REVIEW

JBMR®

Machine Learning Solutions for Osteoporosis—A Review

Julien Smets,[1] Enisa Shevroja,[1] Thomas Hügle,[2] William D Leslie,[3] and Didier Hans[1]

[1]Center of Bone Diseases, Bone and Joint Department, Lausanne University Hospital, Lausanne, Switzerland
[2]Department of Rheumatology, Lausanne University Hospital, Lausanne, Switzerland
[3]University of Manitoba, Winnipeg, Canada

ABSTRACT

Osteoporosis and its clinical consequence, bone fracture, is a multifactorial disease that has been the object of extensive research. Recent advances in machine learning (ML) have enabled the field of artificial intelligence (AI) to make impressive breakthroughs in complex data environments where human capacity to identify high-dimensional relationships is limited. The field of osteoporosis is one such domain, notwithstanding technical and clinical concerns regarding the application of ML methods. This qualitative review is intended to outline some of these concerns and to inform stakeholders interested in applying AI for improved management of osteoporosis. A systemic search in PubMed and Web of Science resulted in 89 studies for inclusion in the review. These covered one or more of four main areas in osteoporosis management: bone properties assessment (n = 13), osteoporosis classification (n = 34), fracture detection (n = 32), and risk prediction (n = 14). Reporting and methodological quality was determined by means of a 12-point checklist. In general, the studies were of moderate quality with a wide range (mode score 6, range 2 to 11). Major limitations were identified in a significant number of studies. Incomplete reporting, especially over model selection, inadequate splitting of data, and the low proportion of studies with external validation were among the most frequent problems. However, the use of images for opportunistic osteoporosis diagnosis or fracture detection emerged as a promising approach and one of the main contributions that ML could bring to the osteoporosis field. Efforts to develop ML-based models for identifying novel fracture risk factors and improving fracture prediction are additional promising lines of research. Some studies also offered insights into the potential for model-based decision-making. Finally, to avoid some of the common pitfalls, the use of standardized checklists in developing and sharing the results of ML models should be encouraged. © 2021 American Society for Bone and Mineral Research (ASBMR).

KEY WORDS: OSTEOPOROSIS; FRACTURE PREDICTION; RISK ASSESSMENT; MACHINE LEARNING; ARTIFICIAL INTELLIGENCE

Introduction

The goal of "teaching" machines to behave as if they are thinking, and thus appear intelligent, is commonly known as artificial intelligence (AI). Technically, AI comprises a wide variety of algorithmic tools intended to mimic human reasoning. Machine Learning (ML), a subfield of AI, involves the use of statistical methods to identify empirical patterns ("learn") from data (Fig. 1).[1–3] Specifically, ML uses the data as examples to guide learning process toward a given goal. Deep Learning (DL), a subfield of ML, provides multilayered ("deep") model architectures able to handle complex non-linear relationships between the input and output variables.[4] With the Big Data Era and improvement of computing power, AI has revolutionized industries and, relatively recently, these approaches have been implemented in the field of medicine to find solutions to complex medical scenarios and multifactorial conditions. Significant successes have been found in diagnostic imaging,[5–7] disease risk prediction,[8,9] optimal treatment strategies decision-making,[10] and prediction of molecular shape or activity.[11,12]

Osteoporosis is a complex disease in which the quantity and quality of bone are diminished, causing an increase in bone fragility. The clinical outcome of osteoporosis—fracture—occurs with an incidence of 8.9 million worldwide every year with considerable health, societal, and economic burden.[13] Osteoporosis diagnosis is primarily based on bone mineral density (BMD) as assessed by dual-energy X-ray absorptiometry (DXA), but this does not capture the important contributions of clinical risk factors or other bone measures (eg, trabecular bone score, geometry). Preventing fractures is the main purpose in osteoporosis management. The main clinical risk factors for osteoporotic fractures and altered bone metabolism include older age, sex (with predisposition for women), ethnicity, heredity, previous fracture, malnutrition, alcohol consumption, current smoking, vitamin D deficiency, physical inactivity, various medications and medical disorders. FRAX is the most widely used and validated clinical tool for fracture prediction as based on many of the clinical risk factors referred to earlier combined with/without BMD.[14] FRAX determines the 10-year probability of having a fracture by using classical statistical tools and is calibrated from country-specific

Received in original form October 28, 2020; revised form February 4, 2021; accepted March 16, 2021. Accepted manuscript online March 21, 2021.
Address correspondence to: Didier Hans, PhD, Center of Bone Diseases, Bone and Joint Department (DAL – RHU) Lausanne University Hospital (CHUV) and University of Lausanne, Av. Pierre Decker 4, 1011 Lausanne, Switzerland. E-mail: didier.hans@chuv.ch; didier.hans@ascendys.ch
Additional Supporting Information may be found in the online version of this article.

Journal of Bone and Mineral Research, Vol. 36, No. 5, May 2021, pp 833–851.
DOI: 10.1002/jbmr.4292

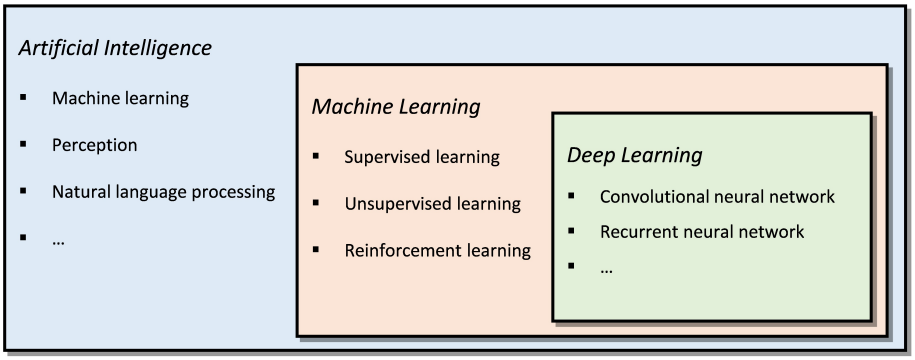


Fig 1. Hierarchical classification with examples of artificial intelligence, machine learning, and deep learning.

data. AI/ML is finding novel application in the deeper investigation of osteoporosis, including diagnosis and fracture prediction from biological testing, imaging, or clinical data.

In this article, we review the state-of-the-art of ML methods and their utility in osteoporosis diagnosis and fracture prediction. We aim to provide a general understanding of the ML methods used and address some of the methodological issues of ML in the medical context. The structure of the article is as follows: (i) a brief overview of ML methods, models, and their optimal practice; (ii) a literature review of ML use in osteoporosis, including bone properties assessment, osteoporosis diagnosis, and fracture detection and prediction; and iii) an overall discussion of issues of the ML application and interpretation identified from this review.

Machine Learning in Practice

ML is typically categorized as supervised learning, unsupervised learning, and reinforcement learning. Supervised learning uses a given set of input features and one or more outcomes as the basis for model training. The learning is accomplished by tuning internal model parameters. The model is iteratively trained to minimize prediction error when comparing samples drawn from the data with a target reference standard, also called ground truth. Unsupervised learning does not use any labeling information and aims at grouping data by shared properties. Practically, it helps to discover structure in the data such as identifying clusters of patients at similar risk or selecting variables most strongly correlated with an outcome. Reinforcement learning is based on a decision-making sequence with a long-term goal, which makes it well suited for treatment management in health care.[10] Supervised learning to predict an outcome is the most widely used ML approach at present. Thus, in the explanatory section of this article, we focus on the supervised learning. Fig. 2 presents the steps of a typical ML pipeline through which a learning-based model is trained to identify a target health condition.

Task definition

Conceptually, the definition of the task is the initial step of the ML workflow. In bone health, this task may consist of the identification of a fracture or osteoporosis diagnosis. The availability of large data sets is crucial for optimal ML performance. Ensuring high-quality data collection, labeling, and processing is of equal importance.

Train/test split

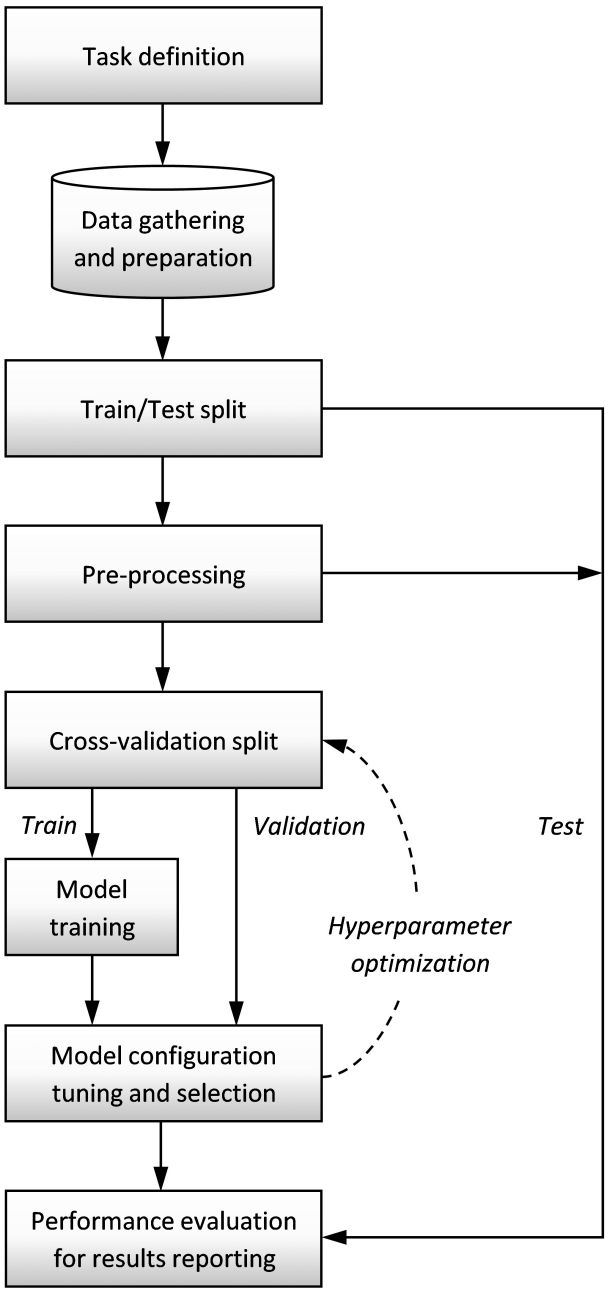
The data processing stage starts with the separation of the available data set into two distinct subsets: the subset from which the model will be developed (train set) and the subset on which this model will be validated and which must remain unseen (test set) (Supplemental Fig. S1). This early split avoids the possible biases introduced during the development of the model.

Pre-processing

The pre-processing stage is usually applied to the data to ensure that it is consistently formatted for the learning. Pre-processing consists of data modifications such as normalization (for continuous data), "one-hot" encoding (for categorical data), and image reshaping. Approaches are used to avoid overfitting when the quantity of data is limited, such as data augmentation, which creates copies of the initial data with random variations (eg, position, rotation, orientation), or synthetic data generation. Model development in itself consists of three main steps: cross-validation, model training, and performance estimation.

Cross-validation

To ensure replication and generalization, ML must learn an accurate representation of the data.[15] If simple models fail in mapping complex relationships between the inputs and output(s), namely underfitting, overly complex models lead to overfitting by paying excessive attention to specific details in the data. This trade-off is in part controlled by training multiple model configurations using different training and/or pre-processing parameters, called hyperparameters, which affect the model's ability to learn and therefore its final performance (cf. dotted line in Fig. 2). Cross-validation involves dividing the data set into intermediate subsets. For example, a validation set, extracted from the train set, helps in tuning model configuration and in selecting the best-performing architecture and hyperparameters. K-fold cross-validation can be used in estimating a model's accuracy and variability by averaging performances across repeated splits. Other methods commonly used include regularization to limit the model's ability to memorize details specific to the data set, data augmentation as defined above, pre-training which uses a model trained in a different context,[16] or ensembling by combining multiple models to increase robustness.



Task definition: ML project starts with a definition of a target outcome and a given set of inputs as predictors.
Data preparation: Collection, formatting, selection and preparation of the data.
Train set: Subset used to develop model by tuning internal parameters with respect to the ground truth labels (e.g. disease presence).
Test set: Subset isolated from the training set to validate model's performance and report results. This sample must be unseen by the model to remain unbiased.
Pre-processing: The processing and transformation of the input data into features (standardization, etc.) by using the training set. Some of these steps can subsequently be applied on the test set.
Cross-validation split: A range of techniques for estimating the performance of a model configuration on unknown data by splitting the data into train and validation subsets. These are often applied to configure model hyperparameters and select the best configuration (tiled line). One common method is K-fold cross-validation that averages performance on multiple run splits. In addition, it can be used for train/test split with caution given the risk of overfitting.
Model training: Tune model parameters using examples from the training set to learn some representation in inferring the desired outcome.
Performance evaluation: Assess model's accuracy using the test subset for performance comparison.

Fig 2. Model development flow chart for supervised learning.

Model training

As described above, the training set is used to optimize the model's internal parameters with respect to the ground truth. Linear and logistic regressions are classical statistical tools used in research and are simple forms of supervised learning. More complex algorithms were then developed such as tree-based models (eg, classification and regression tree [CART]), which provide a straightforward visualization of the decision-making process,[17] or artificial neural networks (ANN). More sophisticated DL models such as convolutional neural networks (CNN) (Supplemental Fig. S2) and recurrent neural networks take advantage of strong local and temporal correlations in the data, and are particularly suitable for medical images[4,18] and longitudinal health records, respectively.

Model performance evaluation

ML studies commonly report performance metrics such as sensitivity, specificity, accuracy, and area under the curve of the receiver operating characteristic (AUC). When several models

are under consideration, the model with the best performance on the validation set is selected. This procedure may be iterative in a search for the optimal architecture and training hyperparameters. Once training has been completed, the final model is applied to the test set and performance measures from the test set are reported. Visualizing a model's decisions with techniques such as feature importance (SHAP[19]) or visual heat maps (Grad-Cam[20]) contributes to their validation in a qualitative manner.

Literature Review

Literature selection

A literature search was performed in the electronic databases of PubMed and Web of Science. Studies published from January 2015 until December 17, 2020, were included. A restriction for English language has been applied. Searches relevant to the use of AI/ML methods to measure, detect, classify, or predict outcomes related to osteoporosis or fracture were made. The search strategy is shown in detail in the Supplemental Material. Records identified from searches were de-duplicated and the titles and abstracts were screened for inclusion by JS using Rayyan.[21] Full-text records were collated for the articles chosen to be included and were further screened by the same author. We performed a qualitative synthesis of the included studies because the wide diversity of objectives, methods, and metrics precluded a quantitative approach or formal meta-analysis.

Quality assessment of the studies

A simplified version of the MI-CLAIM (minimum information about clinical artificial intelligence modeling) checklist was employed for the quality assessment of the eligible studies.[22] This checklist is shown in the Supplemental Table S1. It consists of 12 items across six main quality-assessment domains: study design; data preparation and partitioning; model development, optimization, and final model selection; model performance; model examination; and reproducibility and transparency of the study. For each item, a value of ''complete,'' ''incomplete,'' or ''non-applicable'' was assigned. The final numerical quality score for each study was the sum of these 12 items, where ''complete'' was counted as 1 and ''incomplete'' as 0 (maximum value of 12). The quality assessment tables are available in Supplemental Tables S2 to S5.

Results

Literature selection

The literature searches in PubMed and Web of Science identified 486 records. After the exclusion of duplicates, defined as citations found in both databases, 319 records were screened. A total of 89 of these records were considered eligible for inclusion in the current literature review. A detailed flow diagram of the results from the literature search and the study selection is shown in Fig. 3.

General characteristics of the studies

The 89 articles included in this narrative review fell into four broad areas (Fig. 4): bone properties assessment (13 studies), osteoporosis diagnosis (34 studies), fracture detection (32 studies), and risk prediction (14 studies). Four studies addressed for than one area. The general characteristics and results of the included studies are shown in Tables 1 to 4, grouped according to these areas.

The majority of the studies, 78% (69/89), were published after 2018. A progressive increase over time with 12, 12, 25, and 32 studies published in 2017, 2018, 2019, and 2020, respectively, was observed. Asia, Europe, and USA contributed to 79% of the studies. The top three countries were USA (19/89, 21%), South Korea (10/89, 11%), and China (8/89, 9%). In 36 of the 89 studies, images were used as input, whereas in others, data from varying sources were used. The three most common data sources were X-rays (29/89, 33%), database (23/89, 26%), and computed tomography (CT) (18/89, 20%). DXA modality was studied in only five studies, although some others have used databases with BMD assessed by DXA scans. Almost all studies applied directly to images (26/29) used DL models. AUC was the most frequently reported metric (56/89, 63%).

In regard to the studies' quality, the studies on bone properties assessment presented an average quality score of mode 8 of 11 points (range 5 to 10), studies on osteoporosis classification had an average quality score of mode 7 of 11 (range 2 to 9), studies of fracture detection 8 of 12 (range 5 to 11), and studies on supervised risk prediction had an average quality score of 9 of 12 (range 6 to 10). The 12 points being evaluated through the quality assessment score were not always applicable in each research task. For example, 7 points were applicable for the studies on unsupervised risk prediction. In general, from the 12 points being evaluated during the quality assessment of each study, the most absent were the study reproducibility and transparency, the reporting of the examination technique use, the use of a state-of-the-art method, and the model configuration.

Studies on bone properties assessment

Thirteen studies investigated bone properties such as vertebral fracture load,[23–25] microarchitecture parameters,[26,27] vertebral height,[28] or BMD[29–35] (Table 1). The main objective of these efforts was to improve the diagnosis of osteoporosis. Vertebral fracture load was assessed from donors[24,25] or using finite element analysis.[23] Microarchitecture parameters were determined using simulations or data collected from human cadavers.[26,27] Among the BMD studies, five assessed vertebral BMD,[30–34] one vertebral and hip BMD,[29] and one total body BMD.[35] Internal validation was performed in each study. External validation was performed in two studies from different local institutions in the same country with comparable results.[23,32] Feature selection,[23] data augmentation,[32] and transfer learning from pre-trained models[28,30] were applied to control overfitting. However, critical limitations in the use of ML methods were present: the gold standard, BMD as estimated from DXA, was not used as the ground truth (BMD from DXA is the reference technique used in practice, whereas BMD as estimated from CT scans was used as the ground truth);[30–32,34] model parameter selection was not reported;[23–25,29,33,35] discussion of crucial limitations regarding generalization was absent;[24,25,27,28,32,35] and the inclusion of the training set in model evaluation[33] or the use of 25 times more input features than data,[35] both indicating an increased risk of overfitting. Of note, Pang and colleagues reported their methods in detail and shared the code used for their analysis.[28]

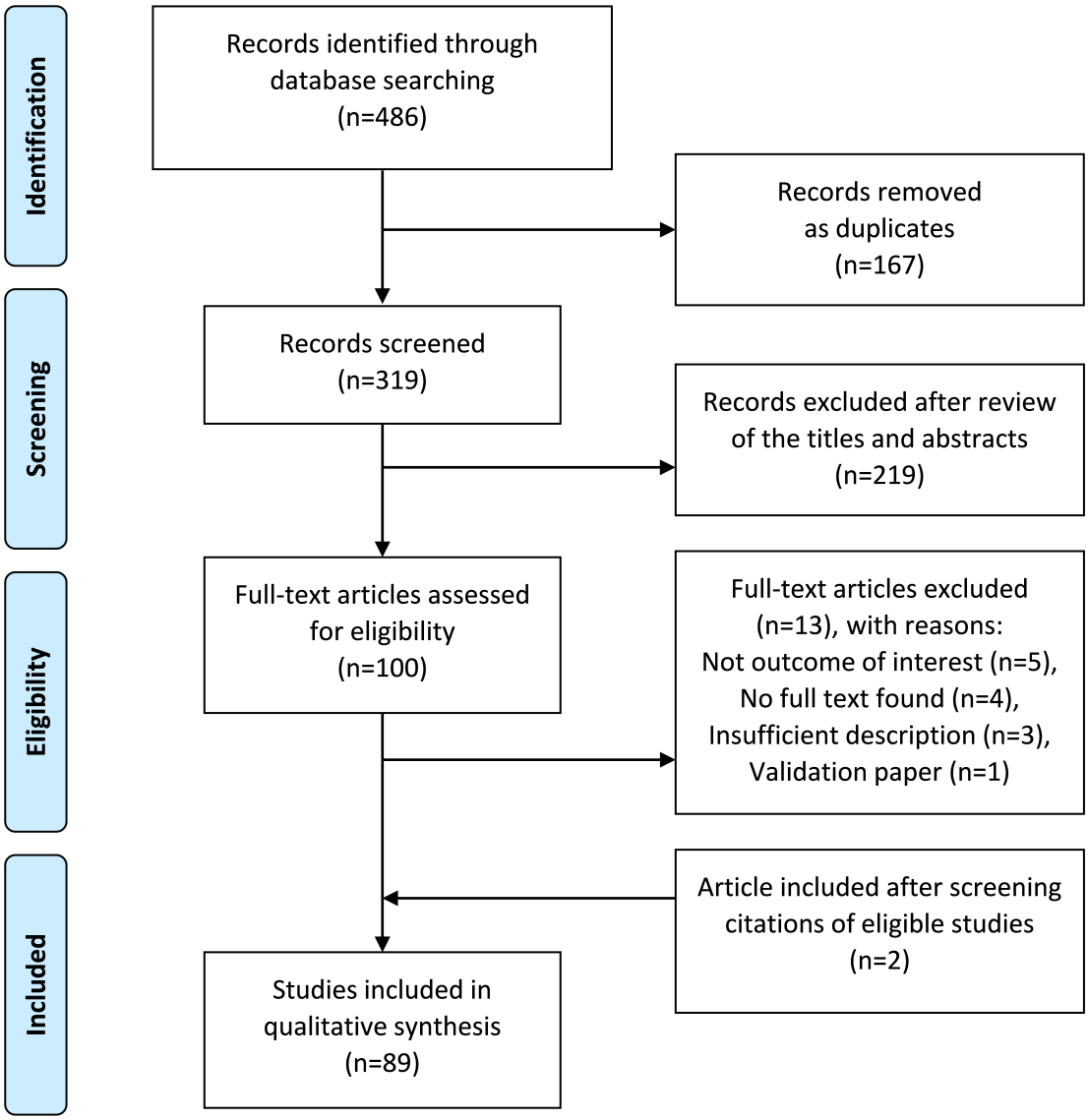


Fig 3. Flow chart of the literature selection in PubMed and Web of Science.

Studies on osteoporosis classification

Thirty-four studies explored osteoporosis diagnosis from data and images (Table 2).[29,32–34,36–65] Osteoporosis classification was made based on lumbar BMD,[32–34,37,51] hip BMD,[38,50,58] lumbar and hip BMD,[29,39–42,46–48,53,59,60] other non-standard assessments,[43,44,49,54–56,65] or unspecified.[36,45,52,57,61–64] Studies identified osteoporosis based on opportunistic imaging from CT,[32–34] X-ray,[37,38,43–45,55–59,63,64] or dental imaging;[36,47–49,53,54,60,62] other studies used data from patient characteristics,[40,41,50,51,61,65] bone biomarkers,[29,39] or acoustical responses.[42,52] As outcome, studies classified osteoporotic versus normal patients,[29,36,39,40,43,49,50,52,54–57,62] osteoporotic versus non-osteoporotic patients (based on a BMD *T*-score threshold of –2.5 SD),[34,38,44,64] normal versus abnormal subjects (based on the BMD *T*-score threshold of –1 SD),[33,41,42,45,47,48,58–60,65] experimented multiple classifications,[46,63] or assigned to three classes: osteoporosis (BMD *T*-score ≤ –2.5 SD), osteopenia (–2.5 < BMD *T*-score ≤ –1), and normal (BMD *T*-score > –1 SD).[32,37,51,61] The models were internally validated in almost each study, and in two of them the models were also externally validated.[32,50] Twelve studies validated their model using accuracy with an average performance of 90.1% (range 70.0% to 98.9%); 22 validated their model using AUC with a mean of 0.90 (range 0.74 to 1.00). Surprisingly, six reported near perfect AUCs (AUC ≥0.99),[29,43,48,54,56,60] indicating potential overfitting with risk for poor generalization. Other studies presented a clear risk of overfitting by using data with important bias between case/control groups,[54,58] model selection based on the testing set,[29,40,48,51,60] reporting high discrepancies between training and test sets,[37,59] or including a part of the training data set in the testing data.[33,63,65] Several studies did not report characteristics of their data set[48,55,56,59–64] or the model selection process.[33,39,41,42,45,46,50,51,53,56–61,64,65] Performance was significantly impacted by case prevalence where accuracy

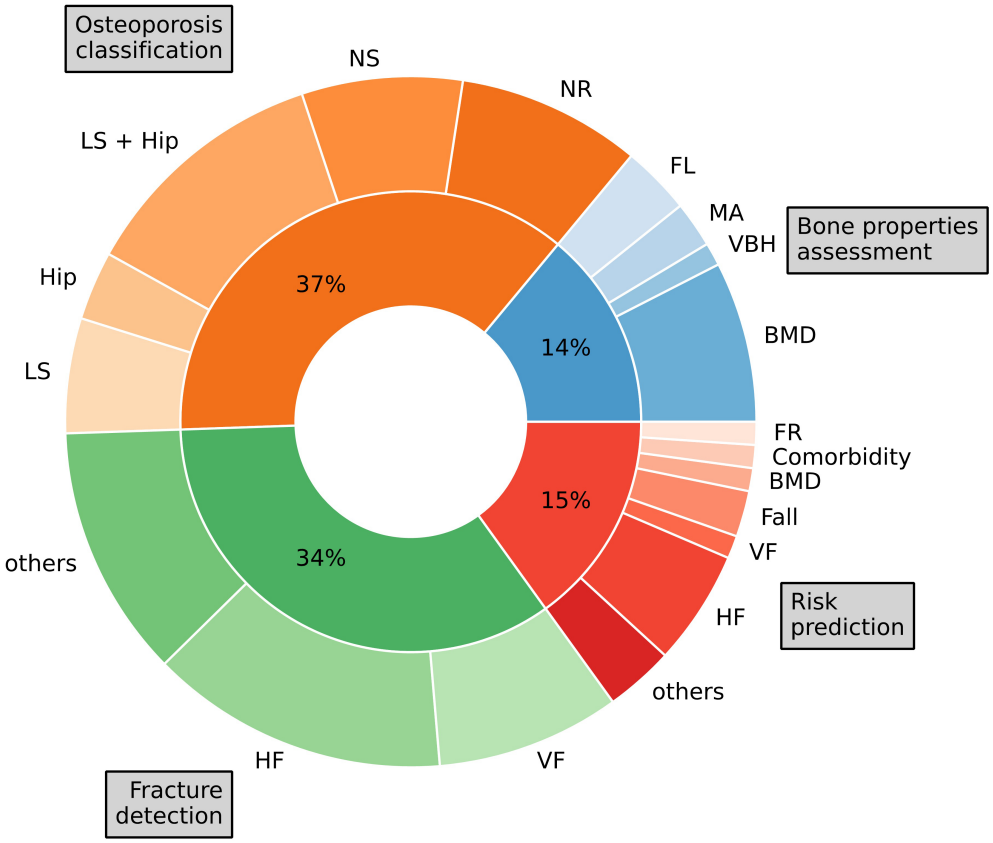


Fig 4. Study tasks related to osteoporosis investigated in this review. Bone properties assessment included estimations of fracture load (FL), microarchitecture (MA), vertebral bone height (VBH), and bone mineral density (BMD). Osteoporosis classification was performed based on lumbar spine (LS), hip, lumbar spine and hip, non-standard assessment technique (NS), or unspecified (NR). Fracture detection has been studied in the vertebrae (VF), the hip (HF), or different body sites. Risk prediction involved the risks of bone density loss (BMD), fragility fractures (FR, VF, HF), falls, or comorbidities.

dropped from 94.0% to 88.4% when tested on 13% and 50% positive (osteoporotic) cases, respectively.[44] An image enhancement and standardization step, and combining multiple features, was able to considerably improve results in two studies, respectively.[43,61] Overfitting was addressed by cropping images into regions,[32,34,37,38,44–46,53,57,59] dimensionality reduction to select best features,[39,40,43,56,64] data augmentation,[40,54] or by using a previously trained model.[38,49,56] Curiously, a simple base model using age and body mass index (BMI) did not perform better than random guess (AUC ≈ 0.5), suggesting that the population had undergone a selection process resulting in an imbalance for these variables.[29] One study proposed a simple decision rule set to diagnose osteoporosis using a categorized version of the features, thus helping the understanding of the decision-making process.[48] The regions of interest for DL model decision-making were visualized using class activation maps (Grad-CAM[20]) to improve its interpretation.[36,38]

Studies on fracture detection

Among the 32 studies that investigated fracture detection (Table 3),[66–97] 11 were on vertebral fractures,[66–76] 17 hip fractures,[74–90] and 10 other fracture sites such as humerus or wrist.[75,76,90–97] Nineteen studies developed CNN models for image analysis.[66,67,71,72,77–86,90,91,93–95] Others used features extracted from images or collected from non-imaging data.[68–70,73–77,87–89,92,96,97] Studies reported average best AUC of 0.92 (range 0.63 to 1.00) and average best accuracy of 89.8% (range 78.4% to 99.1%]. Surprising findings were that hospital-related variables, such as the scanner device model, were better predictors of fractures than patients' characteristics or their images (AUC = 0.89 [95% confidence interval (CI) 0.87–0.91] versus 0.79 [95% CI 0.75–0.82] and 0.78 [95% CI 0.74–0.81], respectively),[77] which the authors acknowledged as a potential bias induced by the triage process. Their model was evaluated on unbalanced and balanced cases, 3% to 50% hip fractures, and demonstrated a considerable decrease in area under the precision-recall curve (AUPRC), from 0.74 to 0.11, despite similar AUCs.[77] Twelve studies compared ML model performance versus human experts.[66,71,78–83,85,91,93,94] In four of these studies, ML outperformed human experts significantly.[80,83,85,91] Thirteen studies applied transfer learning based on pre-defined CNN architectures, pre-trained on the ImageNet data set[77,79–83,85,90,91,93,95] or on a radiography image database.[78,94] Some of these were fine-tuned on the last layers,[82] on the penultimate layer,[77,79,83,93,95] or were unclear.[78,80,81,85,90,91,94] Model decision-making was highlighted using heat maps such as Grad-CAM.[66,67,78,79,84,90,94] However, several studies did not use techniques that allowed for adequate understanding of the models' decision-making (eg, data visualization).[68,70–76,80,82,83,85–89,92,93,95,97] In some studies, class unbalancing

Table 1. Bone Properties Assessment Studies Using Machine Learning Methods, Their General Characteristics and Results

Reference	Country (population)	Task	Modality	Data amount	Inputs (no.)	Model	CV train/ validation/test	Best result	Quality score (max 12)
Zhang et al.[23]	China	FL	QCT	100	5, 9	GRNN, **SVM**	58% train (10-fold CV), 22% test, 20% external test	MAPE = 4.0%	8
Nagarajan et al.[24]	Germany	FL	QCT	28	6	RBF-ANN	80% train, 20% test	RMSE = 1.02	5
Checefsky et al.[25]	Germany	FL	QCT	12	19	linreg, **SVM**	70% train, 30% test	RMSE = 0.82	5
Xiao et al.[26]	USA	Microarchitecture	QCT	1249	IMG	3-layer CNN	64% train, 16% valid, 20% test	r = 1.00	8
Mohanty et al.[27]	NR	Microarchitecture	Acoustics	964	8	ANN	64% train, 16% valid, 20% test	r^2 = 0.99	6
Pang et al.[28]	NR	VBH	MRI	235	IMG	CARN	5-fold CV train/test	MAE = 1.22 mm	10
Zhang et al.[29]a	China	BMD	Database	9053	6	SVM	2-fold CV train/test	MSE = 0.00	8
Fang et al.[30]	China	BMD	CT	1449	IMG slices	DenseNet	40% train (5-fold CV), 60% test	r^2 = 0.99	8
González et al.[31]	USA	BMD	CT	9925	IMG	3-layer CNN	80% train, 10% valid, 10% test	r = 0.94	8
Yasaka et al.[32]a	Japan	BMD	CT	2045	IMG slices	4-layer CNN	81% train, 9% test, 10% external test	r = 0.84	7
Krishnaraj et al.[33]a	NR	BMD	CT	1693	IMG	**linreg**, SVM, logreg	20% train, 100% test	r = 0.85	6
Nam et al.[34]a	South Korea	BMD	CT	198	3	linreg	80% train, 20% test	MSE = 4.46	5
Mohamed et al.[35]	Egypt	BMD	DXA	3000	256, 77365	ANN	60% train, 15% valid, 25% test	RMSE = 0.04	5

NR = not reported; FL = fracture load; VBH = vertebral bone height; BMD = bone mineral density; OP = osteoporosis; LS = lumbar spine; NS = non-standard; VF = vertebral fracture; HF = hip fracture; HuF = humerus fracture; WF = wrist fracture; OF = other fracture location; FR = fracture risk; CT = computed tomography; QCT = quantitative computed tomography; MRI = magnetic resonance imaging; DXA = dual-energy X-ray absorptiometry; DPR = dental panoramic radiography; IMG = image; RBF = radial basis function; ANN = artificial neural network; GRNN = general regression neural network; SVM = support-vector machine; CARN = cascade amplifier regression network; CNN = convolutional neural network; BVLC = Berkeley vision and learning center CNN; NIN = network in network; linreg = linear regression; logreg = logistic regression; RF = random forests; kNN = k-nearest neighbors; MBO = monarch butterfly optimization; SC-CNN = single-column CNN; MC-CNN = multi-column CNN; CART = classification and regression trees; GA = genetic algorithm; LSTM = long short-term memory; HAC = hierarchical agglomerative clustering; LDA = latent Dirichlet allocation; PDM = Poisson Dirichlet model; CV = cross-validation; train = training set; valid = validation set; test = testing set; ROC = receiver operating characteristic; AUC = area under the ROC curve; MSE = mean squared error; RMSE = root mean squared error; MAPE = mean absolute percentage error; r^2 = coefficient of determination; MAE = mean absolute error; r = Pearson correlation coefficient; ACC = accuracy; SEN = sensitivity.

[a]Multipurpose study.

Numerical inputs denote tabular features and IMG represent unstructured images.

Best-performing models are highlighted in bold.

Results indicated by percentages are reported as rounded to one decimal place and others using two decimal places.

Each table is grouped by task and sorted by quality score.

Table 2. Osteoporosis Classification Studies Using Machine Learning Methods, Their General Characteristics and Results

Reference	Country (population)	Task	Modality	Data amount (% OP)	Inputs (no.)	Model	CV train/ validation/test	Best result	Quality score (max 12)
Lee et al.[36]	South Korea	NR OP	DPR	680 (44)	IMG	3-layer CNN, **VGG-16**	80% train (5-fold CV), 20% test	AUC = 0.86	9
Zhang et al.[37]	China	LS OP	X-ray	1820 (39)	IMG	5-layer CNN	64% train, 8% valid, 28% test	AUC = 0.81	9
Yamamoto et al.[38]	Japan	Hip OP	X-ray	1131 (53)	IMG + 4	ResNet-18, resNet-34, GoogleNet, **EfficientNet b3**, EfficientNet b4	80% train, 10% valid, 10% test	AUC = 0.94	9
Wang et al.[39]	China	LS + Hip OP	Database	320 (33)	5, 9	RF	70% train, 30% test	AUC = 0.83	9
Shim et al.[40]	South Korea	LS + Hip OP	Database	1792 (34)	19, 9	kNN, logreg, decision tree, RF, gradient boosting, SVM, **ANN**	76% train (5-fold CV), 24% test	AUC = 0.74	8
Erjiang et al.[41]	Ireland	LS + Hip OP	Database	13577 (18)	30	CatBoost, **XGBoost**, ANN, linear discriminant, RF, logreg, SVM	80% train (5-fold CV), 20% test	AUC = 0.83	8
Vogl et al.[42]	Switzerland	LS + Hip OP	Acoustics	40 (63)	7	SVM	10x6 nested CV train/ test	AUC = 0.83	8
Singh et al.[43]	France	NS OP	X-ray	174 (50)	6	ANN, **SVM**, naive Bayes	46% train, 54% test	AUC = 1.00	8
Tecle et al.[44]	USA	NS OP	X-ray	2358 (45)	IMG	AlexNet	79% train, 11% test	ACC = 94.0%	7
Areeckal et al.[45]	India	NR OP	X-ray	117 (51)	3, 13	ANN	62% train, 12% valid, 26% test	ACC = 88.5%	7
Rastegar et al.[46]	NR	LS + Hip OP	DXA	147 (18)	54	kNN, **RF**, **ensmble**	10-fold CV train/test	AUC = 0.78	7
Kavitha et al.[47]	South Korea	LS + Hip OP	DPR	141 (15)	18	naive Bayes, kNN, **SVM**	5-fold CV and leave-one-out train/test	AUC = 0.95	7
Kavitha et al.[48]	South Korea	LS + Hip OP	DPR	141 (15)	10	fuzzy classifier	5-fold CV train/test	AUC = 0.99	7
Chu et al.[49]	NR	NS-OP	DPR	108 (48)	IMG	Siamese-CNN	LOO CV train/test	ACC = 89.8%	7
Zhang et al.[29]a	China	LS + Hip OP	Database	9053 (27)	6	SVM	2-fold CV train/test	AUC = 1.00	7
Meng et al.[50]	China	Hip OP	Database	2061 (26)	2	ANN	51% train (10-fold CV), 25% test, 24% external test	AUC = 0.83	7
Yasaka et al.[32]a	Japan	LS OP	CT	2045 (NR)	IMG slices	4-layer CNN	81% train, 9% test, 10% external test	AUC = 0.97	7
Krishnaraj et al.[33]a	NR	LS OP	CT	1693 (NR)	IMG	**linreg**, SVM, logreg	20% train, 100% test	ACC = 82.0%	6
Kilic and Hosgormez[51]	Turkey	LS OP	DXA	350 (26)	24	Bagging, boosting, **Random subspace method**	10-fold CV train/test	ACC = 98.9%	6
Scanlan et al.[52]	UK	NR OP	Acoustics	110 (31)	378	ANN	50% train, 25% valid, 25% test	ACC ≈ 70.0%	6
Hwang et al.[53]	South Korea	LS + Hip OP	DPR	454 (50)	19	decision tree, **SVM**	10-fold CV train/test	ACC = 96.9%	6

(Continues)

Table 2. Continued

Reference	Country (population)	Task	Modality	Data amount (% OP)	Inputs (no.)	Model	CV train/ validation/test	Best result	Quality score (max 12)
Lee et al.[54]	South Korea	NS OP	DPR	1268 (50)	IMG	**SC-CNN, MC-CNN**	84% train, 16% test	AUC = 1.00	6
Oulhaj et al.[55]	France	NS OP	X-ray	174 (50)	48	SVM	10-fold CV train/test	AUC = 0.93	6
Zheng et al.[56]	France	NS OP	X-ray	174 (50)	723	12 ML and DL models (**Bag of Keypoints**)	LOO CV train/test	AUC = 1.00	6
Nasser et al.[57]	France	NR OP	X-ray	87 (45)	IMG	Auto-encoder ANN + SVM	10-fold CV train/test	ACC = 95.5%	6
Ashok Kumar et al.[58]	India	Hip OP	X-ray	56 (48)	14	ANN	75% train, 25% test	ACC = 87.5%	5
Lee et al.[59]	South Korea	LS + Hip OP	X-ray	334 (51)	IMG	AlexNet, **VGG-16**, Inception-v3, ResNet-50, SVM, kNN, **RF**	70% train (5-fold CV), 30% test	AUC = 0.74	5
Devikanniga and Raj[60]	South Korea	LS + Hip OP	DPR	141 (15)	10	MBO-ANN	10-fold CV train/test	AUC = 1.00	5
Iliou et al.[61]	Greece	NR OP	Database	589 (33)	33	**ANN**, Naïve Bayes, logreg, SVM, kNN	10-fold CV	AUC = 0.95	5
Nam et al.[34]a	South Korea	LS OP	CT	198 (NR)	4	logistic classifier	80% train, 20% test	AUC = 0.90	5
Abu Marar et al.[62]	NR	NR OP	CBCT	120 (50)	7	ANN	60% train, 40% test	ACC = 97.9%	5
Liu et al.[63]	China	NR OP	X-ray	89 (35)	IMG, 20	**U-net CNN**, SVM	NR	AUC = 0.89	4
Bhattacharya et al.[64]	NR	NR-OP	X-ray	232 (50)	7	kNN, **SVM**	NR	ACC = 95.6%	2
Ragini et al.[65]	India	NS OP	Database	162 (22)	7, 11	ANN	70% train, 100% test	ACC = 87.5%	2

NR = not reported; FL = fracture load; VBH = vertebral bone height; BMD = bone mineral density; OP = osteoporosis; LS = lumbar spine; NS = non-standard; VF = vertebral fracture; HF = hip fracture; HuF = humerus fracture; WF = wrist fracture; OF = other fracture location; FR = fracture risk; CT = computed tomography; QCT = quantitative computed tomography; MRI = magnetic resonance imaging; DXA = dual-energy X-ray absorptiometry; DPR = dental panoramic radiography; IMG = image; RBF = radial basis function; ANN = artificial neural network; GRNN = general regression neural network; SVM = support-vector machine; CARN = cascade amplifier regression network; CNN = convolutional neural network; BVLC = Berkeley vision and learning center CNN; NIN = network in network; linreg = linear regression; logreg = logistic regression; RF = random forests; kNN = k-nearest neighbors; MBO = monarch butterfly optimization; SC-CNN = single-column CNN; MC-CNN = multi-column CNN; CART = classification and regression trees; GA = genetic algorithm; LSTM = long short-term memory; HAC = hierarchical agglomerative clustering; LDA = latent Dirichlet allocation; PDM = Poisson Dirichlet model; CV = cross-validation; train = training set; valid = validation set; test = testing set; ROC = receiver operating characteristic; AUC = area under the ROC curve; MSE = mean squared error; RMSE = root mean squared error; MAPE = mean absolute percentage error; r^2 = coefficient of determination; MAE = mean absolute error; r = Pearson correlation coefficient; ACC = accuracy; SEN = sensitivity.

[a]Multipurpose study.

Numerical inputs denote tabular features and IMG represent unstructured images.

Best-performing models are highlighted in bold.

Results indicated by percentages are reported as rounded to one decimal place and others using two decimal places.

Each table is grouped by task and sorted by quality score.

Table 3. Fracture Detection Studies Using Machine Learning Methods, Their General Characteristics and Results

Reference	Country (population)	Task	Modality	Data amount (% fractures)	Inputs (no.)	Model	CV train/ validation/test	Best result	Quality score (max 12)
Tomita et al.[66]a	USA	VF	CT	1432 (50)	IMG slices	ResNet-LSTM	80% train, 10% valid, 10% test	ACC = 89.2%	11
Derkatch et al.[67]	Canada	VF	DXA	12,742 (17)	IMG	Inception-v2, DenseNet, **ensemble**	60% train, 10% valid, 30% test	AUC = 0.94	11
Mehta and Sebro[68]	USA	VF	DXA	307 (35)	53	**SVM** (**linear**, polynomial, RBF, sigmoid)	80% train (10-fold CV), 20% test	AUC = 0.90	9
Valentinitsch et al.[69]	Germany	VF	CT	154 (34)	32,768	RF	4-fold CV train/test	AUC = 0.88	8
Burns et al.[70]	USA	VF	CT	150 (50)	21	SVM	10-fold CV train/test	SEN = 95.7%	7
Murata et al.[71]a	Japan	VF	X-ray	300 (50)	IMG	VisualRecognitionV3 CNN	5-fold CV train/test	AUC = 0.91	7
Raghavendra et al.[72]	NR	VF	CT	1120 (63)	IMG	3-layer CNN	70% train, 30% test	ACC = 99.1%	6
Frighetto-Pereira et al.[73]	Brazil	VF	MRI	191 (47)	27	kNN, **naive Bayes**, RBF-ANN	10-fold CV train/test	AUC = 0.97	5
Chen et al.[74]	Taiwan	VF, HF	Database	11,645 (10)	10	GA-SVM	50% train (10-fold CV), 50% test	AUC = 0.76	8
Minonzio et al.[75]	France	VF, HF, HuF, WF, OF	Acoustics	250 (45)	32	SVM	80% train (5-fold CV), 20% test	ACC = 92.3%	8
Ferizi et al.[76]	USA	VF, HF, HuF, WF, OF	MRI	460 (35)	40	**logreg**, SVM, decision tree, kNN, **ensemble**	23-fold CV train/test	AUC ≈ 0.63	6
Badgeley et al.[77]	USA, Australia	HF	X-ray	23,557 (3)	IMG, 30	**Inception-v3**, naive Bayes, **logreg**	75% train (10-fold CV), 25% test	AUC = 0.91	11
Cheng et al.[78]a	Taiwan	HF	X-ray	3705 (55)	IMG	DenseNet-121	78% train, 19% valid, 3% test	AUC = 0.98	11
Yu et al.[79]a	USA	HF	X-ray	617 (50)	IMG	Inception-v3	60% train, 20% valid, 20% test	AUC = 0.99	10
Yamada et al.[80]a	Japan	HF	X-ray	2903 (68)	IMG	Xception	90% train, 10% test	ACC = 98.0%	10
Jimenez Sanchez et al.[81]a	Germany	HF	X-ray	1347 (58)	IMG	ResNet-50	70% train, 10% valid, 20% test	AUC = 0.98	9
Adams et al.[82]a	Australia	HF	X-ray	805 (50)	IMG	AlexNet, **GoogLeNet**	64% train, 16% valid, 20% test	AUC = 0.98	9
Mawatari et al.[83]a	Japan	HF	X-ray	600 (50)	IMG	GoogLeNet	92% train, 8% test	AUC = 0.91	9
Mutasa et al.[84]	USA	HF	X-ray	1063 (69)	IMG	21-layer CNN	72% train, 18% valid, 10% test	AUC = 0.92	8
Urakawa et al.[85]a	Japan	HF	X-ray	2246 (53)	IMG	VGG-16	80% train, 10% valid, 10% test	AUC = 0.98	8
Beyaz et al.[86]	Turkey	HF	X-ray	234 (64)	IMG	5-layer CNN	5-fold CV train/test	ACC = 79.3%	7

(Continues)

Table 3. Continued

Reference	Country (population)	Task	Modality	Data amount (% fractures)	Inputs (no.)	Model	CV train/ validation/test	Best result	Quality score (max 12)
Carballido-Gamio et al.[87]	China	HF	QCT	143 (65)	NR	logreg	10-fold CV train/test	AUC = 0.93	7
Villamor et al.[88]	Spain	HF	Database	137 (65)	19	**RBF-SVM**, logreg, ANN, RF	74% train (10-fold CV), 26% test	ACC = 78.4%	7
Nadal et al.[89]	Spain	HF	Database	144 (58)	11	GA	80% train, 20% test	AUC = 0.77	6
Kitamura[90]	USA	HF, OF	X-ray	7337 (53)	IMG	DenseNet-121	70% train, 30% test	AUC = 0.95	9
Chung et al.[91]a	Korea	HuF	X-ray	1891 (69)	IMG	ResNet-152	90% train, 10% test	AUC = 1.00	10
Demir et al.[92]	Turkey	HuF	X-ray	115 (90)	512	kNN, **SVM**, **linear discriminant**, bagging	LOO CV train/test	ACC = 99.1%	6
Olczak et al.[93]a	Sweden	WF, OF	X-ray	256,458 (56)	IMG	BVLC, VGG-S, **VGG-16**, VGG-19, NIN	70% train, 20% valid, 10% test	ACC = 83.0%	10
Lindsey et al.[94]a	USA	WF	X-ray	135,409 (NR)	IMG	U-net CNN	80% train, 10% valid, 10% test	AUC = 0.99	10
Kim and MacKinnon[95]	UK	WF	X-ray	1489 (50)	IMG	Inception-v3	75% train, 9% valid, 9% test +7% external test	AUC = 0.95	8
Gebre et al.[96]	Finland	OF	CT	214 (50)	8, 10	Bayes logreg, ElasticNet	10-fold CV train/test	AUC = 1.00	8
Korfiatis et al.[97]	Italy	OF	micro-CT	63713 (3)	29	**ANN**, SVM, bagging and boosting	10-fold CV train/test	AUC = 0.92	6

NR = not reported; FL = fracture load; VBH = vertebral bone height; BMD = bone mineral density; OP = osteoporosis; LS = lumbar spine; NS = non-standard; VF = vertebral fracture; HF = hip fracture; HuF = humerus fracture; WF = wrist fracture; OF = other fracture location; FR = fracture risk; CT = computed tomography; QCT = quantitative computed tomography; MRI = magnetic resonance imaging; DXA = dual-energy X-ray absorptiometry; DPR = dental panoramic radiography; IMG = image; RBF = radial basis function; ANN = artificial neural network; GRNN = general regression neural network; SVM = support-vector machine; CARN = cascade amplifier regression network; CNN = convolutional neural network; BVLC = Berkeley vision and learning center CNN; NIN = network in network; linreg = linear regression; logreg = logistic regression; RF = random forests; kNN = k-nearest neighbors; MBO = monarch butterfly optimization; SC-CNN = single-column CNN; MC-CNN = multi-column CNN; CART = classification and regression trees; GA = genetic algorithm; LSTM = long short-term memory; HAC = hierarchical agglomerative clustering; LDA = latent Dirichlet allocation; PDM = Poisson Dirichlet model; CV = cross-validation; train = training set; valid = validation set; test = testing set; ROC = receiver operating characteristic; AUC = area under the ROC curve; MSE = mean squared error; RMSE = root mean squared error; MAPE = mean absolute percentage error; r^2 = coefficient of determination; MAE = mean absolute error; r = Pearson correlation coefficient; ACC = accuracy; SEN = sensitivity.

[a]Studies that compared their fracture detection model to human expert level.

Numerical inputs denote tabular features and IMG represent unstructured images.

Best-performing models are highlighted in bold.

Results indicated by percentages are reported as rounded to one decimal place and others using two decimal places.

Each table is grouped by task and sorted by quality score.

Table 4. Risk Prediction Studies Using Machine Learning Methods, Their General Characteristics and Results

Reference	Country (population)	Task (prediction interval)	Modality	Data amount (% cases)	Inputs(no.)	Model	CV train/ validation/test	Best result	Quality score (max 12)
Kruse et al.[98]	Denmark	FR	Database	10,775 (–)	44	HAC	Not required	Not required	6[a]
Wang et al.[99]	USA	Comorbidity	Database	388 (100)	32	LDA, PDM, HAC, K-means, Birch	Not required	Not required	6[a]
Shioji et al.[100]	Japan	BMD (10-year)	Database	135 (NR)	11	linreg, **ANN**	5-fold CV train/test	r = 0.93	6
Ye et al.[101]	USA	Fall (1-year)	Database	265,225 (2)	10,198	RF, Lasso, SVM, kNN, **XGBoost**	67% train, 33% test	AUC = 0.81	9
Cuaya Simbro et al.[102]	USA	Fall (6-month)	Database	253 (15)	4, 6, 10	Decision tree, **SVM**, ANN, dynamic Bayesian network	10-fold CV train/ test	ACC = 88.2%	7
Muehlematter et al.[103]	Switzerland	VF (∼8-month)	CT	120 (50)	29	ANN, RF, **SVM**	67% train (10-fold CV), 33%	AUC = 0.97	10
Kong et al.[104]b	South Korea	VF, HF (∼7.5-year)	Database	2227 (26)	35	**CatBoost**, SVM, logreg	3-fold CV train/test	AUC = 0.69	10
Wu et al.[105]	USA	VF, HF, WF, OF (∼4.5-year)	Database	5130 (9)	14	RF, **gradient boosting**, ANN, logreg	80% train (10-fold CV), 20% test	AUC = 0.71	9
Almog et al.[106]	USA	VF, HF, WF, OF (1-, 2-year)	Database	63,0445 (7)	>40000	Word2Vec, Doc2Vec, LSTM, XGBoost, **ensemble**	70% train (3-fold CV), 30% test	AUC = 0.82	9
Su et al.[107]b	USA	HF (10-year)	Database	5977 (3)	15	CART	10-fold CV train/ test	AUC = 0.73	10
Kruse et al.[108]	Denmark	HF (5-year)	Database	5439 (6)	1255	24 ML models (**XGBoost**)	75% train (5-fold CV), 25% test	AUC = 0.92	9
Engels et al.[109]	Germany	HF (4-year)	Database	288,086 (3)	26	**logreg**, RF, SVM, RUSBoost, XGBoost, ensemble	80% train (10-fold CV), 20% test	AUC = 0.70	9
Jiang et al.[110]	USA	HF (10-year)	Database	5002 (2)	20	SVM	58% train (10-fold CV), 42% test	AUC = 0.88	9
Ho-Le et al.[111]	Australia	HF (10-year)	Database	1167 (8)	11	ANN	60% train (5-fold CV), 40% test	AUC = 0.94	8

NR = not reported; FL = fracture load; VBH = vertebral bone height; BMD = bone mineral density; OP = osteoporosis; LS = lumbar spine; NS = non-standard; VF = vertebral fracture; HF = hip fracture; HuF = humerus fracture; WF = wrist fracture; OF = other fracture location; FR = fracture risk; CT = computed tomography; QCT = quantitative computed tomography; MRI = magnetic resonance imaging; DXA = dual-energy X-ray absorptiometry; DPR = dental panoramic radiography; IMG = image; RBF = radial basis function; ANN = artificial neural network; GRNN = general regression neural network; SVM = support-vector machine; CARN = cascade amplifier regression network; CNN = convolutional neural network; BVLC = Berkeley vision and learning center CNN; NIN = network in network; linreg = linear regression; logreg = logistic regression; RF = random forests; kNN = k-nearest neighbors; MBO = monarch butterfly optimization; SC-CNN = single-column CNN; MC-CNN = multi-column CNN; CART = classification and regression trees; GA = genetic algorithm; LSTM = long short-term memory; HAC = hierarchical agglomerative clustering; LDA = latent Dirichlet allocation; PDM = Poisson Dirichlet model; CV = cross-validation; train = training set; valid = validation set; test = testing set; ROC = receiver operating characteristic; AUC = area under the ROC curve; MSE = mean squared error; RMSE = root mean squared error; MAPE = mean absolute percentage error; r^2 = coefficient of determination; MAE = mean absolute error; r = Pearson correlation coefficient; ACC = accuracy; SEN = sensitivity.

[a]Unsupervised tasks do not apply for some points of the quality assessment checklist.

[b]Studies that compared their fracture risk model to FRAX. Both studies reported better performance than FRAX. Numerical inputs denote tabular features and IMG represent unstructured images. Best-performing models are highlighted in bold. Results indicated by percentages are reported as rounded to one decimal place and others using two decimal places. Each table is grouped by task and sorted by quality score.

was handled using sampling methods,[74,97] and overfitting was controlled with data augmentation.[80,83,84,95] Mutasa and colleagues demonstrated how data augmentation using digitally reconstructed and generated images was able to improve the test AUC from 0.80 (95% CI, not reported) to 0.92 (95% CI, not reported).[84] On the other hand, Adams and colleagues reported no significant improvement using data augmentation, whereas larger data sets had positive effects on the hip fracture detection accuracy.[82] Several studies did not report the confidence intervals or standard errors for the performance metrics that they reported.[66,68,69,72–74,79,81,83,84,86,88,89,92] Four studies obtained both high overall quality scores and near perfect AUCs (AUC ≥ 0.99).[78,79,91,94] Possible reasons for these unusual performances are the use of images' regions of interest[79,91] or the presence of potential data set bias,[78] which can significantly simplify the task. They clearly mention the limitations of their study, a quality that the authors should be acknowledged for. DL models were not used in any of the studies comprising less than 300 samples given that large sample sizes are a prerequisite for the DL model implementation.[69,70,73,75,86–89,92,96] Three studies shared their code for model development replication.[67,75,77]

Studies on risk prediction

Predicting the risk of bone loss, osteoporotic fractures, falls, or comorbidities in osteoporotic patients over time was investigated in 14 studies (Table 4).[98–111] Two of them used unsupervised learning to identify fracture and comorbidity risk groups, respectively.[98,99] Kruse and colleagues developed a fracture risk clustering model to categorize subgroups of patients at risk.[98] Wang and colleagues investigated osteoporotic patients' subgroups and their related comorbidity risk.[99] The 12 remaining studies used supervised learning for the prediction of risk of osteoporosis by bone density loss at 10 years,[100] incident falls at 6 months[102] and 1 year,[101] incident vertebral fracture at ≈ 8 months,[103] hip fracture prediction at 4, 5, or 10 years,[107–111] vertebral or hip fractures at ≈ 7.5 years,[104] major osteoporotic fractures (hip, spine, wrist, or humerus) at ≈ 4.5 years,[105] and all sort of fracture sites at 1 and 2 years.[106] Because unsupervised learning is not intended to predict a predetermined outcome, no performance metrics were reported. However, interesting features were identified, such as antiresorptive treatment response compliance, which could improve osteoporosis treatment.[98] Shioji and colleagues accurately predicted BMD loss over 10 years.[100] They further demonstrated osteoporosis classification with an AUC of 0.87 (95% CI, not reported). Falls were accurately predicted, and further investigations were carried out on the fracture risk categories and comorbidity factors.[101] In the fracture prediction studies included, an average AUC of 0.82 (range 0.69 to 0.97) was observed. Kong and colleagues and Su and colleagues outperformed the FRAX score for fracture prediction with AUC 0.69 (95% CI 0.69–0.69) versus 0.66 (95% CI, NR), statistically significant and AUC 0.73 (95% CI 0.69–0.76) versus 0.70 (95% CI 0.67–0.74), statistically insignificant, respectively.[104,107] Muehlmatter and colleagues compared their ML model's performance with human visual assessment of trabecular bone texture, where the latter had a slightly better than chance performance.[103] By taking advantage of ML models applied to sequential data, Almog and colleagues investigated the use of natural language processing techniques, such as recurrent neural networks, to predict short-term incident fractures in electronic health records (EHRs).[106] They observed a better performance of their model in the prediction of subsequent fractures compared with first fractures. Three studies clearly documented methods for correcting class imbalance caused by the very low incidence of positive cases.[105,106,109] Ye and colleagues and Kruse and colleagues calibrated their prediction models by aligning the predicted with the observed probabilities, to aid its understandability and applicability in clinical practice.[101,108] Good visualization of predictors was obtained using SHAP values.[104] Curiously, one study found unconventional highly ranked fracture predictors, such as general practitioner or dentist expenses, whereas DXA-related predictors were ranked number nine.[108]

Discussion

In this qualitative review, we provide a summary of recent studies that investigated osteoporosis or fracture by using ML approaches. The majority of the included articles used opportunistic data—particularly imaging. A wide variety of features were explored, and novel ones were proposed in certain studies. The ML approaches followed were diverse and many of them had promising performances, indicating the potential of ML use in the overall osteoporosis management. The following sections delve into the overall quality of the studies reviewed, the performance and reliability of the ML models that were used, and the actual clinical implications of ML models.

Overall quality of the studies

The ultimate goal of developing ML models or solutions is to propose improved clinical approaches. The clinical implication of a study heavily depends on its quality, which determines the extent at which its findings can be trusted. In this review, only articles published in peer-review journals were included, thus assuming a minimum acceptable quality. Nevertheless, AI researchers may also publish their work on preprint servers. This potentially excludes these contributions from the scientific mainstream, deprives them from the quality upgrade that usually accompanies peer-reviews, and eventually ceases their contribution to the assembly a solid body of evidence in the field.

In this review, a modified MI-CLAIM checklist was used to assess the overall quality of each study.[22] We simplified the original MI-CLAIM checklist by removing the points concerning direct clinical relevance and incorporating it within the study design part of the checklist. Based on this checklist, the most critical methodological and reporting limitations observed were presence of overfitting risk,[29,33,35,37,40,48,51,54,58–60,63,65] lack of model selection configuration reporting,[23–25,29,33,35,39,41,42,45,46,50,51,53,56–61,64,65] reporting of almost perfect model performances,[29,43,48,51,54,56,60,72,92] or lack of confidence intervals around point estimates.[33,34,44,52–65,72,73,86,88,92,102] In general, the studies where these issues were observed were not adequately acknowledged.[29,51,60,72,92] Acknowledging potential biases or limitations is fundamental for the accurate interpretation and especially for claims related to clinical implication. Nevertheless, several studies showed a greater awareness of the potential existence of bias and made efforts to address this either directly by controlling for the risks of overfitting,[23,28,30,32,34,37–40,43–46,49,53,54,56,57,59,64,77–85,90,91,93–95] indirectly by visualizing decision-making,[36,38,48,66,67,78,79,84,90,94,104] or by using gold standard strategies for comparison.[66,71,78–83,85,91,93,94,104,107]

Clear and transparent reporting of methodological approaches and results underpins confidence in a study. Underreporting of key information in medical studies has been noted

previously[112,113] and is observed in the present review. In some studies, it was unclear whether completely independent training and test sets were used.[47,63–65] In the comparison and selection of hyperparameter configurations, it is recommended to create a validation set from the training set to avoid overfitting. However, several studies evaluated their performance on the entire data set using k-fold cross-validation. One study employed confusing terminology by using cross-validation for the intermediate validation set,[35] a previously reported pitfall.[2] ML approaches in osteoporosis research is relatively new and thus lacks a well-defined pipeline for clinical embedding. Therefore, it is particularly important that the methodology followed in each study be reported in sufficient detail to help build up a robust standardized pipeline. Lack of reporting transparency further impedes effective model comparison and reproducibility. Two recent systematic reviews of ML reported high rates of bias among medical studies due to this.[114,115] Suggested solutions include statements and checklists to improve and standardize the quality of reporting in ML studies.[22,116] Peer-review journals have obligatory checklists to complete or follow before submission, such as STROBE for human observational studies,[117] CONSORT for randomized controlled trials,[118] or PRISMA for systematic reviews and meta-analyses.[119] Similarly, standardized obligatory checklists should be incorporated into the AI peer-review process.

External validation and replication of the proposed ML models are prerequisites for their clinical implementation. However, of 89 included studies, only four validated their model externally.[23,32,50,95] It is encouraging to note that some authors reported their methodology in great detail and/or freely shared their data or code, allowing for future replication.[28,37,42,56,67,75,77] Because the use of ML approaches to address uncertainties and issues in osteoporosis is relatively recent, we are still on time to avoid a replication crisis in the field. Besides transparency and high-quality reporting, data and code sharing should be encouraged. In an era where failure to reproduce scientific results is all too common, publishers, researchers, and readers should be critical of reports that do not provide a sufficiently clear description of the model's details for independent replication.

ML models' performance

The majority of the reviewed studies used ML models, and several used DL models or multiple types of models. The identification of the most promising model requires good knowledge of its properties and assumptions and a transparent justification for its choice. Interestingly, in this review, only Lee and colleagues clearly described the rationale for their chosen ML model.[36]

Best-performing ML models were very diverse, with the most successful being SVM, ANN, and logistic regression. However, these are also among the most available ML models and their superior performance might be linked to their popularity. On the other hand, DL models were also widely used, and preconfigured models, such as ResNets, obtained some of the best results. Most DL models used X-ray images, likely reflecting the fact that this is the most readily available modality. Nevertheless, the highest reported results in regression and classification tasks achieved almost perfect performance, which cannot be explained solely by their popularity.[26,27,29,30,35,43,48,51,54,56,60,72,78,79,91,92,94] Indeed, certain limitations, such as potential bias in the input data[78] or the use of a region of interest instead of full images,[79,91] help to explain these impressive performances. Although our review presents simple comparisons between studies, any model performance metrics reported and direct comparisons should be interpreted with caution.

ML model performance is assessed differently depending on the task. For instance, estimating bone properties is a regression task, where performance is assessed by error or correlation measures. Most common metrics in regression tasks are mean absolute or squared error (MAE/MSE, respectively), where MAE gives equal weights in error magnitudes and MSE highly penalizes outliers. Several articles included in this review used the Pearson correlation coefficient to depict correlation between the predicted and actual measures. In general, the studies including classification tasks, whether in fracture detection or prediction tasks, determined model performances using accuracy or AUC. However, as highlighted by Tecle and colleagues,[44] accuracy is sensitive to prevalence and tends to favor the majority class.[120] Consequently, this metric often overestimates the real performance of the model. Therefore, one study mainly reported their screening sensitivity.[70] The AUC may also be subject to this in clinical scenarios where population prevalence is an important factor, as demonstrated by Badgeley and colleagues.[77] In such settings, reporting multiple metrics, such as sensitivity, specificity, and area under precision-recall curve (AUPRC) is recommended.[2,22,121]

Efforts to quantify fracture risk prediction typically evaluated models' performance using AUC. The currently widely used model to assess 10-year fracture risk prediction, FRAX, has an AUC ranging between 0.74 and 0.79.[122] Therefore, fracture risk prediction using ML models have room for improvement.[123] However, single performance metrics such as AUC are insufficient to recommend models' implementation into clinical practice.[124,125] A calibration step is necessary to align the predicted probability of the event with the observed occurrence of the event. In addition, osteoporotic fracture risk prediction depends on multiple factors, the majority of which are time-dependent.[8] Thus, ML models using temporal sequences of data, such as recurrent neural networks, could better predict fracture risk. However, only two studies have applied methods based on this type of information,[66,106] a gap in the current approaches yet an opportunity for future work in osteoporosis management.

ML model reliability

Sample size, data quality, and the rigor in results' reporting are major determinants of models' reliability. In the current review, sample sizes varied from less than 50 to hundreds of thousands. Depending on the complexity of the ML model architecture, a robust model training might need from a hundred to millions of samples. However, in practice, it is often difficult to estimate the sample size needed to appropriately handle a given task. State-of-the-art DL studies commonly use data sets with thousands of images to train complex DL models containing millions of parameters. In practice, access to labeled databases with a well-defined outcome definition is challenging. Expectedly, the studies reviewed here present issues related to data quantity, and one highlighted such issue.[52] Data augmentation becomes particularly essential in avoiding overfitting and inconsistent predictions. Xiao and colleagues showed the importance of the number and resolution of images to identify microstructural bone characteristics.[26] In addition, model pre-training can also

be of great importance in small data sets. In Lee and colleagues, the model's performance for osteoporosis screening was significantly improved after experimenting with several transfer learning methodologies, among which fine-tuning proved to the be most successful.[36] Moreover, the use of unlabeled data—more available and accessible—in conjunction with a limited amount of labeled data, known as semi-supervised learning,[126] further improves a model's performance. To the best of our knowledge, all the studies included in this review used complete data, either by removing incomplete data or using completion strategies.

Our review shows that Big Data sources, such as opportunistic imaging or EHRs, hold promise for tackling the issue of limited data availability. In general, heterogeneity in sources of information present noise and uncertainty. This gives rise to a fundamental question of how to establish ground truth. Ground truth, as reference labels data used to train the model, should ideally be assessed using gold standard techniques. However, the use of gold standards is not always possible because of limited data availability. For example, the gold standard for osteoporosis diagnosis is the DXA-BMD *T*-score thresholds as based on the WHO guidelines. In this review, several studies used proxies for osteoporosis diagnosis, such as cortical thickness measured from dental panoramic radiography or from hand X-ray.[43,44,49,54–56,65] The use of such unconventional tools limits the clinical application of the ML models derived from these efforts. Some studies did not even clearly state their definition of ground truth, a likely cause for some extremely optimistic and unrealistic results.[36,45,52,57,61–64] Lack of a reference standard makes it difficult to accurately evaluate the clinical implications of the proposed ML solution.

Clinical implications

The implementation of ML algorithms in clinical practice requires additional precautions. AI/ML models might suggest novel solutions to certain issues in osteoporosis or might present improvements to existing solutions. Models providing novelty in clinical practice, such as opportunistic screening for fracture detection or osteoporosis diagnosis, necessitates a well-detailed pathway before their clinical application. This pathway is as yet still not clearly defined. However, characteristics such as high accuracy, positive clinical consequences, and cost-effectiveness are essential for clinical integration.[127] Models improving existing tools, such as fracture risk prediction as assessed by FRAX, have to at least demonstrate equivalence with the existing tool.[104,107] For tasks such as fracture detection, patient and societal acceptance may be a barrier to implementation even if the algorithm is scientifically validated, and raises subjective issues such as "trust," duty of care, and medicolegal liability.

Clinical decision-making is the core competence of physicians, a competence slowly acquired during years of training and clinical experience. AI represents the training of machines to develop expertise from the analysis of data. The dilemma that will inevitably arise is "Should we trust the clinician or the AI?" The trustworthiness of ML models largely depends on transparency from their developers and the embedding ability to explain the model's findings and/or recommendations to clinicians. For instance, only two of the seven articles related to osteoporosis classification using DL made an effort to illustrate the decision-making of their models.[36,38] Several DL efforts in fracture detection visualized the findings of their models using activation maps (Grad-Cam[20]).[66,67,78,79,84,90,94] Other methods have also been applied such as t-SNE (t-distributed stochastic neighbor embedding[128]) to visualize high-dimensional data representations[77,81,99] or SHAP values (shapley additive explanations).[104] To assess the reliability of heat maps on images, Arun and colleagues explored their capability in localizing region of interest in medical imaging DL models.[129] None of the explanatory methods tested passed all validity checks, so a careful interpretation of these visualization tools is required. Simplified illustrations of the sophisticated ML solutions will help clinicians evaluate their actual potential for incorporation into routine clinical practice and dispel the "black box" perception of AI; hence, such illustrative approaches should be encouraged and ideally included in the standard checklist for ML studies.

Suresh and colleagues investigated trust on tasks difficult for humans.[130] The authors demonstrated an increase in recommendation agreement when additional information was provided on the model's decision. They also observed that people overrely on ML incorrect recommendations, although those with higher mathematical skills and ML knowledge are less concerned. From an ethical point of view and in the absence of precise measures of trust, the expertise and judgment of the clinician must remain at the heart of the decision-making process. The legal liability from a bad decision arising from ML remains an unresolved issue.[131] Nevertheless, valuable evidence can be identified by the AI to support the clinician in the highly complex decision-making process, thus giving rise to an "enhanced" clinician. Four of the reviewed studies demonstrated an improvement in hip and humerus fracture identification from X-rays when using DL.[80,83,85,91] In addition, a recent study further validated one of the reviewed studies[78] and demonstrated how AI recommendations were able to improve the identification of hip fractures in clinical routine.[132] Prediction and diagnosis models increasingly prove to be more accurate than experts, expanding on the limits of human performance. This raises the ethical question whether AI decisions may have undesired side effects. Beyond AI/ML performance and liability, other principles such as fairness, privacy, and safety are important.[133] Examples are the problems of ensuring representative (unbiased) data, safeguarding data protection, and preventing non-beneficial effects, respectively. According to the US Food and Drug Administration (FDA), ML-based algorithms belong to the category of medical devices and must therefore follow the same rules. It is therefore essential to regulate these systems properly for the benefit of the patients.[134,135]

The cost-effectiveness of the use of AI/ML solutions in clinical practice is going to be required from payers. Ideally, the proposed model should be effective when its use in practice comes without significant additional direct or indirect costs.

The complexity of the AI concept is also mirrored in the multidimensionality of the problems that it raises and of the infrastructure that it requires to ultimately fulfill its aim to be helpful in clinical practice. The story of OsteoDetect, a DL solution to automate wrist fractures detection, reassures us that the pathway to a validated clinical use of AI approaches in osteoporosis overall management is not unknown.[136]

Limitations and related work

This review has limitations that must be acknowledged. The search was limited to PubMed and Web of Science, which provide access to multiple databases simultaneously. However, some literature may have been overlooked. We included studies that could directly improve osteoporosis as a primary outcome, thus omitting technical publications related to bone

segmentation, genetics, or biomarkers. We prioritized the reporting of results in the form of AUC and accuracy. In addition, despite our best efforts, the proposed quality assessment questioner may be subject to certain limitations. For example, we considered acceptable all validation strategies reported, although external validation allows a more robust demonstration of clinical utility versus simple internal train/test cross-validation. Nevertheless, a large diversity of ML applications was obtained, which highlights common themes along the research pathways. Our review complements previous investigations in this subject. Cruz and colleagues discussed AI in the identification of risk groups.[137] Kruse reviewed the recent advances using Big Data in diabetes mellitus and osteoporosis.[138] Others have presented applications of ML and DL to joint and musculoskeletal imaging.[139–142] Hung and colleagues reviewed the use of AI in maxillofacial radiology and its use in opportunistic diagnosis of osteoporosis.[143] Hügle and colleagues presented an overview of ML applied in rheumatology.[144] Wani and Arora presented a comprehensive review of computer-assisted diagnostic systems, including some based on AI, for osteoporosis.[145]

Conclusion

ML is an exciting approach that holds the promise of advancing research in the field of osteoporosis, but this enthusiasm needs to be tempered by the potential for hidden biases and the need for researchers to acquire expertise in this new discipline in order to avoid the many pitfalls. This review provides a state-of-the-art overview of current efforts using ML to address issues across the osteoporosis management spectrum. The use of images for opportunistic osteoporosis diagnosis or fracture detection emerged as a promising tool in osteoporosis screening and a major advance that ML can bring to the osteoporosis field. Efforts to develop ML-based models for identifying novel fracture risk predictors and improving fracture prediction are additional promising lines of research. Adhering to a predefined detailed pipeline for ML implementation and reporting are essential for accurately assessing results and their clinical implications. Nevertheless, the majority of the studies reviewed in the current article suffered from the lack of a standardized approach in conducting and/or reporting the ML methodology. There is a need for journals to develop and require authors to follow standard checklists as part of the peer-review process. Hopefully the insights presented in this qualitative review will help clinicians and researchers alike, particularly those working in the field of bone health, to better understand the current use of ML in osteoporosis, along with its promises and pitfalls.

Disclosures

All authors state that they have no conflicts of interest.

Acknowledgments

We thank the Swiss National Fund (project number 32473B_156978) for the financial support. Sincere thanks to Cecile Jacques, Medical Library of Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland, for her help with the literature search strategy.

Author Roles: JS: PhD student, ES: Post-doc researcher, TH: head of rhumatology, WL: head of department, DH: head of research and development group.

Author contributions: Julien Smets: Conceptualization; investigation; writing-original draft. Enisa Shevroja: Conceptualization; writing-original draft; writing-review & editing. Thomas Hügle: Writing-review & editing. William Leslie: Writing-review & editing. Didier Hans: Conceptualization; supervision; validation; writing-review & editing.

Peer review

The peer review history for this article is available at https://publons.com/publon/10.1002/jbmr.4292.

References

1. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347-1358.
2. Liu Y, Chen PC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. JAMA. 2019;322(18):1806-1816.
3. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319(13):1317-1318.
4. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-444.
5. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402-2410.
6. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115-118.
7. Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med. 2019;25(6):954-961.
8. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018;1:18.
9. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One. 2017;12(4).
10. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med. 2018;24(11):1716-1720.
11. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020;577(7792):706-710.
12. Stokes JM, Yang K, Swanson K, et al. A deep learning approach to antibiotic discovery. Cell. 2020;180(4):688-702.
13. Johnell O, Kanis JA. An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. Osteoporos Int. 2006;17(12):1726-1733.
14. Kanis JA, Borgstrom F, De Laet C, et al. Assessment of fracture risk. Osteoporos Int. 2005;16(6):581-589.
15. Lever J, Krzywinski M, Altman N. Model selection and overfitting. Nat Methods. 2016;13(9):703-704.
16. Erhan D, Bengio Y, Courville A, Manzagol P-A, Vincent P, Bengio S. Why does unsupervised pre-training help deep learning? J Mach Learn Res. 2010;11:625-660.
17. Shevroja E, Lamy O, Kohlmeier L, Koromani F, Rivadeneira F, Hans D. Use of trabecular bone score (TBS) as a complementary approach to dual-energy X-ray absorptiometry (DXA) for fracture risk assessment in clinical practice. J Clin Densitom. 2017;20(3):334-345.
18. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60-88.
19. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56-67.
20. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis. 2020;128(2):336-359.

21. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev. 2016;5(1):210.
22. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020;26(9):1320-1324.
23. Zhang M, Gong H, Zhang K, Zhang M. Prediction of lumbar vertebral strength of elderly men based on quantitative computed tomography images using machine learning. Osteoporos Int. 2019;30(11): 2271-2282.
24. Nagarajan MB, Checefsky WA, Abidin AZ, et al. Characterizing trabecular bone structure for assessing vertebral fracture risk on volumetric quantitative computed tomography. Proc SPIE. 2015:9417: 94171E.
25. Checefsky WA, Abidin AZ, Nagarajan MB, Bauer JS, Baum T, Wismuller A. Assessing vertebral fracture risk on volumetric quantitative computed tomography by geometric characterization of trabecular bone structure. In Tourassi GD, Armato SG, eds. Medical imaging 2016: computer-aided diagnosis. Proceedings of SPIE. 9785. Spie-Int Soc Optical Engineering: Bellingham; 2015.
26. Xiao P, Zhang T, Dong XN, Han Y, Huang Y, Wang X. Prediction of trabecular bone architectural features by deep learning models using simulated DXA images. Bone Rep. 2020;13:100295.
27. Mohanty K, Yousefian O, Karbalaeisadegh Y, Ulrich M, Grimal Q, Muller M. Artificial neural network to estimate micro-architectural properties of cortical bone using ultrasonic attenuation: a 2-D numerical study. Comput Biol Med. 2019;114:103457.
28. Pang S, Su Z, Leung S, et al. Direct automated quantitative measurement of spine by cascade amplifier regression network with manifold regularization. Med Image Anal. 2019;55:103-115.
29. Zhang T, Liu P, Zhang Y, et al. Combining information from multiple bone turnover markers as diagnostic indices for osteoporosis using support vector machines. Biomarkers. 2019;24(2):120-126.
30. Fang Y, Li W, Chen X, et al. Opportunistic osteoporosis screening in multi-detector CT images using deep convolutional neural networks. Eur Radiol. 2021;31(4):1831-1842.
31. González G, Washko GR, Estépar RSJ. Deep learning for biomarker regression: application to osteoporosis and emphysema on chest CT scans. Proc SPIE. 2018;10574:105741H.
32. Yasaka K, Akai H, Kunimatsu A, Kiryu S, Abe O. Prediction of bone mineral density from computed tomography: application of deep learning with a convolutional neural network. Eur Radiol. 2020;30 (6):3549-3557.
33. Krishnaraj A, Barrett S, Bregman-Amitai O, et al. Simulating dual-energy X-ray absorptiometry in CT using deep-learning segmentation cascade. J Am Coll Radiol. 2019;16(10):1473-1479.
34. Nam KH, Seo I, Kim DH, Lee JI, Choi BK, Han IH. Machine learning model to predict osteoporotic spine with Hounsfield units on lumbar computed tomography. J Korean Neurosurg Soc. 2019;62(4):442-449.
35. Mohamed EI, Meshref RA, Abdel-Mageed SM, Moustafa MH, Badawi MI, Darwish SH. A novel morphological analysis of DXA-DICOM images by artificial neural networks for estimating bone mineral density in health and disease. J Clin Densitom. 2019;22(3):382-390.
36. Lee KS, Jung SK, Ryu JJ, Shin SW, Choi J. Evaluation of transfer learning with deep convolutional neural networks for screening osteoporosis in dental panoramic radiographs. J Clin Med. 2020;9(2):392.
37. Zhang B, Yu K, Ning Z, et al. Deep learning of lumbar spine X-ray for osteopenia and osteoporosis screening: a multicenter retrospective cohort study. Bone. 2020;140:115561.
38. Yamamoto N, Sukegawa S, Kitamura A, et al. Deep learning for osteoporosis classification using hip radiographs and patient clinical covariates. Biomolecules. 2020;10(11):1534.
39. Wang J, Yan D, Zhao A, et al. Discovery of potential biomarkers for osteoporosis using LC-MS/MS metabolomic methods. Osteoporos Int. 2019;30(7):1491-1499.
40. Shim JG, Kim DW, Ryu KH, et al. Application of machine learning approaches for osteoporosis risk prediction in postmenopausal women. Arch Osteoporos. 2020;15(1):169.
41. Erjiang E, Wang T, Yang L, et al. Machine learning can improve clinical detection of low BMD: the DXA-HIP study. J Clin Densitom. 2020 Oct 20. https://www.sciencedirect.com/science/article/pii/S1094695020301207?via%3Dihub.
42. Vogl F, Friesenbichler B, Hüsken L, Kramers-de Quervain IA, Taylor WR. Can low-frequency guided waves at the tibia paired with machine learning differentiate between healthy and osteopenic/osteoporotic subjects? A pilot study. Ultrasonics. 2019;94:109-116.
43. Singh A, Dutta MK, Jennane R, Lespessailles E. Classification of the trabecular bone structure of osteoporotic patients using machine vision. Comput Biol Med. 2017;91:148-158.
44. Tecle N, Teitel J, Morris MR, Sani N, Mitten D, Hammert WC. Convolutional neural network for second metacarpal radiographic osteoporosis screening. J Hand Surg. 2020;45(3):175-181.
45. Areeckal AS, Jayasheelan N, Kamath J, Zawadynski S, Kocher M, David SS. Early diagnosis of osteoporosis using radiogrammetry and texture analysis from hand and wrist radiographs in Indian population. Osteoporos Int. 2018;29(3):665-673.
46. Rastegar S, Vaziri M, Qasempour Y, et al. Radiomics for classification of bone mineral loss: a machine learning study. Diagn Interv Imaging. 2020;101(9):599-610.
47. Kavitha MS, An SY, An CH, et al. Texture analysis of mandibular cortical bone on digital dental panoramic radiographs for the diagnosis of osteoporosis in Korean women. Oral Surg Oral Med Oral Pathol Oral Radiol. 2015;119(3):346-356.
48. Kavitha MS, Kumar PG, Park SY, et al. Automatic detection of osteoporosis based on hybrid genetic swarm fuzzy classifier approaches. Dentomaxillofac Radiol. 2016;45(7):13.
49. Chu P, Bo C, Liang X, et al. Using octuplet siamese network for osteoporosis analysis on dental panoramic radiographs. Annu Int Conf IEEE Eng Med Biol Soc. 2018;2018:2579-2582.
50. Meng J, Sun N, Chen Y, et al. Artificial neural network optimizes self-examination of osteoporosis risk in women. J Int Med Res. 2019;47 (7):3088-3098.
51. Kilic N, Hosgormez E. Automatic estimation of osteoporotic fracture cases by using ensemble learning approaches. J Med Syst. 2016;40(3):61.
52. Scanlan J, Li FF, Umnova O, Rakoczy G, Lövey N, Scanlan P. Detection of osteoporosis from percussion responses using an electronic stethoscope and machine learning. Bioengineering (Basel, Switzerland). 2018;5(4):107.
53. Hwang JJ, Lee JH, Han SS, et al. Strut analysis for osteoporosis detection model using dental panoramic radiography. Dentomaxillofac Radiol. 2017;46(7):20170006.
54. Lee JS, Adhikari S, Liu L, Jeong HG, Kim H, Yoon SJ. Osteoporosis detection in panoramic radiographs using a deep convolutional neural network-based computer-assisted diagnosis system: a preliminary study. Dentomaxillofac Radiol. 2019;48(1):20170344.
55. Oulhaj H, Rziza M, Amine A, et al. Anisotropic discrete dual-tree wavelet transform for improved classification of trabecular bone. IEEE Trans Med Imaging. 2017;36(10):2077-2086.
56. Zheng K, Harris CE, Jennane R, Makrogiannis S. Integrative block-wise sparse analysis for tissue characterization and classification. Artif Intell Med. 2020;107:101885.
57. Nasser Y, El Hassouni M, Brahim A, Toumi H, Lespessailles E, Jennane R. Diagnosis of osteoporosis disease from bone X-ray images with stacked sparse autoencoder and SVM classifier. In ElHassouni M, Karim M, BenHamida A, BenSlima A, Solaiman B, eds. New York: IEEE; 2017 pp 408-412.
58. Ashok Kumar D, Anburajan M, Snekhalatha U. Evaluation of low bone mass and prediction of fracture risk using metacarpal radiogrammetry method: a comparative study with DXA and X-ray phantom. Int J Rheum Dis. 2018;21(7):1350-1371.
59. Lee S, Choe EK, Kang HY, Yoon JW, Kim HS. The exploration of feature extraction and machine learning for predicting bone density from simple spine X-ray images in a Korean population. Skeletal Radiol. 2020;49(4):613-618.
60. Devikanniga D, Joshua Samuel Raj R. Classification of osteoporosis by artificial neural network based on monarch butterfly optimisation algorithm. Healthc Technol Lett. 2018;5(2):70-75.
61. Iliou T, Anagnostopoulos CN, Stephanakis IM, Anastassopoulos G. A novel data preprocessing method for boosting neural network performance: a case study in osteoporosis prediction. Inf Sci. 2017;380:92-100.

62. Abu Marar RF, Uliyan DM, Al-Sewadi HA. Mandible bone osteoporosis detection using cone-beam computed tomography. Eng Technol Appl Sci Res. 2020;10(4):6027-6033.
63. Liu J, Wang J, Ruan W, Lin C, Chen D. Diagnostic and gradation model of osteoporosis based on improved deep U-net network. J Med Syst. 2019;44(1):15.
64. Bhattacharya S, Nair D, Bhan A, Goyal A. Computer based automatic detection and classification of osteoporosis in bone radiographs. New York, NY: IEEE; 2019 pp 1047-1052.
65. Ragini B, Subramaniyan KSA, Sanchana K, Anburajan M. Evaluation of low bone mineral mass using a combination of peripheral bone mineral density and total body composition variables by neural network, eds. Procedia Computer Science, vol. 57; 2015; pp. 1115-1123.
66. Tomita N, Cheung YY, Hassanpour S. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. Comput Biol Med. 2018;98:8-15.
67. Derkatch S, Kirby C, Kimelman D, Jozani MJ, Davidson JM, Leslie WD. Identification of vertebral fractures by convolutional neural networks to predict nonvertebral and hip fractures: a registry-based cohort study of dual X-ray absorptiometry. Radiology. 2019;293(2):405-411.
68. Mehta SD, Sebro R. Computer-aided detection of incidental lumbar spine fractures from routine dual-energy X-ray absorptiometry (DEXA) studies using a support vector machine (SVM) classifier. J Digit Imaging. 2020;33(1):204-210.
69. Valentinitsch A, Trebeschi S, Kaesmacher J, et al. Opportunistic osteoporosis screening in multi-detector CT images via local classification of textures. Osteoporos Int. 2019;30(6):1275-1285.
70. Burns JE, Yao J, Summers RM. Vertebral body compression fractures and bone density: automated detection and classification on CT images. Radiology. 2017;284(3):788-797.
71. Murata K, Endo K, Aihara T, et al. Artificial intelligence for the detection of vertebral fractures on plain spinal radiography. Sci Rep. 2020; 10(1):20031.
72. Raghavendra U, Bhat NS, Gudigar A, Acharya UR. Automated system for the detection of thoracolumbar fractures using a CNN architecture. Futur Gener Comput Syst. 2018;85:184-189.
73. Frighetto-Pereira L, Rangayyan RM, Metzner GA, de Azevedo-Marques PM, Nogueira-Barbosa MH. Shape, texture and statistical features for classification of benign and malignant vertebral compression fractures in magnetic resonance images. Comput Biol Med. 2016;73:147-156.
74. Chen YF, Lin CS, Wang KA, et al. Design of a clinical decision support system for fracture prediction using imbalanced dataset. J Healthc Eng. 2018;2018:9621640.
75. Minonzio JG, Cataldo B, Olivares R, et al. Automatic classifying of patients with non-traumatic fractures based on ultrasonic guided wave spectrum image using a dynamic support vector machine. IEEE Access. 2020;8:194752-194764.
76. Ferizi U, Besser H, Hysi P, et al. Artificial intelligence applied to osteoporosis: a performance comparison of machine learning algorithms in predicting fragility fractures from MRI data. J Magn Reson Imaging. 2019;49(4):1029-1038.
77. Badgeley MA, Zech JR, Oakden-Rayner L, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ Digit Med. 2019;2:31.
78. 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.
79. 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.e1–9.
80. 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;12:1-6.
81. Jimenez-Sanchez 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.
82. 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.
83. 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.
84. 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.
85. 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. Skelet Radiol. 2019;48(2):239-244.
86. 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.
87. Carballido-Gamio J, Yu A, Wang L, et al. Hip fracture discrimination based on statistical multi-parametric modeling (SMPM). Ann Biomed Eng. 2019;47(11):2199-2212.
88. Villamor E, Monserrat C, Del Río L, Romero-Martín JA, Rupérez MJ. Prediction of osteoporotic hip fracture in postmenopausal women through patient-specific FE analyses and machine learning. Comput Methods Programs Biomed. 2020;193:105484.
89. Nadal E, Munoz D, Vivo N, Lucas I, Rodenas JJ. Evaluation of hip fracture risk using a hyper-parametric model based on the locally linear embedding technique. Compte Rendus Mecanique. 2019;347(11):856-862.
90. Kitamura G. Deep learning evaluation of pelvic radiographs for position, hardware presence, and fracture detection. Eur J Radiol. 2020; 130:109139.
91. Chung SW, Han SS, Lee JW, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop. 2018;89(4):468-473.
92. Demir S, Key S, Tuncer T, Dogan S. An exemplar pyramid feature extraction based humerus fracture classification method. Med Hypotheses. 2020;140:109663.
93. Olczak J, Fahlberg N, Maki A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthop. 2017;88(6):581-586.
94. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A. 2018;115 (45):11591-11596.
95. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol. 2018;73(5):439-445.
96. Gebre RK, Hirvasniemi J, Lantto I, Saarakkala S, Leppilahti J, Jämsä T. Discrimination of low-energy acetabular fractures from controls using computed tomography-based bone characteristics. Ann Biomed Eng. 2020;49:367-381.
97. Korfiatis VC, Tassani S, Matsopoulos GK, Korfiatis VC, Tassani S, Matsopoulos GK. A new ensemble classification system for fracture zone prediction using imbalanced micro-CT bone morphometrical data. IEEE J Biomed Health. 2018;22(4):1189-1196.
98. Kruse C, Eiken P, Vestergaard P. Clinical fracture risk evaluated by hierarchical agglomerative clustering. Osteoporos Int. 2017;28(3):819-832.
99. Wang Y, Zhao Y, Therneau TM, et al. Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records. J Biomed Inform. 2020;102:103364.
100. Shioji M, Yamamoto T, Ibata T, Tsuda T, Adachi K, Yoshimura N. Artificial neural networks to predict future bone mineral density and bone loss rate in Japanese postmenopausal women. BMC Res Notes. 2017;10(1):590.
101. Ye C, Li J, Hao S, et al. Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm. Int J Med Inform. 2020;137:104105.
102. Cuaya-Simbro G, Perez-Sanpablo AI, Munoz-Melendez A, Uriostegui IQ, Morales-Manzanares EF, Nunez-Carrera L. Comparison of machine learning models to predict risk of falling in osteoporosis elderly. Found Comput Decis Sci. 2020;45(2):65-77.
103. Muehlematter UJ, Mannil M, Becker AS, et al. Vertebral body insufficiency fractures: detection of vertebrae at risk on standard CT

images using texture analysis and machine learning. Eur Radiol. 2019;29(5):2207-2217.

104. Kong SH, Ahn D, Kim BR, et al. A novel fracture prediction model using machine learning in a community-based cohort. JBMR Plus. 2020;4(3).
105. Wu Q, Nasoz F, Jung J, Bhattarai B, Han MV. Machine learning approaches for fracture risk assessment: a comparative analysis of genomic and phenotypic data in 5130 older men. Calcif Tissue Int. 2020;107(4):353-361.
106. Almog YA, Rai A, Zhang P, et al. Deep learning with electronic health records for short-term fracture risk identification: crystal bone algorithm development and validation. J Med Internet Res. 2020;22(10).
107. Su Y, Kwok TCY, Cummings SR, Yip BHK, Cawthon PM. Can classification and regression tree analysis help identify clinically meaningful risk groups for hip fracture prediction in older American men (the MrOS cohort study)? JBMR Plus. 2019;3(10).
108. Kruse C, Eiken P, Vestergaard P. Machine learning principles can improve hip fracture prediction. Calcif Tissue Int. 2017;100(4):348-360.
109. Engels A, Reber KC, Lindlbauer I, et al. Osteoporotic hip fracture prediction from risk factors available in administrative claims data—a machine learning approach. PLoS One. 2020;15(5).
110. Jiang P, Missoum S, Chen Z. Fusion of clinical and stochastic finite element data for hip fracture risk prediction. J Biomech. 2015;48 (15):4043-4052.
111. Ho-Le TP, Center JR, Eisman JA, Nguyen TV, Nguyen HT. Prediction of hip fracture in post-menopausal women using artificial neural network approach. Annu Int Conf IEEE Eng Med Biol Soc. 2017; 2017:4207-4210.
112. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594.
113. Ferizi U, Honig S, Chang G. Artificial intelligence, osteoporosis and fragility fractures. Curr Opin Rheumatol. 2019;31(4):368-375.
114. 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.
115. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368:m689.
116. Collins GS, Moons KGJTL. Reporting of artificial intelligence prediction models. Lancet. 2019;393(10181):1577-1579.
117. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008;61(4):344-349.
118. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c332.
119. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med. 2009;6(7).
120. Branco P, Torgo L, Ribeiro RP. A survey of predictive modeling on imbalanced domains. ACM Comput Surv. 2016;49(2):1-50.
121. Ozenne B, Subtil F, Maucort-Boulch D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J Clin Epidemiol. 2015;68(8):855-859.
122. Marques A, Ferreira RJO, Santos E, Loza E, Carmona L, da JAP S. The accuracy of osteoporotic fracture risk prediction tools: a systematic review and meta-analysis. Ann Rheum Dis. 2015;74(11):1958-1967.
123. El Miedany Y. FRAX: re-adjust or re-think. Arch Osteoporos. 2020;15 (1):150.
124. Pencina MJ, D'Agostino RB Sr. Evaluating discrimination of risk prediction models: the C statistic. JAMA. 2015;314(10):1063-1064.
125. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. JAMA. 2017;318(14):1377-1384.
126. Sohn K, Berthelot D, Li C-L, et al. Fixmatch: simplifying semi-supervised learning with consistency and confidence. arXiv preprint arXiv. 2020:200107685.
127. Kolanu N, Silverstone EJ, Ho BH, et al. Clinical utility of computer-aided diagnosis of vertebral fractures from computed tomography images. J Bone Miner Res. 2020;35(12):2307-2312.
128. van der Maaten L, Hinton G. Viualizing data using t-SNE. J Mach Learn Res. 2008;9:2579-2605.
129. Arun N, Gaw N, Singh P, et al. Assessing the (un) trustworthiness of saliency maps for localizing abnormalities in medical imaging. arXiv preprint arXiv. 2020:200802766.
130. Suresh H, Lao N, Liccardi I. Misplaced trust: measuring the interference of machine learning in human decision-making. 12th ACM conference on web science. Southampton, UK: Association for Computing Machinery; 2020. p. 315–24.
131. Price WN 2nd, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. JAMA. 2019;322(18):1765-1766. https://doi.org/10.1001/jama.2019.15064.
132. Cheng CT, Chen CC, Cheng FJ, et al. A human-algorithm integration system for hip fracture detection on plain radiography: system development and validation study. JMIR Med Inform. 2020;8(11): e19416.
133. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell. 2019;1(9):389-399.
134. Hwang TJ, Kesselheim AS, Vokinger KN. Lifecycle regulation of artificial intelligence- and machine learning-based software devices in medicine. JAMA. 2019;322(23):2285-2286. https://doi.org/10.1001/jama.2019.16842.
135. Eaneff S, Obermeyer Z, Butte AJ. The case for algorithmic stewardship for artificial intelligence and machine learning technologies. JAMA. 2020;324(14):1397-1398. https://doi.org/10.1001/jama.2020.9371.
136. 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.
137. Cruz AS, Lins HC, Medeiros RVA, Filho JMF, da Silva SG. Artificial intelligence on the identification of risk groups for osteoporosis, a general review. Biomed Eng Online. 2018;17(1):12.
138. Kruse C. The new possibilities from "Big Data" to overlooked associations between diabetes, biochemical parameters, glucose control, and osteoporosis. Curr Osteoporos Rep. 2018;16(3):320-324.
139. Pedoia V, Majumdar S, Link TM. Segmentation of joint and musculoskeletal tissue in the study of arthritis. MAGMA. 2016;29(2):207-221.
140. Gyftopoulos S, Lin D, Knoll F, Doshi AM, Rodrigues TC, Recht MP. Artificial intelligence in musculoskeletal imaging: current status and future directions. AJR Am J Roentgenol. 2019;213(3):506-513.
141. Gorelik N, Gyftopoulos S. Applications of artificial intelligence in musculoskeletal imaging: from the request to the report. Can Assoc Radiol. 2021;72(1):45-59.
142. Burns JE, Yao J, Summers RM. Artificial intelligence in musculoskeletal imaging: a paradigm shift. J Bone Miner Res. 2020;35(1):28-35.
143. Hung K, Montalvao C, Tanaka R, Kawai T, Bornstein MM. The use and performance of artificial intelligence applications in dental and maxillofacial radiology: a systematic review. Dentomaxillofac Radiol. 2020;49(1):20190107.
144. Hügle M, Omoumi P, van Laar JM, Boedecker J, Hügle T. Applied machine learning and artificial intelligence in rheumatology. Rheumatol Adv Pract. 2020;4(1):rkaa005.
145. Wani IM, Arora S. Computer-aided diagnosis systems for osteoporosis detection: a comprehensive survey. Med Biol Eng Comput. 2020; 58(9):1873-1917.