15. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV: IEEE; 2016:770-778. doi:10.1109/CVPR.2016.90. 16. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE; 2017:618-626. doi:10.1109/ICCV.2017.74. 17. Kirillov A, Mintun E, Ravi N, et al. Segment Anything. In: 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE; 2023:4015-4026. doi:10.1109/ICCV51070.2023.00371. 18. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang TS. Free-form image inpainting with gated convolution. In: 2019 IEEE/CVF

International Conference on Computer Vision. Seoul, Republic of Korea: IEEE; 2019:4471-4480. doi:10.1109/ICCV.2019.00457. 19. Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. Am Stat. 2016;70(2):129-133. doi:10.1080/00031305.2016.1154108. 20. Wang Y, Ma JX, Yin T, et al. Correlation between reduction quality of femoral neck fracture and femoral head necrosis based on biomechanics. Orthop Surg. 2019;11(2):318-324. doi:10.1111/os.12458. 21. Xu JL, Liang ZR, Xiong BL, et al. Risk factors associated with osteonecrosis of femoral head after internal fixation of femoral neck fracture: a systematic review and meta-analysis. BMC Musculoskelet Disord. 2019;20(1):632. doi:10.1186/s12891-019-2990-5. 22. Dolatowski FC, Adampour M, Frihagen F, Stavem K, Utvag SE, Hoelsbrekken SE. Preoperative posterior tilt of at least 20° increased the risk of fixation failure in Garden-I and -II femoral neck fractures: 322 patients followed for a mean of 3 years. Acta Orthop. 2016;87(3):252-256. doi:10.3109/17453674.2016.1155253. 23. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. doi:10.1186/s12916-019-1466-7.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1SupplementaryMaterial.docx