JAMA Network Open

Original Investigation | Orthopedics

Artificial Intelligence for Hip Fracture Detection and Outcome Prediction
A Systematic Review and Meta-analysis

Johnathan R. Lex, MBChB; Joseph Di Michele, MD; Robert Koucheki, MD; Daniel Pincus, MD, PhD; Cari Whyne, PhD; Bheeshma Ravi, MD, PhD

Abstract

IMPORTANCE Artificial intelligence (AI) enables powerful models for establishment of clinical diagnostic and prognostic tools for hip fractures; however the performance and potential impact of these newly developed algorithms are currently unknown.

OBJECTIVE To evaluate the performance of AI algorithms designed to diagnose hip fractures on radiographs and predict postoperative clinical outcomes following hip fracture surgery relative to current practices.

DATA SOURCES A systematic review of the literature was performed using the MEDLINE, Embase, and Cochrane Library databases for all articles published from database inception to January 23, 2023. A manual reference search of included articles was also undertaken to identify any additional relevant articles.

STUDY SELECTION Studies developing machine learning (ML) models for the diagnosis of hip fractures from hip or pelvic radiographs or to predict any postoperative patient outcome following hip fracture surgery were included.

DATA EXTRACTION AND SYNTHESIS This study followed the Preferred Reporting Items for Systematic Reviews and Meta-analyses and was registered with PROSPERO. Eligible full-text articles were evaluated and relevant data extracted independently using a template data extraction form. For studies that predicted postoperative outcomes, the performance of traditional predictive statistical models, either multivariable logistic or linear regression, was recorded and compared with the performance of the best ML model on the same out-of-sample data set.

MAIN OUTCOMES AND MEASURES Diagnostic accuracy of AI models was compared with the diagnostic accuracy of expert clinicians using odds ratios (ORs) with 95% CIs. Areas under the curve for postoperative outcome prediction between traditional statistical models (multivariable linear or logistic regression) and ML models were compared.

RESULTS Of 39 studies that met all criteria and were included in this analysis, 18 (46.2%) used AI models to diagnose hip fractures on plain radiographs and 21 (53.8%) used AI models to predict patient outcomes following hip fracture surgery. A total of 39 598 plain radiographs and 714 939 hip fractures were used for training, validating, and testing ML models specific to diagnosis and postoperative outcome prediction, respectively. Mortality and length of hospital stay were the most predicted outcomes. On pooled data analysis, compared with clinicians, the OR for diagnostic error of ML models was 0.79 (95% CI, 0.48-1.31; $P = .36$; $I^2 = 60\%$) for hip fracture radiographs. For the ML models, the mean (SD) sensitivity was 89.3% (8.5%), specificity was 87.5% (9.9%), and F1 score was 0.90 (0.06). The mean area under the curve for mortality prediction was 0.84 with ML models compared with 0.79 for alternative controls ($P = .09$).

(continued)

Key Points

Question For patients with hip fractures, how well do current artificial intelligence algorithms perform at diagnosing fractures and predicting postoperative outcomes?

Findings This systematic review and meta-analysis of 39 studies identified similar error rates of hip fracture diagnosis between artificial intelligence models and expert clinicians. There was minimal advantage of machine learning models over traditional regression techniques for postoperative outcome prediction.

Meaning These findings suggest that artificial intelligence has the potential to automate hip fracture diagnosis; however, complicated, uninterpretable models may not provide benefit over traditional, interpretable models for patient-specific outcome prediction.

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Abstract (continued)

CONCLUSIONS AND RELEVANCE The findings of this systematic review and meta-analysis suggest that the potential applications of AI to aid with diagnosis from hip radiographs are promising. The performance of AI in diagnosing hip fractures was comparable with that of expert radiologists and surgeons. However, current implementations of AI for outcome prediction do not seem to provide substantial benefit over traditional multivariable predictive statistics.

JAMA Network Open. 2023;6(3):e233391. doi:10.1001/jamanetworkopen.2023.3391

Introduction

The number of artificial intelligence (AI) algorithms in the medical literature and health care industry is increasing rapidly. This increase is due to relatively recent advances in computational power, data accessibility, and model complexity through mathematical and computer science research.[1,2] Correspondingly, there has been an increasing number of AI algorithms and AI-enabled medical devices approved by the US Food and Drug Administration (79 and 343, respectively).[3-5] There are a large number of potential applications for AI; however, most of the models developed have focused on the interpretation of medical imaging and clinical decision support systems.[3] Across the literature, AI models are beginning to show the ability to automate and potentially improve clinicians' diagnostic and clinical decision-making.[6,7] However, preceding research and literature reviews have predominantly been conducted in the fields of radiology, pathology, oncology, and ophthalmology, with a smaller proportion of research being conducted within the surgical specialties and particularly orthopedic surgery.[6-10]

The most prominent domain within orthopedic surgery in which AI research has been conducted is in hip fractures. Among elderly populations, hip fractures make up more than 14% of total fractures, although they represent a disproportionate 72% of fracture-related health care costs.[11,12] Approximately 300 000 hip fractures occur per year in the US alone.[13,14] Despite prevention efforts, this number is steadily increasing due to an aging population.[13,15,16] Worldwide, this number is expected to reach 6.3 million hip fractures at a cost of $131.5 billion per year by 2050.[17,18] In addition to their significant prevalence and economic impact, hip fractures are also associated with significant individual morbidity and mortality, with a 1-year mortality rate of approximately 25% to 30%.[19-22] Therefore, technology to improve the efficiency of managing this condition has the potential to improve patient outcomes and provide economic benefit to health care systems.

Improvement of the efficiency of hip fracture diagnosis and surgery has received considerable attention in recent years. Expedited management and comprehensive care pathways have been proven to improve outcomes, including survival rate, for these patients.[22-24] These circumstances provide an ideal use case for this novel technology should its performance be equal or superior to human performance. Image analysis and clinical decision support systems powered by AI may automate sections of the diagnostic pathway and improve outcome prediction accuracy.[6-8] Expedited diagnosis by leveraging this technology would lead to rapid diagnosis and access to surgical care. Perioperative risk stratification for clinicians and hospitals caring for these patients can assist in decision-making, accurate expectation management, and financial and resource planning. Moreover, these applications have the potential to reduce errors secondary to physician fatigue from repetitive cognitive demands and improve informed decision-making for patients and families.[25-27] In this systematic review, we sought to evaluate the literature and performance of AI algorithms designed to improve the management of hip fractures in elderly patients across 2 domains: (1) to evaluate the performance of AI compared with health care professionals for detection of hip fractures on medical imaging and (2) to determine the accuracy of AI at predicting various postoperative clinical outcomes compared with traditional statistical methods.

Methods

Search Strategy and Study Selection

A systematic review of the literature was performed using MEDLINE, Embase, and the Cochrane Library for all articles published from database inception to January 23, 2023. The Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) 2020 reporting guideline was used to design the review.[28] The inclusion criteria and analysis plan were decided a priori and registered on PROSPERO (CRD42022351255). In each database, the following keywords were combined to identify relevant articles: *hip* OR *neck of femur* OR *femoral neck* OR *intertrochanteric* OR *pertrochanteric* OR *subtrochanteric* AND *fracture* OR *broken* AND *artificial intelligence* OR *AI* OR *machine learning* OR *ML* OR *computer vision* OR *neural network*. A manual reference search of included articles was also undertaken to identify any additional relevant articles. This review had 2 groups: (1) studies that developed any machine learning (ML) or deep learning model for the diagnosis of hip fractures using medical imaging and (2) studies that developed a model designed to predict any postoperative patient outcome following hip fracture surgery.

Eligibility Criteria and Data Extraction

Artificial intelligence models evaluating the radiographic diagnosis or outcome prediction of hip fractures, including femoral neck, intertrochanteric, and subtrochanteric fractures, were included, with isolated fractures of any other anatomical site being excluded (acetabular, pelvic, femoral head, midshaft femur, or distal femoral fractures). Studies designed to diagnose hip fractures from medical imaging were included if they were based on anteroposterior and lateral hip or anteroposterior pelvic plain radiographs. Ground truth must have been based on image review by a consensus medical expert group, radiologist report and image review by a staff radiologist, surgical confirmation, or cross-sectional imaging (computed tomography or magnetic resonance imaging) confirmation. All level 3 or higher studies, including randomized clinical trials, prospective studies, and retrospective studies, were included. Studies were not excluded based on the presence or absence of a comparator group or language of publication. Case reports, literature reviews, abstracts, unpublished studies, and nonhuman studies were excluded. Authors were contacted if needed to retrieve copies of unavailable manuscripts.

Screening of search results based on titles and abstracts was performed by 2 independent reviewers (J.D.M. and R.K.), with conflicts resolved by inclusion of a third reviewer (J.R.L.). Three reviewers independently assessed the eligibility following abstract screening for study inclusion according to the inclusion and exclusion criteria. In cases of conflict, decisions were made through consensus agreement among the 3 reviewers. Eligible full-text articles were evaluated, and relevant data were extracted independently by 2 reviewers (J.D.M. and R.K.) using a template data extraction form, with conflicts resolved by inclusion of a third reviewer (J.R.L.).

When studies generated multiple AI models to predict the same outcome, we recorded performance data from the best-performing model on the test or out-of-sample data subset. This process ensured that only the best-performing model would be chosen for use in clinical practice. When reported, the accuracy of staff (consultant-level) orthopedic surgeons or radiologists at diagnosing hip fractures was compared with AI models. Performance of resident physicians was not included. For the comparator groups evaluating clinician performance, if multiple clinicians were evaluated, their mean accuracy score was calculated from their performance on the same out-of-sample data subset. A mean score was calculated for clinician performance because this most closely resembles the diagnostic accuracy of the current workflow, with different radiologists or surgeons making the diagnosis, depending on the day.

For studies that predicted postoperative outcomes, the performance of traditional predictive statistical models, either multivariable logistic or linear regression, was recorded and compared with the performance of the best ML model on the same out-of-sample data set. This method of statistical analysis is most typically used to generate predictive or prognostic clinical models and was therefore

used as a baseline performance indicator. Any more advanced algorithm type, including regression with regularization, was considered a form of ML.

Statistical Analysis

Odds ratios (ORs) with 95% CIs were used for dichotomous outcome measures. Heterogeneity was assessed using the I^2 statistic, with $I^2 \geq 75\%$ indicating considerable heterogeneity. A random-effects model to pool the data was planned to be used if considerable heterogeneity was found ($I^2 \geq 50\%$); otherwise, a fixed-effect model and a Mantel-Haenszel statistical method were used. Sensitivity and specificity of the diagnostic AI model's performance were plotted and compared with the performance of medical experts, with a pooled 95% CI around the mean. Youden index scores were calculated from sensitivity and specificity when reported. The area under the curve (AUC) of each predictive statistical and AI model were compared using a 2-tailed, unpaired *t* test. Microsoft Excel (Microsoft Corp) was used to extract data. For pooled data analysis, Review Manager (RevMan), version 5.4 (The Cochrane Collaboration) and Stata software, version 16.1 (StataCorp LLC) were used.

Results

Study Selection

Of 39 studies that met all criteria and were included in this analysis, 18 studies[29-46] (46.2%) used AI models to diagnose hip fractures on plain radiographs and 21 studies[47-67] (53.8%) used AI models to predict patient outcomes following hip fracture surgery. A PRISMA flowchart of included studies is displayed in eFigure 1 in Supplement 1. The characteristics of the included studies are given in **Table 1** (diagnostic studies) and **Table 2** (outcome prediction studies). Diagnostic studies were published between 2019 and 2022 and used a total of 39 598 plain radiographs to train, validate, and test AI models (Table 1; eTable 1 in Supplement 1). Outcome prediction studies were published between 2004 and 2022. Mortality followed by length of hospital stay were the most commonly predicted outcomes, with other predicted outcomes of 30-day complications, living situation, postoperative delirium, and modified functional independence measure.[47,58,61] A pooled total of 714 939 hip fractures were used for training, validating, and testing ML models specific to postoperative outcome prediction. All databases used for outcome prediction are listed in eTable 2 in Supplement 1.

Hip Fracture Diagnosis

All included studies developed a form of convolutional neural network model to diagnose fractures (Table 1). Comparative quantitative data for accuracy of hip fracture diagnosis as per plain radiographs were available from 8 studies (44.4%). Based on pooled data analysis, compared with clinicians, the OR for diagnostic error of AI models was 0.79 (95% CI, 0.48-1.31; $P = .36$; $I^2 = 60\%$) (**Figure 1**).

Among the included AI models, the mean (SD) sensitivity was 89.3% (8.5%), specificity was 87.5% (9.9%), and F1 score was 0.90 (0.06). Despite AI models having a higher overall accuracy, there was more variability in the sensitivity and specificity across models compared with clinician performance (**Figure 2**). Sensitivity ranged across studies from 67.0% to 98.0%, and specificity ranged from 70.0% to 98.7% (eTable 3 in Supplement 1). This wide range was predominantly due to 1 outlying study because 13 of 14 models (92.9%) reported a sensitivity greater than 80% and 11 of 14 models (78.6%) reported a specificity greater than 80%. eTable 3 in Supplement 1 also provides a breakdown of the F1 scores, κ scores, and Youden indexes of included AI models.

Postoperative Outcome Prediction

Machine learning models have been developed to predict the outcome of 6 different postoperative outcomes following hip fracture surgery: mortality (15 studies[49,50,52-60,63-67]), length of stay (3 studies[48,51,62]), delirium (1 study[58]), discharge destination (1 study[47]), hospital cost (1 study[51]), 30-day major complications (1 study[64]), and functional independence measure (1 study[61]) (Table 2).

Table 1. Included Studies on Application of Artificial Intelligence for Diagnosis of Hip Fractures

Source (country)	Input imaging	No. of output classes	Output	Algorithm used	No. of radiographs	Training size, %	Validation size, %	Testing size, No. or %	Ground truth
Cheng et al,[29] 2019 (Taiwan)	AP pelvic radiograph	2	Fractured (femoral neck and trochanteric); nonfractured	DCNN	3605	80	20	100	Radiologist report or CT report with each image reviewed by trauma surgeon
Urakawa et al,[30] 2019 (Japan)	AP proximal femoral radiograph	2	Fractured (intertrochanteric); nonfractured	CNN	3346	80	10	10%	Surgically confirmed
Adams et al,[31] 2019 (Australia)	AP hip cropped from AP pelvic radiograph	2	Fractured (femoral neck); nonfractured	AlexNet DCNN; GoogLeNet DCNN	640	80	20	160	Surgically confirmed
Mawatari et al,[32] 2020 (Japan)	Pelvic radiograph	2	Fractured (proximal femur fracture); nonfractured	DCNN	352	85.8	NR	14.2%	CT or MRI confirmed
Jiménez-Sánchez et al,[33] 2020 (Germany, Spain, and France)	AP and lateral pelvic radiograph, images were cropped	2	Fractured; nonfractured	ResNet-50; AlexNet	1347	70	10	20%	Image review by group of experts (1 staff trauma surgeon, 1 staff radiologist, 1 senior trauma resident)
		3	AO type A fracture; AO type B fracture; nonfractured	ResNet-50; AlexNet					
Yu et al,[34] 2020 (US)	AP hip radiograph	2	Fractured; nonfractured	DCNN	627	60	20	20%	Surgically confirmed or CT confirmed
Kitamura,[35] 2020 (US)	Pelvic radiograph	2	Normal; abnormal	Densenet-121 architecture	7337	70	NR	30%	Radiologist report and image review by a staff radiologist
		8	Normal; anterior pelvis; posterior pelvis; pelvic ring; proximal femur; acetabular; femur/acetabular; nonfemoral						
Krogue et al,[36] 2020 (US)	Hip and pelvic radiograph, hips were labeled via bounding boxes	6	Normal; displaced femoral neck fracture; nondisplaced femoral neck fracture; intertrochanteric fracture; previous open reduction and internal fixation; previous arthroplasty	DCNN (DenseNet)	1999	61.1	24.4	14.5%	Consensus by experts; CT, MRI, postoperative imaging in the event of uncertainty
		2	Fractured; nonfractured						
Yamada et al,[37] 2020 (Japan)	AP and lateral pelvic radiograph	3	Femoral neck fractures; trochanteric fractures; nonfractured	CNN	2923	89.7	10.3	NR	Consensus by experts and CT
Mutasa et al,[38] 2020 (US)	AP pelvic radiograph	2	Femoral neck fracture (any Garden fracture); normal	CNN	1063	Unclear	20	Unclear	Image review by a single staff radiologist
		3	Femoral neck fracture (Garden I/II fracture); femoral neck fracture (Garden III/IV fracture); normal						
Beyaz et al,[39] 2020 (Turkey)	AP hip cropped from AP pelvic radiograph, various image sizes assessed	2	Fractured (femoral neck); nonfractured	CNN; GA	234	NR	NR	NR	Unclear
Açıcı et al,[40] 2021 (Turkey)	AP pelvic radiograph	2	Fractured (femoral neck); nonfractured	CNN; GA; PSO; LSTM; BILSTM	64	NR	NR	NR	Unclear
Bae et al,[41] 2021 (Korea)	AP pelvic radiograph	3	Displaced fracture; nondisplaced fracture; nonfractured	CNN; ResNet 18 with CBAM	4189	80	10	10%	CT or MRI confirmed
Cheng et al,[42] 2021 (Taiwan)	AP pelvic radiograph	3	Hip fracture only; pelvic fracture only; no acute finding	PelviXNet (DenseNets + FPN); CNN	5204	100	NR	1888	Image review by group of clinicians
Guy et al,[43] 2021 (France)	AP and lateral pelvic radiograph	3	Femoral neck fracture; trochanteric fracture; nonfractured	Tensorflow deep learning algorithm	1309	80	10	10%	Unclear
Twinprai et al,[44] 2022 (Thailand)	AP hip and pelvic radiograph	2	Fractured; nonfractured	DCNN	1000	90	NR	10%	Consensus by experts and CT/MRI
		3	No fracture; trochanteric; intracapsular						
Murphy et al,[45] 2022 (UK)	AP pelvic radiograph	4	Normal; femoral neck fracture; intertrochanteric fracture; subtrochanteric fracture	CNN	3659	60	20	20%	Consensus by experts
Liu et al,[46] 2022 (China)	AP hip radiograph	2	Nonfractured; fractured (intertrochanteric)	Faster RCNN	700	91.9	NR	8.1%	Unclear

Abbreviations: AO, atlanto-occipital; AP, anteroposterior; BILSTM, bidirectional long short-term memory; CBAM, convolutional block attention module; CNN, convolutional neural network; CT, computed tomography; DCNN, deep convolutional neural network; FPN, feature pyramid network; GA, general algorithm; LSTM, long short-term memory; MRI, magnetic resonance imaging; NR, not reported; PSO, particle swarm optimization; RCNN, region-based convolutional neural network.

Age (18 of 21 studies[47,49-52,54-61,63-67] [85.7%]) and sex (17 of 21 studies [80.9%][47-52,54-58,60,61,64-67]) were the most used features, whereas all other input features varied widely across studies[60] and databases (eTable 4 in Supplement 1).

For 30-day mortality, median accuracy of the ML models was 72.8% (range, 71.0%-93.0%; n = 3), median AUC was 0.80 (range, 0.76-0.93; n = 6), median sensitivity was 73.0% (range,

Table 2. Studies on Application of Machine Learning Models for Predicting Postoperative Outcomes Following Hip Fracture Surgery

Source (country)	Algorithm used	Outcome predicted	No. of features in model	Time points of outcome	No. of output classes	Output	No. of hip fractures	Training size, %	Testing size, %	Ground truth
Ottenbacher et al,[47] 2004 (US)	ANN	Living setting	13	80-180 d After discharge	2	Home; not at home	3708	66.60	33.40	Retrospective database review
Sund et al,[48] 2009 (Finland)	Bayesian nonparametric MLP	Length of stay	22	NR	NR	NR	15 544	NR	NR	NR
Lin et al,[49] 2010 (Taiwan)	ANN	Mortality	12	1 y	2	Die; survive	286	68.88	31.12	Retrospective database review
Shi et al,[50] 2013 (China)	ANN	Mortality	8	1 y	2	Die; survive	2150	66.60	33.40	Retrospective database review
Karnuta et al,[51] 2019 (US, UK)	NB	Length of stay; cost	7	NR	LOS: 3 classes	LOS: 1-3 d; 4-6 d; 7-9 d; ≥10 d	98 562	90	10	Retrospective database review
					Cost: 4 classes	Cost:<$8464; $8464-$26 313;>$26 313				
Chen et al,[52] 2020 (Taiwan)	ANN	Mortality	9	Unclear (has stated both 30 or 90 d)	2	Die; survive	10 534	70.00	15.00	Retrospective database review
Zhang et al,[54] 2020 (China)	BBN	Mortality	16	1 y	2	Die; survive	448	90	10	Retrospective database review
DeBaun et al,[53] 2021 (US)	ANN	Mortality	47	30 d	2	30-d Mortality; survive	19 835	80	20	Retrospective database review
Cao et al,[55] 2021 (Sweden)	CNN	Mortality	6	30 d	2	Die; survive	134 915	80	20	Retrospective database review
Cowling et al,[56] 2021 (UK)	XGBoost algorithm, for tree models	Mortality	5	1 y	2	Die; survive	169 646	NR	NR	Retrospective database review
Forssten et al,[57] 2021 (Sweden, US)	SVM; NB; RF	Mortality	30	1 y	2	Die; survive	124 707	80	20	Retrospective database review
Oosterhoff et al,[58] 2021 (US, the Netherlands)	NN; SGM; SVM; RF; PLR	Postoperative delirium	6	30 d	2	Delirium; no delirium	28 207	80	20	Retrospective database review
Li et al,[59] 2021 (China)	RF	Mortality	10	1 mo; 3 mo; 6 mo; 1 y; 2 y	2	Die; survive	1330	NR	NR	Unclear
Cary et al,[60] 2021 (US)	MLP	Mortality	14	30 d; 1 y	2	Die; survive	17 140	Unclear	Unclear	Retrospective database review
Shtar et al,[61] 2021 (Israel)	AdaBoost; CatBoost; ExtraTrees; KNN; RF; SVM; XGBoost	Motor functional independence measure	25	NR	2	Better than the median score; worse than the median score	1625	Unclear	Unclear	Retrospective database review
Zhong et al,[62] 2021 (China)	PCR; SVM; BP	Length of stay	6	NR	NR	Outputting a specific LOS	182	80	20	Retrospective database review
Xing et al,[63] 2022 (China)	RF; lasso regression	Mortality	7	1 y	2	Die; survive	591	70	30	Retrospective database review
Harris et al,[64] 2022 (US)	Lasso regression	Mortality; major complications	15	30 d	NR	Outputting a specific risk	82 168	90	10	Retrospective database review
Oosterhoff et al,[65] 2022 (the Netherlands, US)	SGB; RF; SVM; NN; Elastic−Net Penalized Logistic Regression	Mortality	10	90 d; 2 y	NR	Outputting a specific risk	2478	80	20	Retrospective database review
Kitcharanant et al,[67] 2022 (Thailand)	GB; RF; ANN; LR; NB; SVM; KNN	Mortality	15	1 y	NR	Outputting a specific risk	492	70	30	Retrospective database review
Lei et al,[66] 2023 (China)	RF; GBM; DT; eXGBoost	In-hospital mortality	6	NR	NR	Outputting a specific risk	391	66.60	165 External validation	Retrospective database review

Abbreviations: ANN, artificial neural network; BBN, bayesian belief network; BP, backpropagation; CNN, convolutional neural network; DT, decision tree; GB, gradient boosting; GBM, gradient boosting machine; KNN, K-nearest neighbor; LOS, length of stay; LR, linear regression; MLP, multilayer perceptron; NB, naive Bayes; NN, neural network; NR, not reported; PCR, principal component regression; PLR, penalized logistic regression; RF, random forest; SGB, stochastic gradient boosting; SGM, semiglobal matching; SVM, support vector machine.

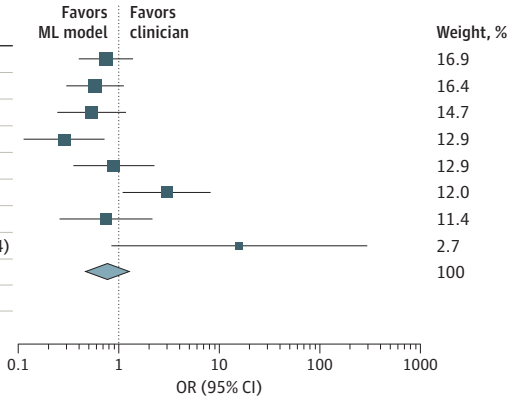
68.0%-94.0%; n = 3), and median specificity was 72.8% (range, 61.0%-97.0%; n = 5). For the ML models, 1-year mortality median accuracy was 85.8% (range, 68.0%-95.0%; n = 3), median AUC was 0.81 (range, 0.72-0.99; n = 9), median sensitivity was 70.0% (range, 68.0%-74.0%; n = 3), and median specificity was 67.2% (range, 61.0%-99.0%; n = 3). In predicting length of stay, 1 ML model reported an AUC of 0.88 and an accuracy of 76.5%[51]; accuracy statistics were not available from the other 2 studies. In predicting hospital costs, accuracy was 79.0% and AUC was 0.89 (n = 1). The AUCs were 0.79 for predicting 30-day postoperative delirium (n = 1), 0.73 for discharge destination (n = 1), and 0.86 for functional independence measure (n = 1).

There were only enough studies to compare the difference in ML prediction to traditional statistical models (multivariable linear or logistic regression) for the postoperative mortality outcome. The mean AUC for ML models was 0.84 compared with 0.79 for alternative controls (P = .09) (eTable 5 in Supplement 1).

Figure 1. Forest Plot Demonstrating the Accuracy of Artificial Intelligence Models Compared With Clinicians in Diagnosing Hip Fractures

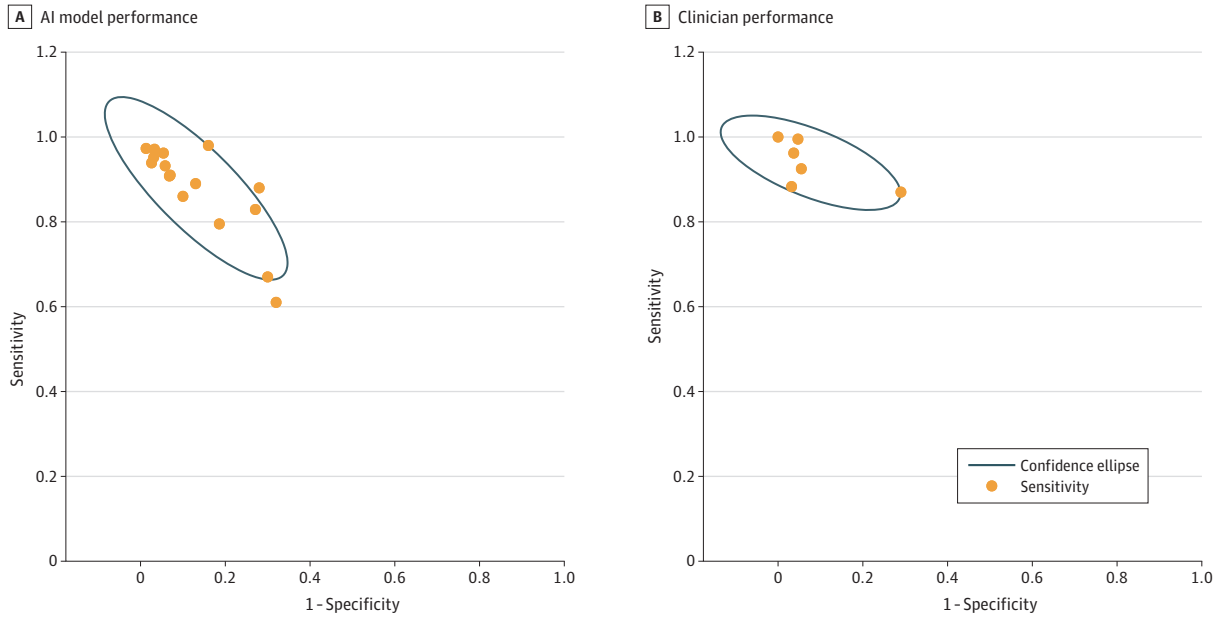
Study or subgroup	ML models Events	ML models Total	Clinician Events	Clinician Total	OR (95% CI)
Jiménez-Sánchez et al,[33] 2020	19	269	25	270	0.74 (0.40-1.39)
Urakawa et al,[30] 2019	15	334	25	334	0.58 (0.30-1.12)
Krogue et al,[36] 2020	18	289	11	100	0.54 (0.24-1.18)
Yamada et al,[37] 2020	6	300	20	300	0.29 (0.11-0.72)
Adams et al,[31] 2019	9	160	10	160	0.89 (0.35-2.26)
Cheng et al,[42] 2021	143	1888	4	150	2.99 (1.09-8.19)
Liu et al,[46] 2022	7	57	9	57	0.75 (0.26-2.16)
Yu et al,[34] 2020	4	125	0	213	15.81 (0.84-296.24)
Total (95% CI)		3422		1584	0.79 (0.48-1.31)
Total events	221		104		

Heterogeneity: $\tau^2 = 0.30$; $\chi^2 = 17.33$, $df = 7$ ($P = .02$); $I^2 = 60\%$
Test for overall effect: $z = 0.91$ ($P = .36$)



ML indicates machine learning; OR, odds ratio.

Figure 2. Sensitivity and Specificity of Artificial Intelligence (AI) Models Used for Diagnosing Hip Fractures on Radiographs and of Clinicians Who Used the Same Test Data Set



Discussion

This study identified a mixed utility of AI models to assist in the management of patients presenting with hip fractures. Multiple studies have developed independent, reliable algorithms specifically to diagnose hip fractures based on routinely collected plain radiographic imaging (eFigure 2 in Supplement 1). In this study, the meta-analysis conducted to compare the accuracy of these ML models revealed that the models are comparable with the mean performance of expert clinicians at diagnosing hip fractures. Across all included studies,[29-67] there was a wider range of sensitivity and specificity compared with clinician performance. However, the range was negatively skewed by 1 study that attempted to classify fractures into 3 different categories despite a relatively small training sample size. This review also compared the prediction of postoperative outcomes following hip fracture surgery of deep AI models compared with standard multivariable logistic or linear regression techniques (eFigure 2 in Supplement 1) and found no significant difference in AUC between the 2 statistical techniques in 60% of studies. Overall, these results must be interpreted with caution as described by the limitations below.

Because up to 10% of suspected hip fractures are not diagnosed on initial pelvic radiograph, techniques aiding the rapid and accurate identification of these fractures is imperative to enable expedient surgical management.[68] Such techniques are particularly prudent for hip fracture management because surgical delays longer than 24 hours are associated with a 20% increased risk of 30-day mortality, twice the risk of medical complications, and greater length of stay and medical costs.[22,69] Therefore, being able to automate the diagnosis and forecast outcomes for these patients can have significant benefits for patients and health care systems. However, despite the broad potential, the safety and reliability must be understood before the adoption of any new technology.[70] Contemporary ML algorithms compare more interactions between the included predictors in the aim to improve the accuracy of outcome prediction. The 2 studies[49,50] in which the ML models significantly outperformed traditional regression were both trained on relatively smaller data sets in the hundreds of training examples compared with thousands to tens of thousands. Studies that included larger sample sizes did not show similar advantages for ML. Nevertheless, because hip fractures are a common and expensive condition to manage, small improvements in prediction accuracy may help hospitals plan resources, such as staffing, beds, budget and implant procurement, improving system efficiency, and reducing expenses.

This study identified excellent potential in using AI to diagnosis hip fractures on routinely acquired plain radiographs; however, there are significant limitations and validation steps required before implementation. Missing a hip fracture (delaying diagnosis and possibly allowing fracture displacement) needs to be minimized by these algorithms (ie, avoiding false-negative results). Models are typically trained to maximize sensitivity and specificity by using an outcome metric that combines both (eg, AUC or F1 score). However, an alternative approach may be to train models to reduce false-negative results, followed by clinician screening of radiographs flagged as positive by the AI model. In the study with the largest training data set (approximately 5136 images), Kitamura[35] trained an algorithm that had 86.0% sensitivity and 90.0% specificity. Although the algorithm was able to detect pelvic imaging position, hardware presence, and pelvic and acetabular fractures, it was most accurate at diagnosing proximal femoral fractures. To reduce errors, training should be performed using a larger number of outlier examples or oversampling techniques, such as random or borderline oversampling or adaptive synthetic sampling.[71] A key consideration in the accuracy of hip fracture diagnoses was that although the performance of most ML models was comparable with the performance of expert radiologists and orthopedic surgeons, ML models consistently outperform trainees and nonexpert clinicians. This finding suggests that ML model use may be considered in assisting diagnoses in remote and resource-poor settings, where expert clinicians are not available.

Studies also demonstrated improvements in the accuracy of hip fracture diagnoses by experts when aided by ML algorithms. Considering that radiologists and orthopedic surgeons are facing increasing volumes of patients requiring radiographic interpretation for hip pathologies, addition of

ML models in aiding expert review can accelerate image interpretation and decrease processing times. The British Orthopaedic Association guidelines on managing hip fractures states that patients presenting to hospitals for hip fractures should be transferred to the ward within 4 hours of presentation.[72] This guideline has been shown to be a marker for quality of care provided by each hospital, and adherence to these guidelines has reduced time from admission to the operation room, reduced mortality at 30 days, and reduced mortality at 1 year.[73] However, there are poor rates of adherence to these guidelines. A potential solution is incorporation of ML models in clinical practice to accelerate diagnosis, prognosis determination, and admission planning, which may ultimately decrease time to transfer to ward and time to the operating room.[46]

All generated algorithms have been studied on retrospective data and evaluated using a holdout sample. It is also essential to perform further external validation at institutions separate from where the algorithms were trained, which no included studies performed. Across all specialties, 75% of algorithms perform moderately or substantially worse during external validation,[74] which may be due to different methods of data collection, imaging scanners, resolution, formatting, or image capture protocols. In addition, algorithms must also be studied on prospective data in a native clinical environment before their regular use. Integrating this technology into routine care pathways will allow the algorithms to be evaluated not only on accuracy but also on other important criteria, such as impact on patient outcomes, user acceptance, efficiency, resource utilization and planning, computational time, and cost. Integration into practice also comes with practical challenges, such as acceptability and ethical and legal responsibilities. Algorithms that provide a confidence level of their decision can be used to help aid clinician decision-making and patient screening to improve the efficiency and reduce the burden of work for clinicians.

Limitations

To our knowledge, this study was the first evaluation of the literature surrounding AI algorithms developed for the management of hip fractures from diagnosis to postoperative outcomes. This review provides insight into the potential impact that applying AI has on the management of one of the most common, resource-intensive, and devastating diagnoses. Nonetheless, there are limitations to the algorithms and methods of included studies, as mentioned previously, as well as to our review. A hierarchal summary receiver operating curve summarizing performance across all studies was unable to be created because studies did not report data in the form of a contingency table. Various AI techniques and predictive features were used across all studies, but we were unable to compare the use of each strategy and the effect this had on algorithm performance due to data and study reporting heterogeneity. Additionally, the quality of studies was unable to be properly evaluated because the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis for Artificial Intelligence (TRIPOD-AI) guidelines are still under development.[75]

Conclusions

This systematic review and meta-analysis helps evaluate the literature surrounding the current development of AI applications for the management of hip fractures. The potential applications regarding the use of AI to aid with diagnosis from hip and pelvic radiographs are promising. However, the use of AI does not seem to provide substantial additional benefit over traditional multivariable predictive statistics. The results of these applications are variable, which may be due to the quality or quantity of data from which these algorithms are developed rather than a true limitation of AI's power. Further studies should focus on evaluating whether these limitations remain with the use of large, accurate, multi-institutional data sets. Moreover, studies externally validating and implementing hip fracture diagnostic algorithms need to be performed to assess the effect on patient care.

ARTICLE INFORMATION

Accepted for Publication: January 31, 2023.

Published: March 17, 2023. doi:10.1001/jamanetworkopen.2023.3391

Corresponding Author: Johnathan R. Lex, MB, ChB, Division of Orthopaedic Surgery, Department of Surgery, University of Toronto, 500 University, Room 602, Toronto, ON M5G 1V7, Canada (johnathanlex@gmail.com).

Author Affiliations: Division of Orthopaedic Surgery, Department of Surgery, University of Toronto, Toronto, Ontario, Canada (Lex, Di Michele, Pincus, Ravi); Institute of Biomedical Engineering, University of Toronto, Toronto, Ontario, Canada (Lex, Koucheki); Orthopaedics Biomechanics Laboratory, Sunnybrook Research Institute, Toronto, Ontario, Canada (Lex, Whyne); Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada (Koucheki); Division of Orthopaedic Surgery, Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada (Pincus, Ravi).

Author Contributions: Drs Lex and Ravi had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Lex, Koucheki, Whyne, Ravi.

Acquisition, analysis, or interpretation of data: Lex, Di Michele, Koucheki, Pincus.

Drafting of the manuscript: Lex, Di Michele, Koucheki, Ravi.

Critical revision of the manuscript for important intellectual content: Di Michele, Koucheki, Pincus, Whyne.

Statistical analysis: Koucheki.

Administrative, technical, or material support: Koucheki.

Supervision: Lex, Di Michele, Pincus, Whyne, Ravi.

Conflict of Interest Disclosures: Dr Lex reported receiving grants from Arthrex Inc outside the submitted work and serving on the Resident Advisory Board for PrecisionOS Technologies. No other disclosures were reported.

Funding/Support: Scholarship support for this project (specifically for Dr Lex) was provided by the William and Suzanne Holland Chair in Musculoskeletal Research and the Queen Elizabeth II/Patty Rigby & John Wedge Graduate Scholarships in Science and Technology.

Role of the Funder/Sponsor: The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Sharing Statement: See Supplement 2.

REFERENCES

1. Thompson NC, Greenewald K, Lee K, Manso GF. The computational limits of deep learning. *arXiV*. Preprint posted online July 10, 2020. doi:10.48550/arXiv.2007.05558

2. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. doi:10.1038/nature14539

3. The Medical Futurist. FDA-approved A.I.-based algorithms. Accessed April 30, 2022. https://medicalfuturist.com/fda-approved-ai-based-algorithms/

4. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med*. 2020;3(1):118. doi:10.1038/s41746-020-00324-0

5. Center for Devices and Radiological Health. US Food and Drug Administration. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. Accessed April 30, 2022. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

6. Vasey B, Ursprung S, Beddoe B, et al. Association of clinician diagnostic performance with machine learning–based decision support systems: a systematic review. *JAMA Netw Open*. 2021;4(3):e211276. doi:10.1001/jamanetworkopen.2021.1276

7. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health*. 2019;1(6):e271-e297. doi:10.1016/S2589-7500(19)30123-2

8. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer*. 2018;18(8):500-510. doi:10.1038/s41568-018-0016-5

9. Kapoor R, Walters SP, Al-Aswad LA. The current state of artificial intelligence in ophthalmology. *Surv Ophthalmol*. 2019;64(2):233-240. doi:10.1016/j.survophthal.2018.09.002

10. Stewart JE, Rybicki FJ, Dwivedi G. Medical specialties involved in artificial intelligence research: is there a leader. *Tasman Medical J*. 2020;2(1):20-27.

11. Burge R, Dawson-Hughes B, Solomon DH, Wong JB, King A, Tosteson A. Incidence and economic burden of osteoporosis-related fractures in the United States, 2005-2025. *J Bone Miner Res*. 2007;22(3):465-475. doi:10.1359/jbmr.061113

12. Wiktorowicz ME, Goeree R, Papaioannou A, Adachi JD, Papadimitropoulos E. Economic implications of hip fracture: health service use, institutional care and cost in Canada. *Osteoporos Int*. 2001;12(4):271-278. doi:10.1007/s001980170116

13. Swayambunathan J, Dasgupta A, Rosenberg PS, Hannan MT, Kiel DP, Bhattacharyya T. Incidence of hip fracture over 4 decades in the Framingham Heart Study. *JAMA Intern Med*. 2020;180(9):1225-1231. doi:10.1001/jamainternmed.2020.2975

14. Swenning T, Leighton J, Nentwig M, Dart B. Hip fracture care and national systems: the United States and Canada. *OTA Int*. 2020;3(1):e073. doi:10.1097/OI9.0000000000000073

15. Organisation for Economic Co-operation and Development. OECD data: elderly population. Accessed April 30, 2022. https://data.oecd.org/pop/elderly-population.htm

16. Leslie WD, O'Donnell S, Jean S, et al; Osteoporosis Surveillance Expert Working Group. Trends in hip fracture rates in Canada. *JAMA*. 2009;302(8):883-889. doi:10.1001/jama.2009.1231

17. Cooper C, Campion G, Melton LJ III. Hip fractures in the elderly: a world-wide projection. *Osteoporos Int*. 1992;2(6):285-289. doi:10.1007/BF01623184

18. Harvey N, Dennison E, Cooper C. Osteoporosis: impact on health and economics. *Nat Rev Rheumatol*. 2010;6(2):99-105. doi:10.1038/nrrheum.2009.260

19. Carpintero P, Caeiro JR, Carpintero R, Morales A, Silva S, Mesa M. Complications of hip fractures: a review. *World J Orthop*. 2014;5(4):402-411. doi:10.5312/wjo.v5.i4.402

20. Prieto-Alhambra D, Moral-Cuesta D, Palmer A, et al. The impact of hip fracture on health-related quality of life and activities of daily living: the SPARE-HIP prospective cohort study. *Arch Osteoporos*. 2019;14(1):56. doi:10.1007/s11657-019-0607-0

21. Morri M, Ambrosi E, Chiari P, et al. One-year mortality after hip fracture surgery and prognostic factors: a prospective cohort study. *Sci Rep*. 2019;9(1):18718. doi:10.1038/s41598-019-55196-6

22. Pincus D, Ravi B, Wasserstein D, et al. Association between wait time and 30-day mortality in adults undergoing hip fracture surgery. *JAMA*. 2017;318(20):1994-2003. doi:10.1001/jama.2017.17606

23. Bhandari M, Swiontkowski M. Management of acute hip fracture. *N Engl J Med*. 2017;377(21):2053-2062. doi:10.1056/NEJMcp1611090

24. HIP ATTACK Investigators. Accelerated surgery versus standard care in hip fracture (HIP ATTACK): an international, randomised, controlled trial. *Lancet*. 2020;395(10225):698-708. doi:10.1016/S0140-6736(20)30058-1

25. Taylor-Phillips S, Stinton C. Fatigue in radiology: a fertile area for future research. *Br J Radiol*. 2019;92(1099):20190043. doi:10.1259/bjr.20190043

26. Bates DW, Levine D, Syrowatka A, et al. The potential of artificial intelligence to improve patient safety: a scoping review. *NPJ Digit Med*. 2021;4(1):54. doi:10.1038/s41746-021-00423-6

27. Choudhury A, Asan O. Role of artificial intelligence in patient safety outcomes: systematic literature review. *JMIR Med Inform*. 2020;8(7):e18599. doi:10.2196/18599

28. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372(71):n71. doi:10.1136/bmj.n71

29. Cheng CT, Ho TY, Lee TY, et al. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. *Eur Radiol*. 2019;29(10):5469-5477. doi:10.1007/s00330-019-06167-y

30. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. *Skeletal Radiol*. 2019;48(2):239-244. doi:10.1007/s00256-018-3016-3

31. Adams M, Chen W, Holcdorf D, McCusker MW, Howe PD, Gaillard F. Computer vs human: deep learning versus perceptual training for the detection of neck of femur fractures. *J Med Imaging Radiat Oncol*. 2019;63(1):27-32. doi:10.1111/1754-9485.12828

32. Mawatari T, Hayashida Y, Katsuragawa S, et al. The effect of deep convolutional neural networks on radiologists' performance in the detection of hip fractures on digital pelvic radiographs. *Eur J Radiol*. 2020;130:109188. doi:10.1016/j.ejrad.2020.109188

33. Jiménez-Sánchez A, Kazi A, Albarqouni S, et al. Precise proximal femur fracture classification for interactive training and surgical planning. *Int J Comput Assist Radiol Surg*. 2020;15(5):847-857. doi:10.1007/s11548-020-02150-x

34. Yu JS, Yu SM, Erdal BS, et al. Detection and localisation of hip fractures on anteroposterior radiographs with artificial intelligence: proof of concept. *Clin Radiol*. 2020;75(3):237.e1-237.e9. doi:10.1016/j.crad.2019.10.022

35. Kitamura G. Deep learning evaluation of pelvic radiographs for position, hardware presence, and fracture detection. *Eur J Radiol*. 2020;130:109139. doi:10.1016/j.ejrad.2020.109139

36. Krogue JD, Cheng KV, Hwang KM, et al. Automatic hip fracture identification and functional subclassification with deep learning. *Radiol Artif Intell*. 2020;2(2):e190023. doi:10.1148/ryai.2020190023

37. Yamada Y, Maki S, Kishida S, et al. Automated classification of hip fractures using deep convolutional neural networks with orthopedic surgeon-level accuracy: ensemble decision-making with antero-posterior and lateral radiographs. *Acta Orthop*. 2020;91(6):699-704. doi:10.1080/17453674.2020.1803664

38. Mutasa S, Varada S, Goel A, Wong TT, Rasiej MJ. Advanced deep learning techniques applied to automated femoral neck fracture detection and classification. *J Digit Imaging*. 2020;33(5):1209-1217. doi:10.1007/s10278-020-00364-8

39. Beyaz S, Açıcı K, Sümer E. Femoral neck fracture detection in X-ray images using deep learning and genetic algorithm approaches. *Jt Dis Relat Surg*. 2020;31(2):175-183. doi:10.5606/ehc.2020.72163

40. Açıcı K, Sümer E, Beyaz S. Comparison of different machine learning approaches to detect femoral neck fractures in x-ray images. *Health Technol*. 2021;11(3):643-653. doi:10.1007/s12553-021-00543-9

41. Bae J, Yu S, Oh J, et al. External validation of deep learning algorithm for detecting and visualizing femoral neck fracture including displaced and non-displaced fracture on plain x-ray. *J Digit Imaging*. 2021;34(5):1099-1109. doi:10.1007/s10278-021-00499-2

42. Cheng CT, Wang Y, Chen HW, et al. A scalable physician-level deep learning algorithm detects universal trauma on pelvic radiographs. *Nat Commun*. 2021;12(1):1066. doi:10.1038/s41467-021-21311-3

43. Guy S, Jacquet C, Tsenkoff D, Argenson JN, Ollivier M. Deep learning for the radiographic diagnosis of proximal femur fractures: limitations and programming issues. *Orthop Traumatol Surg Res*. 2021;107(2):102837. doi:10.1016/j.otsr.2021.102837

44. Twinprai N, Boonrod A, Boonrod A, et al. Artificial intelligence (AI) vs. human in hip fracture detection. *Heliyon*. 2022;8(11):e11266. doi:10.1016/j.heliyon.2022.e11266

45. Murphy EA, Ehrhardt B, Gregson CL, et al. Machine learning outperforms clinical experts in classification of hip fractures. *Sci Rep*. 2022;12(1):2058. doi:10.1038/s41598-022-06018-9

46. Liu P, Lu L, Chen Y, et al. Artificial intelligence to detect the femoral intertrochanteric fracture: the arrival of the intelligent-medicine era. *Front Bioeng Biotechnol*. 2022;10:927926. doi:10.3389/fbioe.2022.927926

47. Ottenbacher KJ, Linn RT, Smith PM, Illig SB, Mancuso M, Granger CV. Comparison of logistic regression and neural network analysis applied to predicting living setting after hip fracture. *Ann Epidemiol*. 2004;14(8):551-559. doi:10.1016/j.annepidem.2003.10.005

48. Sund R, Riihimäki J, Mäkelä M, et al. Modeling the length of the care episode after hip fracture: does the type of fracture matter? *Scand J Surg*. 2009;98(3):169-174. doi:10.1177/145749690909800308

49. Lin CC, Ou YK, Chen SH, Liu YC, Lin J. Comparison of artificial neural network and logistic regression models for predicting mortality in elderly patients with hip fracture. *Injury*. 2010;41(8):869-873. doi:10.1016/j.injury.2010.04.023

50. Shi L, Wang XC, Wang YS. Artificial neural network models for predicting 1-year mortality in elderly patients with intertrochanteric fractures in China. *Braz J Med Biol Res*. 2013;46(11):993-999. doi:10.1590/1414-431X20132948

51. Karnuta JM, Navarro SM, Haeberle HS, Billow DG, Krebs VE, Ramkumar PN. Bundled care for hip fractures: a machine-learning approach to an untenable patient-specific payment model. *J Orthop Trauma*. 2019;33(7):324-330. doi:10.1097/BOT.0000000000001454

52. Chen CY, Chen YF, Chen HY, Hung CT, Shi HY. Artificial neural network and Cox regression models for predicting mortality after hip fracture surgery: a population-based comparison. *Medicina (Kaunas)*. 2020;56(5):243. doi:10.3390/medicina56050243

53. DeBaun MR, Chavez G, Fithian A, et al. Artificial neural networks predict 30-day mortality after hip fracture: insights from machine learning. *J Am Acad Orthop Surg*. 2021;29(22):977-983. doi:10.5435/JAAOS-D-20-00429

54. Zhang Y, Huang L, Liu Y, Chen Q, Li X, Hu J. Prediction of mortality at one year after surgery for pertrochanteric fracture in the elderly via a Bayesian belief network. *Injury*. 2020;51(2):407-413. doi:10.1016/j.injury.2019.11.029

55. Cao Y, Forssten MP, Mohammad Ismail A, et al. Predictive values of preoperative characteristics for 30-day mortality in traumatic hip fracture patients. *J Pers Med*. 2021;11(5):353. doi:10.3390/jpm11050353

56. Cowling TE, Cromwell DA, Bellot A, Sharples LD, van der Meulen J. Logistic regression and machine learning predicted patient mortality from large sets of diagnosis codes comparably. *J Clin Epidemiol*. 2021;133:43-52. doi:10.1016/j.jclinepi.2020.12.018

57. Forssten MP, Bass GA, Ismail AM, Mohseni S, Cao Y. Predicting 1-year mortality after hip fracture surgery: an evaluation of multiple machine learning approaches. *J Pers Med*. 2021;11(8):727. doi:10.3390/jpm11080727

58. Oosterhoff JHF, Karhade AV, Oberai T, Franco-Garcia E, Doornberg JN, Schwab JH. Prediction of postoperative delirium in geriatric hip fracture patients: a clinical prediction model using machine learning algorithms. *Geriatr Orthop Surg Rehabil*. Published online December 13, 2021. doi:10.1177/21514593211062277

59. Li Y, Chen M, Lv H, Yin P, Zhang L, Tang P. A novel machine-learning algorithm for predicting mortality risk after hip fracture surgery. *Injury*. 2021;52(6):1487-1493. doi:10.1016/j.injury.2020.12.008

60. Cary MP Jr, Zhuang F, Draelos RL, et al. Machine learning algorithms to predict mortality and allocate palliative care for older patients with hip fracture. *J Am Med Dir Assoc*. 2021;22(2):291-296. doi:10.1016/j.jamda.2020.09.025

61. Shtar G, Rokach L, Shapira B, Nissan R, Hershkovitz A. Using machine learning to predict rehabilitation outcomes in postacute hip fracture patients. *Arch Phys Med Rehabil*. 2021;102(3):386-394. doi:10.1016/j.apmr.2020.08.011

62. Zhong H, Wang B, Wang D, et al. The application of machine learning algorithms in predicting the length of stay following femoral neck fracture. *Int J Med Inform*. 2021;155:104572. doi:10.1016/j.ijmedinf.2021.104572

63. Xing F, Luo R, Liu M, Zhou Z, Xiang Z, Duan X. A new random forest algorithm-based prediction model of post-operative mortality in geriatric patients with hip fractures. *Front Med (Lausanne)*. 2022;9:829977. doi:10.3389/fmed.2022.829977

64. Harris AHS, Trickey AW, Eddington HS, et al. A tool to estimate risk of 30-day mortality and complications after hip fracture surgery: accurate enough for some but not all purposes? a study from the ACS-NSQIP Database. *Clin Orthop Relat Res*. 2022;480(12):2335-2346. doi:10.1097/CORR.0000000000002294

65. Oosterhoff JHF, Savelberg ABMC, Karhade AV, et al. Development and internal validation of a clinical prediction model using machine learning algorithms for 90 day and 2 year mortality in femoral neck fracture patients aged 65 years or above. *Eur J Trauma Emerg Surg*. 2022;48(6):4669-4682. doi:10.1007/s00068-022-01981-4

66. Lei M, Han Z, Wang S, et al. A machine learning-based prediction model for in-hospital mortality among critically ill patients with hip fracture: an internal and external validated study. *Injury*. 2023;54(2):636-644. doi:10.1016/j.injury.2022.11.031

67. Kitcharanant N, Chotiyarnwong P, Tanphiriyakun T, et al. Development and internal validation of a machine-learning-developed model for predicting 1-year mortality after fragility hip fracture. *BMC Geriatr*. 2022;22(1):451. doi:10.1186/s12877-022-03152-x

68. Dominguez S, Liu P, Roberts C, Mandell M, Richman PB. Prevalence of traumatic hip and pelvic fractures in patients with suspected hip fracture and negative initial standard radiographs: a study of emergency department patients. *Acad Emerg Med*. 2005;12(4):366-369. doi:10.1197/j.aem.2004.10.024

69. Pincus D, Wasserstein D, Ravi B, et al. Medical costs of delayed hip fracture surgery. *J Bone Joint Surg Am*. 2018;100(16):1387-1396. doi:10.2106/JBJS.17.01147

70. Stefanidis D, Fanelli RD, Price R, Richardson W; SAGES Guidelines Committee. SAGES guidelines for the introduction of new technology and techniques. *Surg Endosc*. 2014;28(8):2257-2271. doi:10.1007/s00464-014-3587-6

71. He H, Bai Y, Garcia EA, Li S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE; 2008:1322-1328. doi:10.1109/IJCNN.2008.4633969

72. Marsh D, Currie C, Brown P, et al. *The Care of Patients With Fragility Fracture*. British Orthopaedic Association; 2007.

73. Grant-Freemantle MC, Kenyon RM, Gibbons J, Flynn SO, Davey M, Burke N. Assessing the time to ward transfer in patients presenting to the emergency department with an acute hip fracture: a closed-loop audit. *Cureus*. 2020;12(1):e6794. doi:10.7759/cureus.6794

74. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. *Radiol Artif Intell*. 2022;4(3):e210064. doi:10.1148/ryai.210064

75. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(7):e048008. doi:10.1136/bmjopen-2020-048008

SUPPLEMENT 1.

eFigure 1. PRISMA Flowchart of Included Studies
eFigure 2. Depiction of Typical Machine Learning Models Used for Diagnosing Hip Fractures From Medical Imaging, Representing a Convolutional Neural Network (A), and for Predicting Postoperative Patient Clinical Outcomes From a Multilayer Perceptron (B)
eTable 1. Machine Learning Algorithm Properties of Studies on Hip Fracture Diagnosis Using Plain Radiographs
eTable 2. Databases Used for Outcome Prediction With Availability
eTable 3. Summary of Outcome Performance of Artificial Intelligence Models for Diagnosing Hip Fractures Based on Plain-Film Radiographs
eTable 4. Common Features Used for Postoperative Outcome Prediction by Each Study
eTable 5. Comparison of Area Under the Curve Values Between Machine Learning Models and Traditional Statistical (Multivariable Logistic or Linear Regression) Prediction Models

SUPPLEMENT 2.

Data Sharing Statement