Published in final edited form as:

Curr Opin Rheumatol. 2019 July ; 31(4): 368–375. doi:10.1097/BOR.0000000000000607.

Artificial Intelligence, Osteoporosis and Fragility Fractures

Uran Ferizi^{1,*}, Stephen Honig², Gregory Chang¹

¹Department of Radiology, New York University School of Medicine, USA

²Division of Rheumatology, Department of Medicine, New York University School of Medicine, USA

Abstract

Purpose: The application of Artificial Intelligence tools has found new recent applications in medical applications. These tools have the potential to capture underlying trends and patterns, otherwise impossible with previous modeling capabilities. Both Machine Learning and Deep Learning models have found applications in osteoporosis, both to model the risk of fragility fracture, and to help with the identification and segmentation of images.

Findings: Here we survey the latest research in the AI application to the prediction of osteoporosis, published between January 2017 and March 2019. Around half of the articles covered predict (by classification or regression) an indicator of osteoporosis, such as bone mass or fragility fractures; the other half of studies use tools for automatic segmentation of the images of patients with or at risk of osteoporosis. The data for these studies includes diverse signal sources: acoustics, MRI, CT and of course X-rays.

Summary: New methods for automatic image segmentation, and prediction of fracture risk show promising clinical value. Though these recent developments have had a successful initial application to osteoporosis research, their development is still under improvements, such as accounting for positive-negative class bias. We urge care when reporting accuracy metrics, and when comparing between different studies. We also call for data transparency from study authors.

Keywords

osteoporosis; fragility fracture; Artificial Intelligence

1 INTRODUCTION

In this review, we will survey the latest research in the AI application to the prediction of osteoporosis. The studies considered are those published from January 2017 to March 2019, found through PubMed/Medline, using the search terms (“Bone” AND “Fragility Fracture” OR “Osteoporosis”) AND (“Artificial Intelligence” OR “Deep Learning” OR “Machine Learning”). From this list [1–34] we omitted the articles that do not explicitly use ML for disease prediction. These included studies in the field of Genetics that are related to osteoporosis, and some four other incomplete ones.

*Corresponding Author: Uran Ferizi, NYU Department of Radiology, 660 First Avenue, 4th floor, 10016, New York, NY, USA, Tel: +1 212 263 3398, [uran.ferizi\(at\)nyulangone.org](mailto:uran.ferizi(at)nyulangone.org).

A note on terminology:

because the AI tools we will describe have been developed and employed in many academic fields, such as Statistics, Computer Science, and Psychology, many terms are similar or the same thing, such as models vs. classifiers vs. algorithms; or features vs. covariates vs. predictors vs. parameters.

DIAGNOSIS of osteoporosis—One of major disease of aging, osteoporosis most commonly affects post-menopausal women, but is increasingly recognized as a clinical issue among older men. It is more common in white persons than in the other racial groups [35]. The standard-of-care test to diagnose osteoporosis involves the estimation of bone mineral density (BMD) via dual-energy X-ray absorptiometry (DXA) in the proximal femur, lumbar spine and, in some instances, the forearm. For the diagnosis, the BMD is then compared (by how much lower it is, in terms of standard deviations, or the T- score) with that of a reference group: a gender and ethnicity matched healthy, pre-menopausal adult population [36]. As defined by the World Health Organization (WHO), for osteoporosis, the BMD criterion requires a T-score of less than -2.5 ; normal BMD T-score is higher than -1.0 , and osteopenia is anything in-between T-scores -1 and -2.5 . Low bone density increases the incidence of fragility fractures [37,38]. So, in addition to the BMD criterion for the diagnosis of osteoporosis, the presence of fragility fractures renders a patient with “severe osteoporosis” [38]. In their lifetime, approximately, about 50% of women and 20% of men will incur an osteoporotic fracture [39].

The BMD estimation does correlate with bone weakness and fragility fracture [40]. However, in addition to using X-rays, an ionizing radiation (albeit at a very low dose), BMD via DXA also has other modeling limitations. DXA measurement assumes the presence of only bone and muscle, so fat will inevitably affect this measurement. Additionally, as it is a 2D-projection technique, it cannot fully account for bone geometry, size and microstructure [41]. This has led to many patients not receiving the indicated medical treatment. Moreover, population based studies have shown that the majority of women who incur fragility fractures have the BMD designation of osteopenia [42,43]. In particular, in a 2004 USA-wide study of just-under 150,000 postmenopausal women, part of the National Osteoporosis Risk Assessment (NORA) data, found that 82% of osteoporotic fractures were in women not meeting the diagnostic criterion for osteoporosis on the basis of DXA [44].

2 MODELS for prediction: background

In addition to BMD, other risk factors that contribute to fragility fractures are age, previous fragility fracture, previous occurrence of hip fracture in a parent, and use of corticosteroids [38]. Currently widely used, a fracture risk assessment tool, FRAX [45], is a model that uses twelve input parameters: age, sex, weight, height, previous fracture, parental fracture history, rheumatoid arthritis, use of glucocorticoids, secondary osteoporosis, smoking and alcohol consumption. Its output, the 10-year probability of fracture, has received some criticism. Described limitations include its modeling assumptions (e.g. in FRAX most parameters, such as smoking, are dichotomously entered, regardless of their severity), and its possible

overuse as a tool to justify pharmaceutical treatment for improving BMD [46–48]. Some studies suggest other simpler models that may perform better [49].

Criticizing models is easy. However, to paraphrase the great statistician George Box, *all models are wrong, but some are useful* [50]. Recently, the list of models that aim to capture the risk of fragility fracture has been vastly increased by the application of Artificial Intelligence tools, in the hope of making the predictions ever more useful. These models go beyond the linear or nonlinear combination of all the input risk factors, such as those of linear or Poisson regression. The AI tools have the potential to capture underlying trends and patterns, otherwise impossible with previous modeling tools; they learn from data.

2.1 Machine Learning and Deep Learning

The arsenal of AI tools is vast and ever-expanding. Here we will give a brief outline of the models and methods; however, a more detailed and wonderful description can be taken from classic textbooks on the subject [51–53]. Broadly speaking, the algorithms can be described as being supervised or unsupervised. The class of supervised models, which is the most popular, maps the input variables into an output measure. The relative importance of these input variables in producing the output can be changed by way of *weights*, depending on whether the prediction distance from the outcome increases or decreases during the training stage. The unsupervised models do not have such measures, but instead aim to describe the inputs groupings. Occasionally, some algorithms can straddle both types, and those are described as semisupervised. Depending on the type of the outcome prediction, the task is termed *classification* if the output is categorical; specifically, this is the case when the model is required to classify a group of subjects into those with or without osteoporosis. The task is termed *regression*, if the output is quantitative; this is the case when the model is required to predict the BMD of each subject in the osteoporosis dataset. The third way is to predict ordered categorical variables of outcome: e.g. individuals with normal, osteopenia and osteoporosis T-score.

2.2 The AI algorithms

The well-established **linear and logistic regression** are still used in modelling prediction. However, the choice of models has vastly increased in the recent decades: which Machine Learning algorithm to use depends on the task it is required for, such as prediction requirements (classification/regression/clustering), the richness of the data, or computational time requirements. One of the simplest for classification is **Naïve Bayes**, which generally exhibits high bias and low variance; so, while it is a quick and simple classifier for simple datasets, it may become less appropriate as the dataset becomes richer. **k-Nearest Neighbor** (kNN) is a simple algorithm that performs classification or regression depending on the distances between them; i.e. a classifier of association. In classification problems, the **Support Vector Machine** (SVM) divides the data space into two classes via hyperplanes (types of geometric subspaces of dimension $n-1$ embedded in the space of dimension n); SVM's are good at handling complex, nonlinear scenarios, and tend not to over-fit. Classification or Regression **Decision Trees** makes their decisions from root, all the way up to the branches; they are better than kNN in ignoring irrelevant features, but they are prone to overfitting, with which ensemble learning can help. **Ensemble learning** combines the

strengths of a collection of simpler base models: in *boosting*, the learners evolves over time, in the end producing a weighted result; in bootstrap aggregation or *bagging*, e.g. bagged decision trees, the prediction is averaged over a collection of bootstrap samples of decision trees.

Deep Learning is a collection of algorithms that use neural networks as an architecture, and the leaning is performed through multiple layers of abstraction [54–56]. (Artificial) **Neural Networks** (ANN or NN) are a class of nonlinear statistical algorithms modeled, as the name suggests, on the processing of information done by brain's neurons. The model is defined by how the elements of the network are connected and the weights of these connections. In simplified terms, this learning constructs 'derived parameters' as linear combinations of the input parameters, and then further models the outcome as a nonlinear construction of these derived parameters. Though they are good at handling a large number of inputs, they are rather computationally expensive.

In the **Convolutional Neural Net** (CNN), a very common deep learning algorithm, some of the *weights* (as explained previously) are shared or convolved.

2.3 Stability of Prediction: Cross Validation

Overfitting generally occurs when the complexity of the model is insufficiently supported by the richness of the data; this results in the model not being able to generalize outside of the given dataset. So, in computational modelling, the robustness of fitting to variations in the data is usually tested via statistical subsampling of the dataset, using methods such as bootstrapping or cross-validation. Bootstrapping builds a dataset by sampling with replacement from the original dataset.

Cross-Validation (CV) is rather more popular, and is usually performed in two ways, when training the supervised model and then testing its performance. In the k-fold CV, the model is fitted to the (k-1)/k-th portion of the data, and this fitting's predictions are then compared with the remaining 1/k-th portion of the data. Similarly, in the holdout CV method, one holds a given percentage for testing, and then trains the classifier on the remainder; Leave-One-Out CrossValidation (LOOCV) is an example of such a method.

3 APPLICATIONS to Osteoporosis

Both Machine Learning and Deep Learning models have found applications in osteoporosis. Most of the articles we report on serve either to predict an indicator of OP, such as BMD or fractures, shown in the top half of Tables 1 and 2, or as a tool for automatic segmentation of the images of patients with or at risk of OP. In the first group which consists of eight studies, seven of them use the supervised category of models. Of these, three predict categorical outcomes, specifically the fractures/no-fracture classes [4,8,13] and osteonecrosis [11]; while three predict quantitatively the BMD value. In this group of eight, there also is one unsupervised study, which finds 9 clusters of fracture risk among the population it studies [14].

The second group consists of seven studies, which make more explicit use of image analysis. Often mentioned under texture analysis, the AI models are used to spot pattern in the images, and find underlying connections between groups.

It is encouraging to see a variety of techniques being used to probe bone health, beyond X-ray imaging. Scanlan [22] and Vogl [31] use bone acoustics to infer bone health. Kim [11] and Hwang [10] use dental radiographs to predict jaw osteonecrosis and BMD, respectively. Ferizi [4] and Deniz [57] use MRI too to help with the diagnosis and segmentation of the images, respectively. As the authors suggest, more research is needed to improve such techniques.

While some studies collect data over a long period, especially those concerning multi-center cohorts [3,7,11], two studies are specifically longitudinal by design. Using data from a larger study, the Dubbo Osteoporosis Epidemiology Study (DOES), this 10-year longitudinal study aims to predict the fractures across 1,167 Australian women. Similarly, using the measurements from the femoral and lumbar regions of 135 Japanese women, the models of Shioji [25] predict the BMD and bone mass loss over a ten-year period. Both of these studies conclude that ANN performed best with such tasks.

An interesting question for the clinician concerns the optimal age for osteoporosis screening. The 2011 U.S. Preventive Services Task Force (USPSTF) recommended screening women for osteoporosis at the age of 65, and cites insufficient evidence in the case for men [35]. Kruse [14] identifies 60 as the age that distinguishes *average* from *high fracture* risk; this study, on a Danish population, gives no information on the male/female ratio of the cohort.

Aside from age, two other important attributes are sex and geographical origin. Though osteoporosis affects mostly women, and this is reflected by the fifteen studies under review, where all but one concentrates on women, the disease affects men too. Specifically, six studies include men in the dataset [1,10,11,13,22,28]; one had no information on the population sex [7]).

As for their origin, the subjects under this review were of many nationalities: three studies concerned the USA [4,7,57], two studies for each of South Korea [10,11] and Denmark [13,14] and one each for Australia [8], Japan [25], Switzerland [31], UK [22], France [26], Austria [3], Germany [28] and India [1].

As mentioned earlier, the more complex models require rich datasets for the modeling predictions to be useful; this is especially the case for the Deep Learning (Neural Network) models. We can mention at least three factors that affect this richness. First, the obvious: the size of the dataset. In our studies, the cohort sizes varied from 19 (Dimai [3]) to 10,775 (Kruse [14]). Second, the more pertinent information for each datum/subject, i.e. number of features, the better. In our studies, these varied from the c.75,000 in Kruse [13], which then keeps the most relevant by removing those features that produce low variance to arrive at a smaller c.100 number; even an 11-feature set achieves a good AUC=0.91 for the women cohort, versus AUC=0.92 for nearly ten times more features.

Cross validation can help with overfitting in modest datasets. Most of the studies in the review use some form of CV. However, variations in the application of this technique need to be borne in mind when using the performance metrics. When referring to overfitting, modelers often have the complexity of the model in mind *vis-à-vis* the dataset: too complex a model will tune in the parameters to incorporate spurious effects in the data, thus reducing the predictability of the model outside the given dataset. So, the ability to generalize the results remains dependent on the size of the dataset.

We reiterate here some previous warnings about experiment design and results interpretation.

Label Class Imbalance:

It is not clear if any of the studies account for the label class imbalance, specifically the number of positives and negatives in the dataset (in our review, three studies [1,10,26] do have their classes balanced by design).

Performance Metrics:

The indicators of performance in the articles are mostly expressed in terms of the well-established sensitivity and specificity, as well as accuracy and the area under the Receiver Operating Characteristic (ROC), the AUC. It is easy to see why accuracy alone can give a misleading picture of model performance, especially in label-unbalanced datasets: e.g. a model is *a priori* 75% accurate if it only predicts negatives in a cohort with only 25% positives. However, the tools from the studies we consider are not indented for screening purposes; instead the cohorts are preselected for diagnosis, thus some of the information on the prevalence is lost. But, when such information exists, to properly calibrate the accuracy, prevalence needs to be factored in [58]: $\text{Accuracy} = \text{Sensitivity} * \text{Prevalence} + \text{Specificity} * (1 - \text{Prevalence})$.

The other most popular of the performance metrics, the AUC, reflects the model performance across different thresholds of sensitivity and specificity, on the assumption that this weighting is the same across all diseases [59]. In screening studies, these indicators alone can be misleading, and false-discovery-rate is necessary [58]; in this review, all but one (Scanlan [22]) use it.

4 CONCLUSION

Albeit still in development, recent developments in AI have had a successful application in aid of osteoporosis diagnosis. New methods for automatic image segmentation are a much-needed help to the rheumatologist and radiologist. Care should be taken when reporting the performance metrics and comparing studies as, besides being evaluated on different datasets, they may even be evaluated differently. A greater willingness for data and model transparency will help the field even more, and make the studies more reproducible.

Acknowledgments

Grant support: NIH grants: R01 AR 066008; NIH R01 AR 070131

REFERENCES

- *1. 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):665–73. [PubMed: 29198076] * This pilot study defines as “healthy” all subjects whose BMD does not fall under osteopenia or osteoporosis. The images are pre-selected/segmented, where 21 of poor quality (of the total 138) are discarded. Statistical tests are used to select the 13 most important features from a total of 85 (5 from the radiogram, and 80 from the texture analysis).
2. Ashok Kumar D, Anburajan M, Snehalatha 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):1350–71. [PubMed: 29968333]
- **3. Dimai HP, Ljuhar R, Ljuhar D, Norman B, Nehrer S, Kurth A, et al. Assessing the effects of longterm osteoporosis treatment by using conventional spine radiographs: results from a pilot study in a subcohort of a large randomized controlled trial. *Skeletal Radiol*. 2018.** This study uses data from 19 patients, taken from a larger FREEDOM pivotal trial which studied 7808 postmenopausal women with OP. The patients were prescribed Denosumab, after 8 years of which the overall BMD increased by 42%. The study compares Bone Structure Value (BSV) with BMD. BSV is estimated through a ML procedure, combined with a bone microstructural model based on Fractal Mathematics. The findings were supported by an ex vivo HR-pQCT study performed in human vertebrae.
- *4. Ferizi U, Besser H, Hysi P, Jacobs J, Rajapakse CS, Chen C, 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*. 2018.* Comparison of 15 classifiers using MRI data of different femoral regions: head, neck, intertroch, troch, ward. Both MRI and FRAX data independently add value to the prediction.
5. Gao F, Xu F, Wu D, Cheng J, Xia P. Identification of novel genes associated with fracture healing in osteoporosis induced by Krm2 overexpression or Lrp5 deficiency. *Mol Med Rep*. 2017;15(6):3969–76. [PubMed: 28487939]
6. Gibson D, Drabkin A, Krantz MJ, Mascolo M, Rosen E, Sachs K, et al. Critical gaps in the medical knowledge base of eating disorders. *Eat Weight Disord*. 2018;23(4):419–30. [PubMed: 29681012]
- *7. Gonzalez G, Washko GR, Estepar RSJ. Deep learning for biomarker regression: application to osteoporosis and emphysema on chest CT scans. *Proc SPIE Int Soc Opt Eng*. 2018;10574.* In this extremely rich dataset, the CT scans of individuals with at least 10 pack-years of smoking are used to predict BMD and lung percentage of emphysema. The proposed method requires 2D field-of-view of images for biomarker computation.
- *8. Ho-Le TP, Center JR, Eisman JA, Nguyen TV, Nguyen HT. Prediction of hip fracture in postmenopausal women using artificial neural network approach. *Conf Proc IEEE Eng Med Biol Soc*. 2017;2017:4207–10. [PubMed: 29060825] *A subset of the Dubbo Osteoporosis Epidemiology Study (DOES), this 10-year longitudinal study records the fractures across 1167 women over 60 years of age, using the Neural Network to predict the fracture cases. Additionally, accuracy (using BMD only)=0.82; accuracy (non-BMD only)=0.84; accuracy (BMD and non-BMD)=0.87.
9. Hussain D, Al-Antari MA, Al-Masni MA, Han SM, Kim TS. Femur segmentation in DXA imaging using a machine learning decision tree. *J Xray Sci Technol*. 2018;26(5):727–46. [PubMed: 30056442]
- *10. Hwang JJ, Lee JH, Han SS, Kim YH, Jeong HG, Choi YJ, et al. Strut analysis for osteoporosis detection model using dental panoramic radiography. *Dentomaxillofac Radiol*. 2017;46(7):20170006.* Classifies panoramic dental radiographs into OP/non-OP (this excludes any subject in the osteopenia bracket). The study suggests that the strut method in the endosteal margin area correlates strongly with OP.
- *11. Kim DW, Kim H, Nam W, Kim HJ, Cha IH. Machine learning to predict the occurrence of bisphosphonate-related osteonecrosis of the jaw associated with dental extraction: A preliminary report. *Bone*. 2018;116:207–14. [PubMed: 29698784] * All patients selected were under (past/present) biophosphonate treatment. Parameters include serum CTX level and corticosteroid use.

12. 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):320–4. [PubMed: 29679305]
- *13. Kruse C, Eiken P, Vestergaard P. Machine Learning Principles Can Improve Hip Fracture Prediction. *Calcif Tissue Int*. 2017;100(4):348–60. [PubMed: 28197643] *24 statistical models trained on 75% of the data, via 5-fold, 5-times repeated CV. The analysis was performed separately for data of men and women, gathered during 1996–2006. Uses imputation techniques for missing data. Starts off with 75000-rich parameter list and removes those with low variance, to achieve around 100 parameters. An 11-predictor subset delivers AUC=0.91 for women; a 10-predictor subset delivers AUC=0.86 for men. Parameters include, among others, cholesterol tests, penicillin use, GP and dentist visits.
- *14. Kruse C, Eiken P, Vestergaard P. Clinical fracture risk evaluated by hierarchical agglomerative clustering. *Osteoporos Int*. 2017;28(3):819–32. [PubMed: 27848006] * The study uses Hierarchical Agglomerative Clustering to find 9 clusters across 10775 Danish women involved in the study (size of clusters varies from 26.4 to 1.3% of the cohort.). The dataset is very rich with information, including GP and physical therapy expenses. The study identifies 60 as age that distinguishes average from high fracture risk.
15. Kruse C, Goemaere S, De Buyser S, Lapauw B, Eiken P, Vestergaard P. Predicting mortality and incident immobility in older Belgian men by characteristics related to sarcopenia and frailty. *Osteoporos Int*. 2018;29(6):1437–45. [PubMed: 29569152]
16. Liang BC, Shi XL, Li CW, Shi ZY, He WT, Yao JL, et al. Identification of human serum protein targets of Qianggu Decoction () in primary type I osteoporosis based on tandem mass tag labeling and liquid chromatography-tandem mass spectrometry technology. *Chin J Integr Med*. 2017;23(10):747–54. [PubMed: 27389089]
17. Ma M, Luo S, Zhou W, Lu L, Cai J, Yuan F, et al. Bioinformatics analysis of gene expression profiles in B cells of postmenopausal osteoporosis patients. *Taiwan J Obstet Gynecol*. 2017;56(2):165–70. [PubMed: 28420501]
18. Malhan D, Muelke M, Rosch S, Schaefer AB, Merboth F, Weisweiler D, et al. An Optimized Approach to Perform Bone Histomorphometry. *Front Endocrinol (Lausanne)*. 2018;9:666. [PubMed: 30519215]
19. Muehlematter UJ, Mannil M, Becker AS, Vokinger KN, Finkenstaedt T, Osterhoff G, et al. Vertebral body insufficiency fractures: detection of vertebrae at risk on standard CT images using texture analysis and machine learning. *European Radiology*. 2018.
20. Oulhaj H, Rziza M, Amine A, Toumi H, Lespessailles E, El Hassouni M, et al. Anisotropic Discrete Dual-Tree Wavelet Transform for Improved Classification of Trabecular Bone. *IEEE Trans Med Imaging*. 2017;36(10):2077–86. [PubMed: 28574347]
21. Qaseem A, Forciea MA, McLean RM, Denberg TD. Treatment of Low Bone Density or Osteoporosis to Prevent Fractures in Men and Women: A Clinical Practice Guideline Update From the American College of Physicians. *Ann Intern Med*. 2017;166(11):818–39. [PubMed: 28492856]
- **22. Scanlan J, Li FF, Umnova O, Rakoczy G, Lovey N, Scanlan P. Detection of Osteoporosis from Percussion Responses Using an Electronic Stethoscope and Machine Learning. *Bioengineering (Basel)*. 2018;5(4).** This study uses an electronic stethoscope to map percussion response to BMD/Osteoporosis/FRAX. Though the outcomes could do better, the method is innovative and refreshing. (See also Vogl. et al. [31])
23. Schulze F, Malhan D, El Khassawna T, Heiss C, Seckinger A, Hose D, et al. A tissue-based approach to selection of reference genes for quantitative real-time PCR in a sheep osteoporosis model. *BMC Genomics*. 2017;18(1):975. [PubMed: 29258442]
24. Shigeta H, Mashita T, Kikuta J, Seno S, Takemura H, Ishii M, et al. Bone marrow cavity segmentation using graph-cuts with wavelet-based texture feature. *J Bioinform Comput Biol*. 2017;15(5):1740004.
25. 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. [PubMed: 29126439] * This study looks at the predictability, via ML tools, of Bone Loss Rate (BMD in 2003 minus BMD in 1993) and Bone Mineral Density, for femoral and lumbar regions in 135 over-50-year old Japanese women. The study also performs

model comparison via AIC/BIC (but it is not clear how the complexity of the model, i.e. number of parameters, are estimated). The 11 parameters include age at menopause and body fat percentage.

- *26. Singh A, Dutta MK, Jennane R, Lespessailles E. Classification of the trabecular bone structure of osteoporotic patients using machine vision. *Comput Biol Med.* 2017;91:148–58. [PubMed: 29059592] * This study uses statistical features of the X-ray image as input (e.g. mean and std of pixels across the entire image, or intensity differences across an image).
27. Souza LS, Rochette NF, Pedrosa DF, Magnago RPL, Filho TBF, Vieira FLH, et al. Role of APOE Gene in Bone Mineral Density and Incidence of Bone Fractures in Brazilian Postmenopausal Women. *J Clin Densitom.* 2018;21(2):227–35. [PubMed: 28784276]
- *28. Valentinitich A, Trebeschi S, Kaesmacher J, Lorenz C, Löffler MT, Zimmer C, et al. Opportunistic osteoporosis screening in multi-detector CT images via local classification of textures. *Osteoporos Int.* 2019.* Unlike other studies, the majority of this cohort are men. Diagnosis of OP based on vertebral fractures.
29. van de Peppel J, Strini T, Tilburg J, Westerhoff H, van Wijnen AJ, van Leeuwen JP. Identification of Three Early Phases of Cell-Fate Determination during Osteogenic and Adipogenic Differentiation by Transcription Factor Dynamics. *Stem Cell Reports.* 2017;8(4):947–60. [PubMed: 28344004]
30. Varghese V, Krishnan V, Kumar GS. Evaluating Pedicle-Screw Instrumentation Using Decision-Tree Analysis Based on Pullout Strength. *Asian Spine J.* 2018;12(4):611–21. [PubMed: 30060368]
- **31. Vogl F, Friesenbichler B, Husken 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.* 2018.** In this exciting pilot study, low-frequency (acoustic) waves are used to predict the BMD. The categories are “healthy” and “osteopenia/osteoporosis” subjects, as determined by their BMD. (See also Scanlan.)
32. Wang J, Yan D, Zhao A, Hou X, Zheng X, Chen P, et al. Discovery of potential biomarkers for osteoporosis using LC-MS/MS metabolomic methods. *Osteoporos Int.* 2019.
33. Wang Y, van Assen AHG, Reis CR, Setroikromo R, van Merkerk R, Boersma YL, et al. Novel RANKL DE-loop mutants antagonize RANK-mediated osteoclastogenesis. *Febs j.* 2017;284(15):2501–12. [PubMed: 28627025]
34. Zhu W, Shen H, Zhang JG, Zhang L, Zeng Y, Huang HL, et al. Cytosolic proteome profiling of monocytes for male osteoporosis. *Osteoporos Int.* 2017;28(3):1035–46. [PubMed: 27844135]
35. Force UPST. Screening for osteoporosis: US preventive services task force recommendation statement. *Ann Intern Med.* 2011;154(5):356. [PubMed: 21242341]
36. Kanis JA. Assessment of fracture risk and its application to screening for postmenopausal osteoporosis: synopsis of a WHO report. WHO Study Group. *Osteoporos Int.* 1994;4(6):368–81. [PubMed: 7696835]
37. Consensus A Consensus development conference: diagnosis, prophylaxis, and treatment of osteoporosis. *Am J Med.* 1993;94(6):646–50. [PubMed: 8506892]
38. Kanis JA. Diagnosis of osteoporosis and assessment of fracture risk. *The Lancet.* 2002;359(9321):1929–36.
39. Health UDo, Services H. Bone health and osteoporosis: a report of the Surgeon General. Rockville, MD: US Department of Health and Human Services, Office of the Surgeon General 2004;87.
40. Marshall D, Johnell O, Wedel H. Meta-analysis of how well measures of bone mineral density predict occurrence of osteoporotic fractures. *Bmj.* 1996;312(7041):1254–9. [PubMed: 8634613]
41. Chang G, Honig S, Brown R, Deniz CM, Egol KA, Babb JS, et al. Finite element analysis applied to 3-T MR imaging of proximal femur microarchitecture: lower bone strength in patients with fragility fractures compared with control subjects. *Radiology.* 2014;272(2):464–74. [PubMed: 24689884]
42. Schuit S, Van der Klift M, Weel A, De Laet C, Burger H, Seeman E, et al. Fracture incidence and association with bone mineral density in elderly men and women: the Rotterdam Study. *Bone.* 2004;34(1):195–202. [PubMed: 14751578]
43. Wainwright SA, Marshall LM, Ensrud KE, Cauley JA, Black DM, Hillier TA, et al. Hip fracture in women without osteoporosis. *The Journal of Clinical Endocrinology & Metabolism.* 2005;90(5):2787–93. [PubMed: 15728213]

44. Siris ES, Chen YT, Abbott TA, Barrett-Connor E, Miller PD, Wehren LE, et al. Bone mineral density thresholds for pharmacological intervention to prevent fractures. *Arch Intern Med*. 2004;164(10):1108–12. [PubMed: 15159268]
45. Kanis JA, Johnell O, Oden A, Johansson H, McCloskey E. FRAX™ and the assessment of fracture probability in men and women from the UK. *Osteoporosis International*. 2008;19(4):385–97. [PubMed: 18292978]
46. Collins GS, Michaëlsson K. Fracture Risk Assessment: State of the Art, Methodologically Unsound, or Poorly Reported? *Current Osteoporosis Reports*. 2012;10(3):199–207. [PubMed: 22688862]
47. Järvinen TL, Jokihaara J, Guy P, Alonso-Coello P, Collins GS, Michaëlsson K, et al. Conflicts at the heart of the FRAX tool. *Canadian Medical Association Journal*. 2014;186(3):165–7. [PubMed: 24366895]
48. Silverman SL, Calderon AD. The utility and limitations of FRAX: a US perspective. *Current Osteoporosis Reports*. 2010;8(4):192–7. [PubMed: 20811963]
49. Rubin KH, Friis-Holmberg T, Hermann AP, Abrahamsen B, Brixen K. Risk assessment tools to identify women with increased risk of osteoporotic fracture: complexity or simplicity? A systematic review. *Journal of Bone and Mineral Research*. 2013;28(8):1701–17. [PubMed: 23592255]
50. Box GE. Science and statistics. *Journal of the American Statistical Association*. 1976;71(356):791–9.
51. Bishop CM. *Pattern recognition and machine learning*: springer; 2006.
52. Friedman J, Hastie T, Tibshirani R. *The elements of statistical learning*: Springer series in statistics New York, NY, USA; 2001.
53. Michie DS DJ; Taylor CC *Machine Learning: Neural and Statistical Classification*. Hertfordshire, UK: Ellis Horwood; 1994 290 p.
54. Le QV. A tutorial on deep learning part 1: Nonlinear classifiers and the backpropagation algorithm. Google Inc, Mountain View, CA 2015:18.
55. Le QV. A tutorial on deep learning part 2: Autoencoders, convolutional neural networks and recurrent neural networks. *Google Brain*. 2015:1–20.
56. LeCun Y, Bengio Y, Hinton G. Deep learning. *nature*. 2015;521(7553):436. [PubMed: 26017442]
- *57. Deniz CM, Xiang S, Hallyburton RS, Welbeck A, Babb JS, Honig S, et al. Segmentation of the proximal femur from MR images using deep convolutional neural networks. *Scientific Reports*. 2018;8(1):16485. [PubMed: 30405145] * Using volumetric structured MRI, this work compares 2D/3D CNN's with multiple number of initial feature maps, layers, and dilation rates.
58. Colquhoun D An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society open science*. 2014;1(3).
59. Halligan S, Altman DG, Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach. *European Radiology*. 2015;25(4):932–9. [PubMed: 25599932]

Key Points:

- Machine Learning tools show promise in predicting fragility fracture risk in the people at risk of osteoporosis; for large “Big Data” sets, Deep Learning models can be helpful too.
- The ability of Deep Learning models to search for subtle patterns can help with the segmentation of images.
- A caution that the metrics of performance are only as good as the quality of the training and testing of the models, as well as the data these models are fitted to.

Table 1:

The articles included in the current review.

Author: Title	Publish	Location	Subjects	%Women	%OP/ fract	Mean Age (Years)
Kruse: <i>Clinical fracture risk evaluated by hierarchical agglomerative clustering.</i> [1]	03/2017	Denmark	10,775	100		
Kruse: <i>Machine Learning Principles Can Improve Hip Fracture Prediction.</i> [2]	04/2017	Denmark	5,439	87		men=69 (fracture) men=62 (noFracture) women=75 (fracture) women=60 (noFracture)
Ho-Le: <i>Prediction of hip fracture in postmenopausal women using artificial neural network approach.</i> [3]	07/2017	Australia	1,167	100	23	77(fracture) 69 (noFracture)
Shioji: <i>Artificial neural networks to predict future bone mineral density and bone loss rate in Japanese postmenopausal women.</i> [4]	11/2017	Japan	135	100	-	65
Ferizi: <i>Artificial Intelligence Applied to Osteoporosis: A Performance Comparison of Machine Learning Algorithms in Predicting Fragility Fractures from MRI Data.</i> [5]	09/2018	USA	92	100	35	32(fracture) 60 (noFracture)
Kim: <i>Machine learning to predict the occurrence of bisphosphonate-related osteonecrosis of the jaw associated with dental extraction: A preliminary report.</i> [6]	11/2018	S. Korea	125	95	33	74(necrosis) 71(noNecrosis)
Vogl: <i>Can low-frequency guided waves at the tibia paired with machine learning differentiate between healthy and osteopenic/osteoporotic subjects? A pilot study.</i> [7]	11/2018	Switzerland	41	100	37	61
Scanlan: <i>Detection of Osteoporosis from Percussion Responses Using an Electronic Stethoscope and Machine Learning.</i> [8]	12/2018	UK	110	82	31	68 (OP) 64 (OK)
Hwang: <i>Strut analysis for osteoporosis detection model using dental panoramic radiography.</i> [9]>	10/2017	S. Korea	454	79	50	64 (OP-BMD) 57 (OK)
Deniz: <i>Segmentation of the Proximal Femur from MR Images using Deep Convolutional Neural Networks.</i> [10]	11/2018	USA	86	100	42	(postmenopausal"
Singh: <i>Classification of the trabecular bone structure of osteoporotic patients using machine vision.</i> [11]	12/2017	France	174	100	50	71 (OP and fracture) 69 (OK)
González: <i>Deep learning for biomarker regression: application to osteoporosis and emphysema on chest CT scans.</i> [12]	02/2018	USA	9,925	-	-	-
Areeckal: <i>Early diagnosis of osteoporosis using radiogrammetry and texture analysis from hand and wrist radiographs in Indian population.</i> [13]	03/2018	India	117	58	49	52 (T-score<-1)) 45 (OK)
Dimai: <i>Assessing the effects of long-term osteoporosis treatment by using conventional spine radiographs: results from a pilot study in a sub-cohort of a large randomized controlled trial.</i> [14]	12/2018	Austria	19	100	0	69 (OP)
Valentinitsch: <i>Opportunistic osteoporosis screening in multi-detector CT images via local classification of textures.</i> [15]	03/2019	Germany	154	33	34	67 (OP) 64 (OK)

Table 2:

The specific implementations of the algorithms used in the studies included in the review.

Author	Using Data	Via ML Tools	To Predict	Result (for best model)	CV training/testing
Kruse [1]	BMD spine and hip	agglomerative clustering	clusters of fracture risk	9 clusters, with low/high fracture risk	90% training 10% testing 10,000 times repeated
Kruse [2]	BMD of hip	ensemble (bagging and boosting)	hip fractures	AUC=0.92(women) AUC=0.91(men)	75% training (5-fold, 5-times repeated); 25% testing
Ho-Le [3]	BMD spine and hip	ANN (best), LR, SVM, kNN	hip fractures	AUC=0.94 accuracy=0.87	60% training (5-fold, 5-times repeated); 40% testing (50 times repeated)
Shioji [4]	BMD spine and femur (as of 1993)	ANN (best), MLR	BMD of spine and femur (as of 2003)	R ² =0.93(spine);0.88(femur) sens=0.80(spine);0.81(femur) spec=0.91(spine);0.93(femur)	5-fold
Ferizi [5]	BMD / MRI of hip	EN(best), LR(best), kNN, DT, SVM	fractures	sens=0.62 spec=0.67	23-fold 200 times repeated
Kim [6]	dental extractions	RF (best), ANN, DT, SVM, MLR	biophosphonate-related osteonecrosis of the jaw	AUC=0.97 sens=1.00 spec=0.83	70% training 30% testing
Vogl [7]	acoustic test of tibia	SVM	BMD (lumbar spine, femoral neck; distal radius)	AUC=0.83(radius) AUC=0.71(hip) AUC=0.56(spine)	nested CV: 10-fold (inner) 6-fold (outer)
Scanlan [8]	acoustic test of tibia	ANN	BMD	accuracy=0.70; false-positive-rate=0.30	-
Hwang [9]	dental X-rays	SVM (best), DT	BMD of spine and hip	sens=0.97 spec=0.97 accuracy=0.97	10-fold
Deniz [10]	MRI of hip	deep CNN	manual segmentation	dice-similarity-score=0.95 precision=0.95 recall=0.95	4-fold
Singh [11]	X-rays of calcaneus bone	SVM (best), kNN, NB, ANN	fractures	accuracy=0.97 sens=0.98 spec=0.95	Leave-One-Out CV
González [12]	thoracic CT	CNN for regression	spine BMD (of manually segmented images)	Pearson r=0.94	80% used for training 10% for validating and 10% for testing
Areeckal [13]	X-rays of hand/wrist	ANN	manual segmentation	sens=0.80 spec=1.00 accuracy=0.89	75% training 25% testing
Dimai [14]	X-rays of spine	-	BMD	Pearson correlation (of BMD with BSV) r = 0.51	-
Valentinitsch [15]	CT of spine	RF	BMD	AUC=0.88	4-fold 10 times repeated

Abbreviations

BMD=Bone Mass/Mineral Density
CNN=Convolutional Neural Networks
ANN=Artificial Neural Networks
NN=Neural Networks
DT=Decision Trees
MLR=Multiple Linear Regression
SVM=Support Vector Machine

kNN=k-Nearest Neighbors
NB=Naïve Bayes
OP=Osteoporosis/OK=Normal
CT=Computerized Tomography
BSV=Bone Structure Value
sens=sensitivity
spec=specificity

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript