BMC Medical Imaging

SYSTEMATIC REVIEW Open Access

Deep learning-based evaluation of panoramic radiographs for osteoporosis screening: a systematic review and meta-analysis

Ali Tarighatnia[1], Masoud Amanzadeh[2*], Mahnaz Hamedan[2], Alireza Mohammadnia[2] and Nader D. Nader[3]

Abstract

Background Osteoporosis is a complex condition that drives research into its causes, diagnosis, treatment, and prevention, significantly affecting patients and healthcare providers in various aspects of life. Research is exploring orthopantomogram (OPG) radiography for osteoporosis screening instead of bone mineral density (BMD) assessments. Although this method uses various indicators, manual analysis can be challenging. Machine learning and deep learning techniques have been developed to address this. This systematic review and meta-analysis is the first to evaluate the accuracy of deep learning models in predicting osteoporosis from OPG radiographs, providing evidence for their performance and clinical use.

Methods A literature search was conducted in MEDLINE, Scopus, and Web of Science up to February 10, 2025, using the keywords related to deep learning, osteoporosis, and panoramic radiography. We conducted title, abstract, and full-text screening based on inclusion/exclusion criteria. Meta-analysis was performed using a bivariate random-effects model to pool diagnostic accuracy measures, and subgroup analyses explored sources of heterogeneity.

Results We found 204 articles, removed 189 duplicates and irrelevant studies, assessed 15articles, and ultimately, seven studies were selected. The DL models showed AUC values of 66.8–99.8%, with sensitivity and specificity ranging from 59 to 97% and 64.9–100%, respectively. No significant differences in diagnostic accuracy were found among subgroups. AlexNet had the highest performance, achieving a sensitivity of 0.89 and a specificity of 0.99. Sensitivity analysis revealed that excluding outliers had little impact on the results. Deeks' funnel plot indicated no significant publication bias ($P=0.54$).

Conclusions This systematic review indicates that deep learning models for osteoporosis diagnosis achieved 80% sensitivity, 92% specificity, and 93% AUC. Models like AlexNet and ResNet demonstrate effectiveness. These findings suggest that DL models are promising for noninvasive early detection, but more extensive multicenter studies are necessary to validate their efficacy in at-risk groups.

Keywords Osteoporosis, Panoramic radiography, Deep learning, OPG, Diagnosis

*Correspondence:
Masoud Amanzadeh
M.amanzadeh@arums.ac.ir
[1]Department of Medical Physics, School of Medicine, Ardabil University of Medical Sciences, Ardabil, Iran
[2]Department of Health Information Management, School of Medicine, Ardabil University of Medical Sciences, Ardabil, Iran
[3]Department of Anesthesiology, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, NY, USA

Background

Osteoporosis is a complex, multidisciplinary disease that has led researchers from various medical fields to explore its etiology, detection, diagnosis, treatment, and prevention [1]. This silent disease presents significant challenges not only for patients but also for healthcare providers, with profound occupational, social, and economic implications [2]. As the prevalence of osteoporosis rises, projections estimate that over 263 million individuals will be affected by 2034. Early diagnosis and timely screening are crucial in reducing the serious health complications associated with this condition [3, 4].

Timely diagnosis and screening for osteoporosis are crucial in mitigating its serious health complications [5, 6]. Among the various imaging techniques employed for osteoporosis detection, dual-energy X-ray absorptiometry (DEXA) is recognized as the gold standard for its accuracy, provided it is correctly executed with appropriate quantitative analyses [7]. However, DEXA can present challenges for patients with metal implants, incomplete or inaccurate health information, or those who cannot fully cooperate during the imaging process [8]. In such cases, clinicians may opt to analyze different body parts, such as the forearm, hip, or lumbar regions, or combine them to improve diagnostic precision. Alternative methods, like orthopantomogram (OPG) radiography, are also being investigated for their potential in osteoporosis screening [9].

OPG is a cost-effective imaging technique primarily used for evaluating the upper and lower jaws, diagnosing dental diseases, jaw injuries, and disorders, and screening for osteoporosis. It offers several indices for assessing osteoporosis, including the lamina dura's width, the mandibular cortex index (MCI), and the ante-gonial index [10]. However, the manual measurement of these indices is often labor-intensive and time-consuming, which can hinder effective feature recognition in radiographic images and impact the reproducibility of classification methods. Artificial intelligence (AI) and deep learning (DL), a subset of AI, are transformative tools in medical imaging that enhance and automate diagnostics while accurately analyzing complex medical images. Convolutional neural networks (CNNs), a prevalent DL architecture, are particularly effective at pattern recognition in medical images, making them ideal for osteoporosis screening. By automating the detection and classification of radiographic indices, DL models reduce subjectivity, improve diagnostic consistency, and address the limitations of manual analysis. DL algorithms like AlexNet, ResNet, and VGG have been developed to assess osteoporosis from panoramic radiographs. These models improve the extraction of complex imaging features, enhancing sensitivity and specificity over traditional methods. As deep learning continues to influence medical diagnostics, evaluating its effectiveness in diagnosing osteoporosis is crucial [11].

This systematic review and meta-analysis aim to assess the diagnostic accuracy of DL models in predicting osteoporosis from OPG radiographs. As the first comprehensive review of its kind, it consolidates current evidence on the performance of these models and critically examines their clinical applications. By analyzing the strengths and limitations of DL techniques, this study provides valuable insights for future research focused on high-risk populations, such as postmenopausal women undergoing corticosteroid therapy. We hypothesize that various deep learning models do not significantly differ in accuracy for detecting osteoporosis from panoramic radiographs. Specifically, the null hypothesis (H0) states that these models have comparable accuracy and precision in detecting osteoporosis, allowing us to assess whether deep learning models provide a measurable improvement over traditional manual methods in osteoporosis detection.

Methods

Data resources and search strategy

This study followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [12] (Fig. 1) and was registered in PROSPERO (CRD42024583964). A literature search was conducted in electronic databases, including MEDLINE (PubMed), Scopus, and Web of Science (WOS), to identify relevant articles up to February 10, 2025. The search used the following Medical Subject Headings (MeSH) and keywords: ("deep learning" or "transfer learning" or "convolutional neural network" or "CNN" or " artificial neural networks" or "artificial intelligence" or "automatic learning") and ("osteoporosis" or "bone mineral density" or "BMD" or "bone loss" or "bone quality" or "bone micro-architecture") and ("orthopantomography" or "panoramic radiograph" or "OPG"). The detailed search strategy is provided in Supplementary File 1. All records were managed in Endnote 20, with duplicates removed.

Inclusion and exclusion criteria

Inclusion and exclusion criteria for eligible studies were established using the PICOS (Population, Intervention, Comparison, Outcome, Study Design) framework (Table 1).

Study selection

Two reviewers (M.A. and A.T.) conducted the screening process independently and in duplicate. Each reviewer screened titles and abstracts to identify studies meeting the inclusion criteria, followed by an independent review of full-text articles to confirm eligibility. Discrepancies were discussed and resolved through consensus, and if

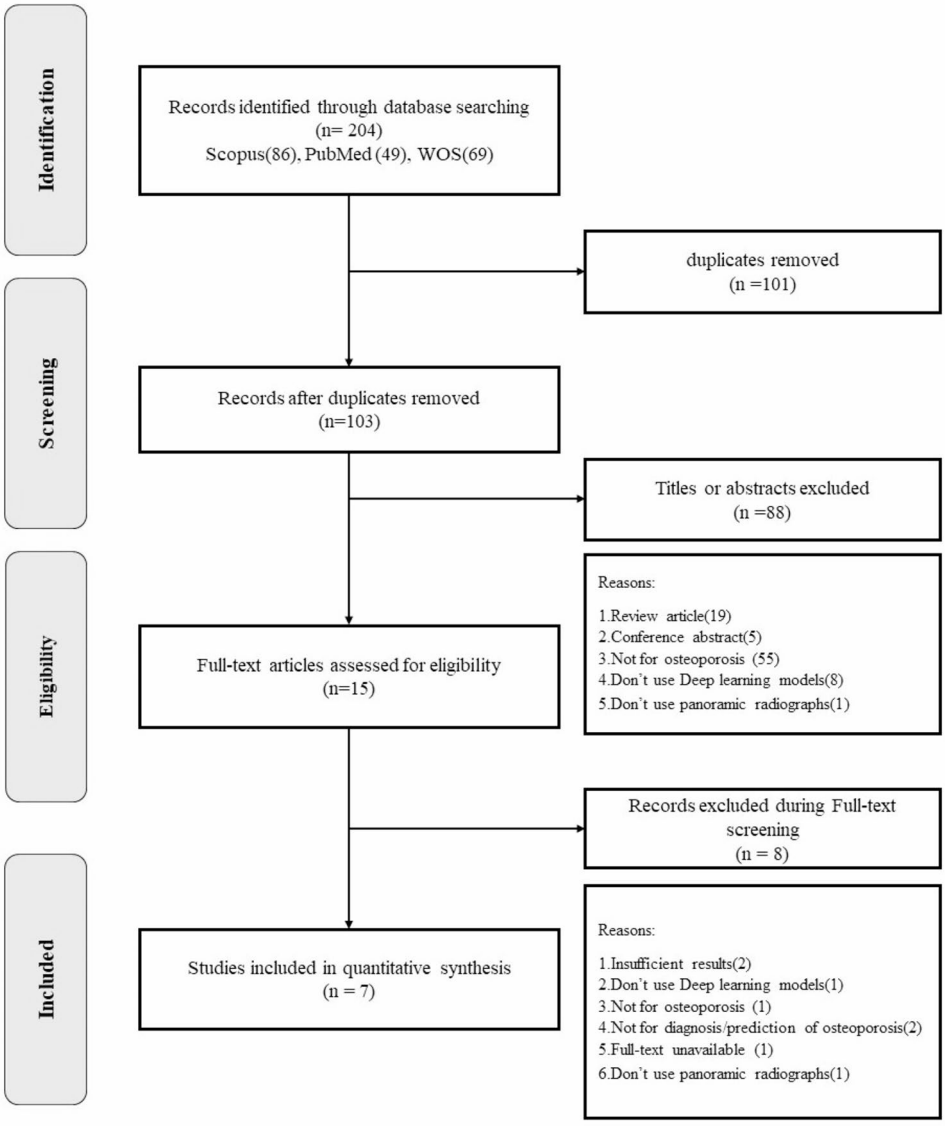

Fig. 1 Preferred reporting items PRISMA flow diagram for systematic reviews and meta-analyses

consensus could not be reached, a third reviewer (M.H.) made the final decision. We documented the articles screened and the reasons for exclusion at each stage.

Data extraction

Two investigators (M.A. and A.T.) independently extracted data from selected studies, including the first author's name, publication year, study design, country, sample size, gender, imaging modality, reference test, DL model, and performance metrics (AUC, sensitivity, specificity, and accuracy). If multiple DL models were developed in one study, data from each model were gathered. For each study, we extracted or calculated the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). When reported directly, we recorded these values as is. If not explicitly provided, we calculated them from the reported sensitivity, specificity, and total number of cases (osteoporosis and non-osteoporosis) using following equations:

- $TP = Sensitivity \times (TP + FN)$.

where $TP + FN$ represents the total number of diseased cases.

- $TN = Specificity \times (TN + FP)$.

where $TN + FP$ represents the total number of non-diseased cases.

- Total number of cases $(N) = TP + FN + TN + FP$.

Table 1 Inclusion and exclusion criteria based on the PICOS framework

	Inclusion criteria	Exclusion criteria
Population (P)	Studies assessing patients with suspected or diagnosed osteoporosis through panoramic radiographs.	Studies of patients with suspected or diagnosed osteoporosis based on images of the hip, femur, or lumbar spine. Studies fewer than 10 participants.
Intervention (I)	Research on deep learning models for osteoporosis diagnosis	Research utilizing non-deep learning techniques (e.g., traditional machine learning and statistical methods)
Comparator (C)	Studies evaluating deep learning models against traditional diagnostic methods, including dual-energy X-ray absorptiometry (DXA), mandibular cortical index (MCI), and expert radiologist assessments.	
Outcome (O)	● Studies reporting diagnostic accuracy metrics, including sensitivity, specificity, and area under the curve (AUC). ● Studies that provide data for calculating these metrics.	Studies that do not report diagnostic accuracy or lack sufficient data for metric calculation.
Study Design (S)	● Original peer-reviewed studies utilizing DL algorithms for osteoporosis prediction or diagnosis. ● published in English language	Review, editorials, commentaries, letters to the editor, case series, case reports, conference abstract and preprint articles

This data was recorded in an Excel spreadsheet. Discrepancies in extraction were resolved through discussion, and a third reviewer (M.H.) was consulted when needed. The data extraction form is provided in Supplemental File 1, Table S2.

Assessment of risk of bias

Two raters (M.A. and A.T.) independently evaluated potential bias risks in selected studies using the QUADAS-2 tool, which comprises four key domains: (1) patient selection, (2) index test, (3) reference standard, and (4) flow and timing [13]. Disagreements were resolved with input from the third author.

Statistical analysis

The included studies employed various metrics, including sensitivity, specificity, and AUC, to evaluate the diagnostic performance of DL models. Sensitivity, specificity, and AUC summaries were calculated using bivariate random-effects models for diagnostic meta-analysis. Forest plots illustrated pooled sensitivity and specificity with a 95% confidence interval (95% CI). To quantify the likelihood of positive or harmful tests, positive/negative likelihood ratios (LR+/LR-) were computed, leading to the calculation of the diagnostic odds ratio (DOR) and its 95% CI, along with a corresponding forest plot. The summary receiver-operating characteristic curve (SROC) was plotted using the bivariate method, and the AUC was calculated. Heterogeneity among studies was assessed using the inconsistency index (I2); an $I^2 > 50\%$ indicated significant heterogeneity, prompting the application of a random effects model.

To explore potential sources of heterogeneity, we conducted sensitivity analyses, meta-regression, and subgroup analyses based on the DL methods (such as AlexNet, VGG, and ResNet). Subgroup analyses also compared the diagnostic accuracy of DL models within each group. Deeks' funnel plot asymmetry test assessed publication bias with a statistically significant P-value of less than 0.05. All statistical analyses were conducted using the Midas and mandi packages in STATA version 17 (StataCorp LP, College Station, TX) and Meta-Disc.

Results

Study selection

Figure 1 outlines our study selection process. We initially identified 204 articles through electronic database searches. After excluding 189 duplicates and unrelated articles, 15 studies were assessed for eligibility. Ultimately, seven articles were selected after a full-text review.

Study characteristics

The selected articles, published between 2019 and 2025, included studies conducted in Japan ($n = 2$), South Korea ($n = 2$), Turkey ($n = 1$), Brazil ($n = 1$), and Germany ($n = 1$). A total of 4217 participants were involved, with 85% female, and the average age of participants exceeded 50 years. In four studies, MCI served as the reference standard for confirming patients with osteoporosis [14–17], while three studies utilized DEXA [11, 18, 19]. The reviewed studies developed 21 DL models with various architectures, often utilizing transfer learning methods such as ResNet, VGG, and EfficientNet. Most studies implemented multiple DL models. Three studies employed k-fold cross-validation [16–18], while the others used simple random sampling for dataset partitioning into training, validation, and testing sets [11, 14, 15, 19]. All studies utilized an internal dataset for validation testing. Regarding hyperparameters, there was notable variability across studies. Optimization algorithms that reported in three studies, including Stochastic Gradient Descent (SGD) [16, 18, 19] and Adam [14]. Reported learning rates ranged from 0.001 [17, 19] to 0.1 [16], with one study using multiple values (0.001 and 0.01) [18]. four studies reported batch sizes, which varied between 16 [19], 24 [14], 32 [16], and 50 [15]. The number of training epochs, reported in four studies, included 20 [15], 30

[16], 100 [17, 18], and 150 [19]. Table 2 summarizes the information from the selected studies.

Risk of bias within studies

Using the QUADAS-2 method, we assessed the quality of the studies and potential bias. Most studies showed a low risk of bias in the "patient selection" and "reference standard" categories, though one was ambiguous. All studies were generally low risk in the "patient selection," "index test," and "reference standard" domains. Four studies had unclear assessments in the "flow and timing" domain. Detailed quality assessment results can be found in Supplemental File 1, Fig. S1, and Table S3.

Diagnostic accuracy of deep learning models

The DL models developed in the included studies exhibited AUC values from 66.8% [11] to 99.8% [14]. Their sensitivity and specificity ranged from 59% [18] to 97% [14] and 64.9% [11] to 100% [14], respectively. Detailed results for these DL models can be found in Supplemental File 1, Table S4. The pooled sensitivity and specificity were 0.80 (95% CI: 0.74–0.86) and 0.92 (95% CI: 0.88–0.95). Figures 2 and 3 present forest plots with a 95% CI for sensitivities and specificities. Both sensitivity (I^2 = 94%, $p < 0.001$) and specificity (I^2 = 97%, $p < 0.001$) showed significant heterogeneity. The bivariate approach yielded a pooled SROC curve with an AUC of 0.93 (95% CI: 0.91–0.95) (Fig. 4), and the DOR for the DL models was 50.42 (95% CI: 23.31–109.05). Additional details on the meta-analysis of diagnostic accuracy are in Supplemental File 2, Figs. S1-S5.

Meta-regression analysis and subgroup analysis

The meta-regression analysis aimed to identify sources of heterogeneity among studies. As illustrated in Fig. 5, validation methods and type of reference standard significantly contributed to this heterogeneity. Additionally, Subgroup analyses were conducted based on DL methods (Table 3). The results have shown that no significant differences in diagnostic accuracy were found between subgroups, and in terms of DL methods, AlexNet, with a sensitivity of 0.89 and specificity of 0.99, had better performance than another method.

Sensitivity analysis and publication bias

Sensitivity analyses were performed to identify sources of heterogeneity. The bivariate box plot (Supplementary file 21, Fig. S3) indicated two outlier models, which were subsequently excluded from the meta-analysis. The sensitivity analysis showed that excluding these outliers had no significant impact (Supplementary File 2, Table S1). Additionally, Deeks' funnel plot asymmetry test was conducted to assess potential publication bias (Fig. 6), revealing no significant bias among the studies ($P = 0.54$).

Discussion

Diagnostic imaging results are assessed to confirm or exclude diseases in patients at clinics and para-clinic services. Radiological tests measure accuracy through sensitivity and specificity in relation to gold standard methods, which often exhibit an inverse relationship. The area under the receiver operating characteristic (ROC) curve indicates combined efficacy and serves as

Table 2 This review includes a comprehensive summary of the data collected from the studies

Study/ Year	Country	Study Design	Number of patients	Mean age	Center	Reference test	validation techniques	DL Methods	External validation test
Nakamoto et al./ 2022 [15]	Japan	Retrospective	100	> 50	single-center	MCI	Random sampling	AlexNet, VGG16, GoogleNet	NO
Lee et al./2020[11]	South Korea	Retrospective	680	> 50	single-center	DXA	Random sampling	CNN, VGG-16,	NO
Sukegawa et al./2022[18]	Japan	Retrospective	778	> 50	single-center	DXA	5-fold cross validation	EfficientNet b3, ResNet50, ResNet152, ResNet18, EfficientNet b0, EfficientNet b7	NO
Lee et al./2019[14]	South Korea	Retrospective	200	> 50	single-center	MCI	Random sampling	AlexNet	NO
Tassoker et al./2022[16]	Turkey	Retrospective	1488	> 50	single-center	MCI	5-fold cross validation	AlexNet, GoogleNet, ResNet50, ShuffleNet, SqueezeNet	NO
Dias et al./2025[17]	Brazil	Retrospective	471	> 50	single-center	MCI	5-fold cross validation	EfficientNet b5 EfficientNet b6 EfficientNet b7	NO
Gaudin et al./2024[19]	Germany	Retrospective	500	> 50	single-center	DXA	Random sampling	Densenet201	NO

Mean age: mean age of participants. Center: data gathering centers (Single center/Multicenter). DL method: deep learning method. MCI: mandibular cortex index. DXA: dual-energy X-ray absorptiometry. DL Methods: Deep learning methods. NA: Not Available

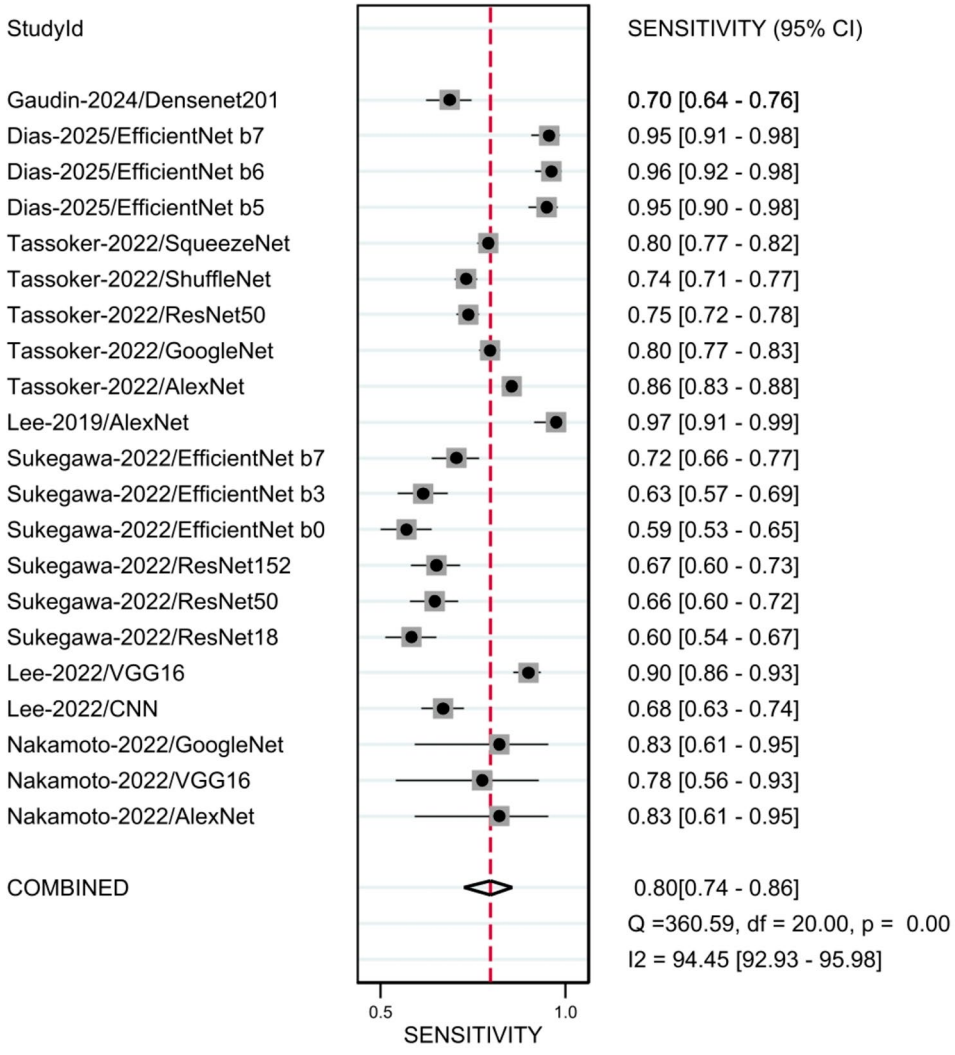

Fig. 2 Forest plot showing the pooled sensitivity of Deep Learning for predicting osteoporosis

a key accuracy metric. These scores are vital for analyzing imaging in both quantitative and qualitative contexts. As a result, researchers are increasingly exploring the sensitivity and specificity of various medical imaging techniques for diagnosing osteoporosis in high-risk populations [20, 21]. This area has seen considerable research activity. While each imaging method has pros and cons, integrating DL has emerged as a promising solution to address these limitations and enhance diagnostic efforts [22]. Some studies emphasize the diagnosis of dental issues, the distinction between primary and secondary tumors, and the poor prognosis associated with distant metastases to the mandible, along with the importance of timely treatment and management strategies. One study specifically demonstrated the correlation between clinical findings and sensitivity to healthy and diseased dental conditions, such as caries and periapical lesions, using an artificial intelligence program. The role of medical imaging and artificial intelligence in recognizing dental diseases has been underscored in these studies [23, 24].

This systematic review and meta-analysis assessed the diagnostic accuracy of DL models in predicting osteoporosis. The results suggest that DL models are valuable tools for aiding radiologists and physicians in the early, non-invasive diagnosis of osteoporosis. This is crucial for improving prognosis, enabling effective treatment, and increasing survival rates.

Additionally, DL algorithms can enhance osteoporosis screening by analyzing panoramic images without disrupting clinical workflows and demonstrate strong performance in managing severe osteoporotic fractures. Yen et al. recently conducted a meta-analysis on DL model performance in diagnosing osteoporosis, reporting high diagnostic accuracy [25]. However, that study had

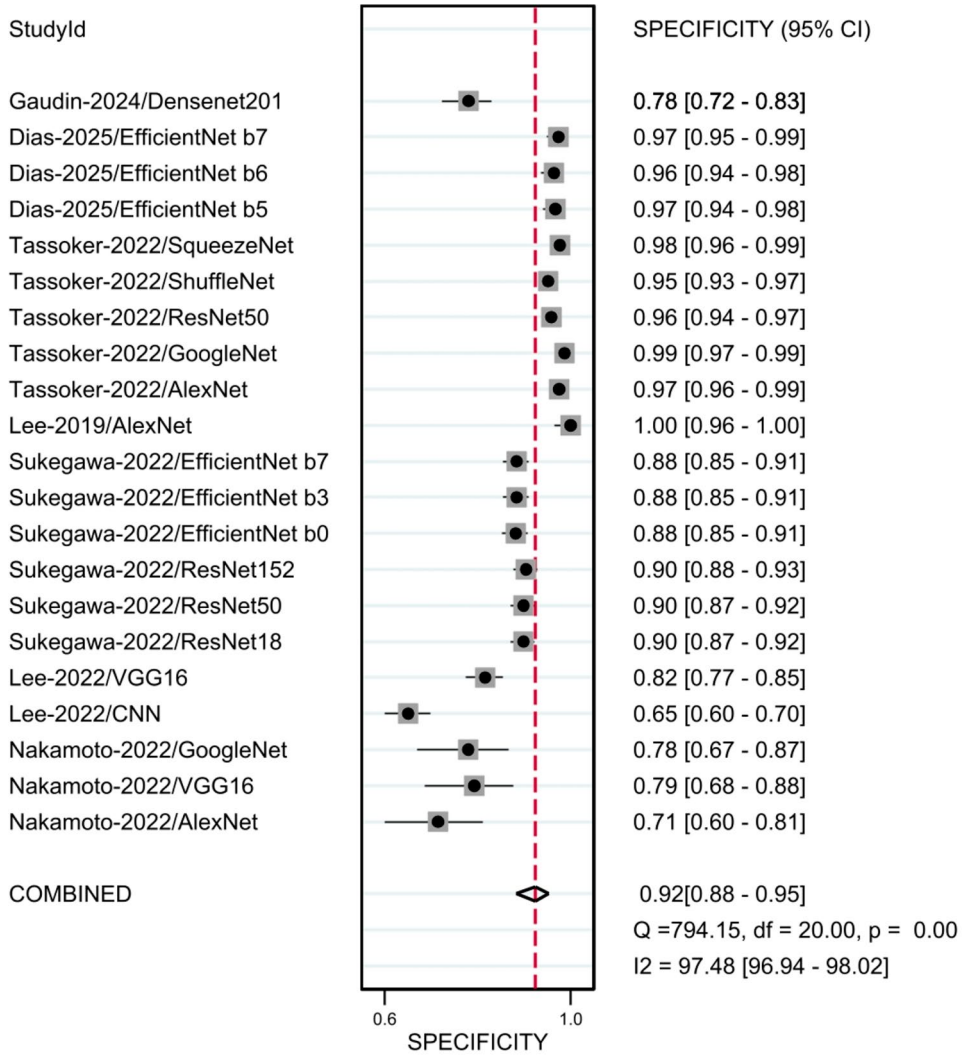

Fig. 3 Forest plot illustrating the pooled specificity of Deep Learning in predictions

limitations, including the absence of specific analyses like meta-regression, subgroup analysis, and publication bias assessment. It focused on pelvic and waist images while providing minimal attention to OPG images and did not explore different DL techniques. Furthermore, it lacked results such as DOR and LR. Our study comprehensively addresses these gaps by evaluating and comparing these parameters without the mentioned limitations.

The studies reviewed demonstrated that the effectiveness of a DL model is assessed using high-accuracy metrics such as AUC, sensitivity, and specificity, which effectively differentiate between patients and healthy individuals. The combined metrics were AUC 0.93 (95% CI: 0.91–0.95), sensitivity 0.80 (95% CI: 0.74–0.86), and specificity 0.92 (95% CI: 0.88–0.95). Similarly, Yen et al. reported an AUROC of 0.88, a sensitivity of 0.81, and a specificity of 0.87 [25]. Additionally, our results show that DL models outperform other machine learning methods in osteoporosis prediction, aligning with Rahim et al.'s study [26]. However, this research area is still emerging, and further studies are necessary to validate the generalizability of these results and enhance DL model performance for clinical applications.

This meta-analysis found a pooled diagnostic odds ratio (DOR) of 50.42 (95% CI: 23–109), indicating that DL is generally superior for diagnosing osteoporosis compared to traditional machine learning techniques. Likelihood ratios (LR) are crucial metrics that reflect disease frequency and enhance clinical judgment [27]. The study reported a pooled positive likelihood ratio (LR+) of 10.67 (range 6.4–17.6), suggesting that predictions of osteoporosis using DL are 10.67 times more likely to be correct than pessimistic predictions, demonstrating a substantial positive predictive value for identifying actual cases of osteoporosis. Additionally, a pooled negative likelihood ratio (LR–) of 0.21 (range 0.15–0.29) was observed, indicating effective identification of individuals without osteoporosis. Rahim et al. reported LR + and LR – rates of

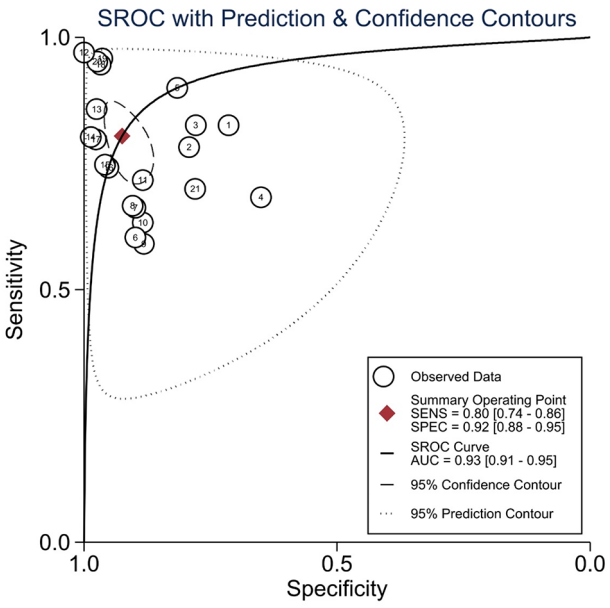

Fig. 4 Summary receiver-operating curves (SROC) using the bivariate approach

3.7 and 0.22 for ML models predicting osteoporosis [26]. Thus, our results suggest that DL outperforms other ML algorithms in this context.

This study provides the first comprehensive evaluation of the diagnostic accuracy of DL models for predicting osteoporosis from panoramic radiographs, serving as a valuable reference for future research. Despite significant challenges posed by heterogeneity in our meta-analysis, we identified and addressed its sources to enhance the robustness of our results. The analysis revealed considerable heterogeneity, indicated by high I2 values. To explore its sources, we conducted meta-regression, finding that variations in characteristics such as validation methods may influence prediction performance across studies. Additionally, we performed a subgroup analysis based on DL techniques, revealing that most studies utilized transfer learning models like AlexNet and ResNet, which are notably effective for osteoporosis diagnosis [28]. However, factors like dataset size, image quality, and training parameters significantly impact performance. Future research should standardize methods and reporting practices to mitigate heterogeneity. We evaluated publication bias in our analysis and found no significant bias, as confirmed by the Deek's funnel plot test, which strengthens the reliability of our findings. Nonetheless, Caution is necessary in interpreting these results due to possible unrecognized biases. External validation of DL models is essential for clinical reliability [29, 30], but all reviewed studies depended exclusively on internal validation. Therefore, implementing external validation is essential for accurately assessing DL model performance. In addition, the few numbers of studies and single-center data were the limitation of our study; therefore, multicenter data and further research is needed to broaden the evidence. A key limitation of this study is the incomplete and inconsistent reporting of hyperparameters in several included studies. While we diligently analyzed the available data concerning critical parameters such as optimization algorithms, learning rates, batch sizes, and epoch, it became evident that many of the studies either neglected to report these essential details entirely or provided information that was deemed inadequate for comprehensive analysis. This lack of transparency poses significant challenges, as it hinders our ability to conduct a thorough evaluation regarding how these various hyperparameters influence model performance. Furthermore, this reporting issue contributes to the considerable heterogeneity observed in our meta-analysis, making it difficult to draw reliable conclusions across the different studies reviewed. To improve the comparability and reproducibility of deep learning models used in osteoporosis prediction, it is imperative that future research endeavors adhere to accepted standardized reporting guidelines for hyperparameters. By following such guidelines, researchers can provide a clearer picture of their methodologies and findings. Drawing from the insights gained and existing limitations highlighted in this study, it is anticipated that future research efforts will increasingly focus on standardizing documentation practices. This standardization is crucial for enabling more rigorous and meaningful evaluations of deep learning model performance. In turn, such improvements are expected to enhance the accuracy and precision of quantitative analyses within the field. Additionally, it would be beneficial for future studies to delve deeper into exploring the effects of specific hyperparameter settings through more structure-oriented subgroup or sensitivity analyses. This deeper exploration could lead to a better understanding of how varying these parameters impacts diagnostic accuracy, ultimately providing clearer insights into optimizing deep learning approaches in osteoporosis prediction and related applications.

While DL shows strong potential in predicting osteoporosis, additional research is necessary to validate these findings through prospective clinical trials. Further efforts should focus on developing and optimizing DL models for clinical integration. Lastly, strategies must be established to address the ethical and societal implications of using DL in osteoporosis prediction.

Conclusions

This review and meta-analysis revealed that DL models for osteoporosis diagnosis achieved pooled sensitivity, specificity, and AUC of 80%, 92%, and 93%, respectively, outperforming other algorithms. Transfer learning

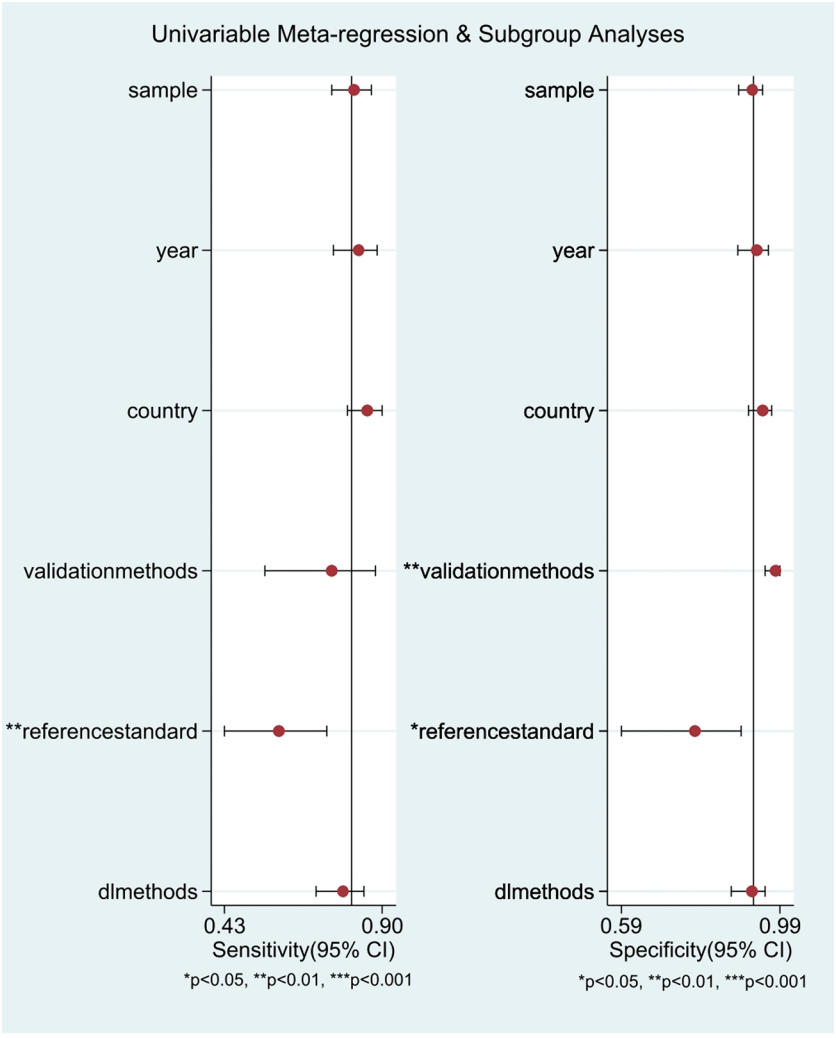

Fig. 5 Meta-regression and subgroup analysis

Table 3 The results of subgroup analyses

Subgroup	DL models *N* (%)	Sen (95%CI)	Spe (95%CI)	PLR (95%CI)	NLR (95%CI)	DOR (95%CI)
Overall	21 (100)	0.80 (0.74–0.86)	0.92 (0.88–0.95)	10.67 (6.4–17.6)	0.21 (0.15–0.29)	50.42 (23–109)
DL Methods						
AlexNet	3(14)	0.89 (0.73–0.96)	0.99 (0.49-1)	94 (-353-541)	0.11 (0.04–0.22)	822 (-3852-5497)
VGG	2(9.5)	0.90 (0.85–0.92)	0.81 (0.77–0.84)	4.73 (3.81–5.65)	0.13 (0.09–0.18)	35 (20–50)
ResNet	4(19)	0.67 (0.62–0.72)	0.92 (0.88–0.94)	8.27 (4.81–11.72)	0.36 (0.29–0.42)	23 (9–37)
EfficientNet	6(29)	0.86 (0.70–0.94)	0.94 (0.89–0.96)	13.57 (4.52–22.6)	0.15 (0.02–0.28)	89 (-47-227)
GoogleNet	2(9.5)	0.8 (0.78–0.83)	0.94 (0.66–0.99)	14 (-14-42.21)	0.2 (0.17–0.25)	67 (-76-210)
Other	4(19)	0.73 (0.68–0.79)	0.90 (0.71–0.97)	7.05 (-1.34-15.46)	0.3 (0.21–0.39)	22 (-11-58)

Sen: sensitivity, Spe: specificity, PLR: positive likelihood ratios, NLR: negative likelihood ratios, DOR: diagnostic odds ratio, DL Methods: deep learning methods

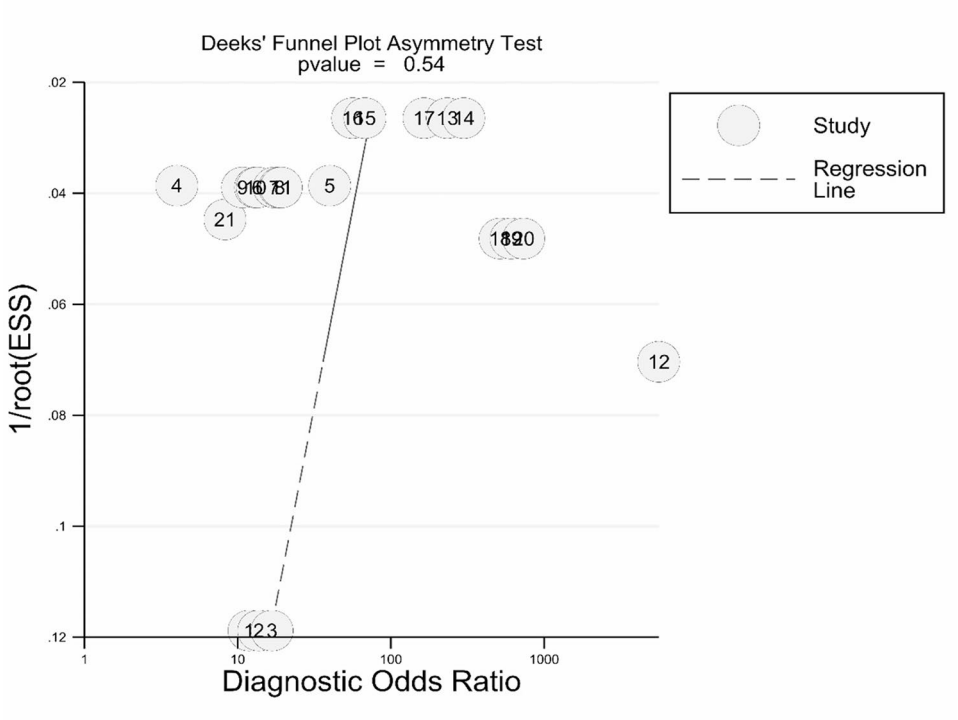

Fig. 6 Deeks' funnel plot asymmetry test for publication

models like VGG and ResNet showed enhanced performance, suggesting their effectiveness. Our findings indicate that DL models can facilitate early detection with high sensitivity and specificity, functioning as promising non-invasive diagnostic tools. However, further research, including more extensive multi-center studies, must refine these algorithms and validate their effectiveness in at-risk populations.

Abbreviations

BMD	Measure bone mineral density
DEXA	Dual-energy X-ray absorptiometry
MCI	Mandibular cortex index
OPG	Orthopantomogram
DL	Deep learning
AI	Artificial intelligence
CNN	Convolutional neural networks
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
TP	True positives
FP	False positives
TN	True negatives
FN	False negatives
AUC	Area under the curve
SGD	Stochastic Gradient Descent
DOR	Diagnostic Odds Ratio
SROC	Summary receiver-operating characteristic curve

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12880-025-01626-z.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Acknowledgements
Not applicable.

Author contributions
MA: Conceptualization; writing – original draft; formal analysis. AT: Conceptualization; formal analysis; writing – review and editing. MH: Conceptualization; formal analysis; writing – review and editing. AM: Conceptualization, writing – review and editing. N.D.N: Conceptualization, review, and editing.

Funding
No funding was received to assist with preparing this manuscript.

Data availability
All data generated or analysed during this study are included in this published article and its supplementary information files.

Declarations

Ethics approval and consent to participate
This study was approved by the ethics committee of Ardabil University of Medical Sciences (approval code: IR.ARUMS.MEDICINE.REC.1403.179).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 27 November 2024 / Accepted: 4 March 2025
Published online: 12 March 2025

References

1. Shevroja E, Cafarelli FP, Guglielmi G, Hans D. DXA parameters, trabecular bone score (TBS) and bone mineral density (BMD), in fracture risk prediction in endocrine-mediated secondary osteoporosis. Endocrine. 2021;74:20–8.
2. Zanker J, Duque G. Osteoporosis in older persons: old and new players. J Am Geriatr Soc. 2019;67(4):831–40.
3. Jafri L, Majid H, Farooqui AJ, Ahmed S, Effendi MUN, Zaman M-u, et al. Developing and piloting an online course on osteoporosis using a multidisciplinary multi-institute approach-a cross-sectional qualitative study. PLoS ONE. 2024;19(2):e0291617.
4. Sabri SA, Chavarria JC, Ackert-Bicknell C, Swanson C, Burger E. Osteoporosis: an update on screening, diagnosis, evaluation, and treatment. Orthopedics. 2023;46(1):e20–6.
5. Banh K. Essentials of Osteoporosis: Early Prevention, Screening, and Management of this Silent Disease. 2022.
6. LeBoff MS, Greenspan S, Insogna K, Lewiecki E, Saag K, Singer A, et al. The clinician's guide to prevention and treatment of osteoporosis. Osteoporos Int. 2022;33(10):2049–102.
7. Snodgrass P, Zou A, Gruntmanis U, Gitajn IL. Osteoporosis diagnosis, management, and referral practice after fragility fractures. Curr Osteoporos Rep. 2022;20(3):163–9.
8. Aibar-Almazán A, Voltes-Martínez A, Castellote-Caballero Y, Afanador-Restrepo DF. Carcelén-Fraile MdC, López-Ruiz E. Current status of the diagnosis and management of osteoporosis. Int J Mol Sci. 2022;23(16):9465.
9. Pallagatti S, Parnami P, Sheikh S, Gupta D, Suppl-1 M. Efficacy of panoramic radiography in the detection of osteoporosis in Post-Menopausal women when compared to dual energy X-Ray absorptiometry. Open Dentistry J. 2017;11:350.
10. Dhanya M, Kumar J, Ramalingam K. Effectiveness of orthopantomograms as a screening tool for osteoporosis: A Case-Control study. Cureus. 2023;15(9).
11. 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).
12. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg. 2010;8(5):336–41.
13. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155:529–36. https://doi.org/10.7326/0003-4819-155-8-201110180-00009
14. 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.
15. Nakamoto T, Taguchi A, Kakimoto N. Osteoporosis screening support system from panoramic radiographs using deep learning by convolutional neural network. Dentomaxillofac Radiol. 2022;51(6):20220135.
16. Tassoker M, Öziç MÜ, Yuce F. Comparison of five convolutional neural networks for predicting osteoporosis based on mandibular cortical index on panoramic radiographs. Dentomaxillofacial Radiol. 2022;51(6).
17. Dias BSS, Querrer R, Figueiredo PT, Leite AF, de Melo NS, Costa LR et al. Osteoporosis screening: leveraging EfficientNet with complete and cropped facial panoramic radiography imaging. Biomed Signal Process Control. 2025;100.
18. Sukegawa S, Fujimura A, Taguchi A, Yamamoto N, Kitamura A, Goto R, et al. Identification of osteoporosis using ensemble deep learning model with panoramic radiographs and clinical covariates. Sci Rep. 2022;12(1):6088.
19. Gaudin R, Otto W, Ghanad I, Kewenig S, Rendenbach C, Alevizakos V et al. Enhanced osteoporosis detection using artificial intelligence: A deep learning approach to panoramic radiographs with an emphasis on the mental foramen. Med Sci. 2024;12(3). https://doi.org/10.3390/medsci12030049
20. Adams JE. Advances in bone imaging for osteoporosis. Nat Rev Endocrinol. 2013;9(1):28–42.
21. de Oliveira MA, Moraes R, Castanha EB, Prevedello AS, Vieira Filho J, Bussolaro FA, et al. Osteoporosis screening: applied methods and technological trends. Med Eng Phys. 2022;108:103887.
22. Tsai D-J, Lin C, Lin C-S, Lee C-C, Wang C-H, Fang W-H. Artificial Intelligence-enabled chest X-ray classifies osteoporosis and identifies mortality risk. J Med Syst. 2024;48(1):12.
23. Ünsal G, Erturk AF, Orhan K. Investigation of the internal structure and radiological characteristics of distant metastases to the jaws: A retrospective study. Eurasian Dent Res. 2024;2(3):70–2.
24. Orhan K, Aktuna Belgin C, Manulis D, Golitsyna M, Bayrak S, Aksoy S, et al. Determining the reliability of diagnosis and treatment using artificial intelligence software with panoramic radiographs. Imaging Sci Dent. 2023;53(3):199–208.
25. Yen T-Y, Ho C-S, Chen Y-P, Pei Y-C. Diagnostic accuracy of deep learning for the prediction of osteoporosis using plain X-rays: A systematic review and Meta-Analysis. Diagnostics. 2024;14(2):207.
26. Rahim F, Zaki Zadeh A, Javanmardi P, Emmanuel Komolafe T, Khalafi M, Arjomandi A, et al. Machine learning algorithms for diagnosis of hip bone osteoporosis: a systematic review and meta-analysis study. Biomed Eng Online. 2023;22(1):68.
27. Davis MP, Soni K, Strobel S. Likelihood ratios: an important concept for palliative physicians to understand. Am J Hospice Palliat Medicine®. 2023;40(8):894–9.
28. Wani IM, Arora S. Osteoporosis diagnosis in knee X-rays by transfer learning based on Convolution neural network. Multimedia Tools Appl. 2023;82(9):14193–217.
29. Bleeker S, Moll H, Steyerberg Ea, Donders A, Derksen-Lubsen G, Grobbee D, et al. External validation is necessary in prediction research:: A clinical example. J Clin Epidemiol. 2003;56(9):826–32.
30. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J. 2021;14(1):49–58.

Publisher's note