



Inteligencia artificial para el cribado de osteoporosis en radiografías panorámicas dentales: una revisión sistemática

Artificial intelligence for osteoporosis screening on dental panoramic radiographs: a systematic review

Neisy Pérez Ramos¹  

¹ Universidad de Ciencias Médicas de Villa Clara, Facultad de Medicina, Departamento de Ciencias Biomédicas. Santa Clara, Cuba

Submitted: 15-05-2025 | Revised: 19-07-2025 | Accepted: 08-09-2025 | Published: 09-09-2025

Corresponding Author: Neisy Pérez Ramos 

ABSTRACT

Introduction: Osteoporosis is underdiagnosed because dual-energy X-ray absorptiometry, the reference standard, is underused, while dental panoramic radiographs are taken routinely and show mandibular cortical signs of low bone mineral density. We synthesized the diagnostic accuracy of artificial-intelligence models that screen for osteoporosis on these radiographs. **Methods:** We searched PubMed, OpenAlex, and Crossref for studies published from 2013 to 2025, screened 321 records after deduplication, retrieved full texts in open access, and included studies that evaluated an artificial-intelligence model against a bone-mineral-density reference or mandibular index. Accuracy was extracted per study and risk of bias appraised. **Results:** Ten studies met the eligibility criteria. Nine used dual-energy X-ray absorptiometry as the reference standard and eight employed deep convolutional networks. Reported area-under-the-curve values ranged from 0.76 to 0.89, and reported accuracy ranged from about 74% to 98%, with sensitivity mostly between 78% and 91% where a bone-density reference was used. Risk of bias was high or unclear in most studies, and only one study reported external validation. **Conclusions:** Artificial-intelligence models can detect osteoporosis on dental panoramic radiographs with promising accuracy, but the evidence is limited by small, single-centre datasets, heterogeneous methods, and a high risk of bias, so translation into opportunistic screening remains premature.

Keywords: osteoporosis; artificial intelligence; deep learning; panoramic radiography; bone density; mass screening.

RESUMEN

Introducción: La osteoporosis está infradiagnosticada porque la absorciometría de rayos X de energía dual, el patrón de referencia, se utiliza poco, mientras que las radiografías panorámicas dentales se obtienen de forma rutinaria y muestran signos corticales mandibulares de baja densidad mineral ósea. Sintetizamos la precisión diagnóstica de los modelos de inteligencia artificial que criban osteoporosis en estas radiografías. **Métodos:** Buscamos en PubMed, OpenAlex y Crossref estudios publicados de 2013 a 2025, cribamos 321 registros tras eliminar duplicados, recuperamos textos completos en acceso abierto e incluimos estudios que evaluaran un modelo de inteligencia artificial frente a una referencia de densidad mineral ósea o un índice mandibular. Se extrajo la precisión por estudio y se evaluó el riesgo de sesgo. **Resultados:** Diez estudios cumplieron los criterios de elegibilidad. Nueve usaron absorciometría de rayos X de energía dual como referencia y ocho emplearon redes convolucionales profundas. El área bajo la curva osciló entre 0,76 y 0,89, la precisión entre aproximadamente 74% y 98%, y la sensibilidad se situó principalmente entre 78% y 91% cuando se usó una referencia de densidad ósea. El riesgo de sesgo fue alto o incierto en la mayoría y solo un estudio comunicó validación externa. **Conclusiones:** Los modelos de inteligencia artificial pueden detectar osteoporosis en radiografías panorámicas dentales con una precisión prometedora, pero la evidencia está limitada por conjuntos pequeños y unicéntricos, métodos heterogéneos y un alto riesgo de sesgo, por lo que su traslación al cribado oportunista sigue siendo prematura.

Palabras clave: osteoporosis; inteligencia artificial; aprendizaje profundo; radiografía panorámica; densidad ósea; tamizaje masivo.

INTRODUCTION

Osteoporosis is a systemic skeletal disease defined by low bone mass and microarchitectural deterioration that raise the risk of fragility fractures, and its clinical importance lies less in the bone loss itself than in the fractures it makes possible. Osteoporosis is common and largely silent: a substantial share of postmenopausal women and older men have bone mineral density below the osteoporotic threshold, yet most are never identified until a fracture occurs, because the diagnosis depends on dual-energy X-ray absorptiometry (DXA), a test confined to specialized centers, underused in routine care, and often ordered only after the harm has already happened.(1,2)

The size of this diagnostic gap is concrete rather than rhetorical, because osteoporotic fractures arise in many adults who had no prior indication for bone-density testing and who therefore reached their first fracture without ever having been assessed.(3)

The mandible participates in the same skeletal process, and the appearance of the mandibular inferior cortex on a dental panoramic radiograph, including cortical thinning, porosity, and erosion, has long been recognized as a marker of low bone mineral density. Because panoramic radiographs are among the most frequently acquired dental images and are taken for reasons unrelated to bone health, they offer an attractive substrate for opportunistic screening, a strategy in which information already present in the image is used to flag patients who should be referred for formal densitometry.(4,5)

The clinical appeal is direct: a dentist could identify an asymptomatic woman at risk of a hip or vertebral fracture during a routine visit, without any additional radiation, cost, or appointment.

Conventional approaches to reading these radiographs rely on radiomorphometric indices, most prominently the mandibular cortical width and the mandibular cortical index, in which an observer grades the mandibular inferior cortex as normal, mildly eroded, or severely eroded.(6,7) These indices correlate with bone mineral density and identify women at higher fracture risk, but their clinical adoption has been limited by the effort and subjectivity of manual measurement and by a sensitivity that is too low for a screening instrument.(8,9)

Early work established that the mandibular cortex differs measurably between women with and without low bone density, and later studies confirmed the correlation across populations and defined how accurately these indices predict densitometric findings, generally with modest sensitivity and stronger specificity.(10-13)

The panoramic mandibular index and related measures have been proposed repeatedly as screening markers, and their diagnostic performance has been summarized with a consistent conclusion.(14-16) Their value is also demonstrably population-dependent, because genetic and epidemiological determinants shape cortical morphology independently of bone density, which complicates any single universal threshold.(17,18) This is precisely the kind of repetitive, pattern-based visual task in which artificial intelligence, and deep convolutional networks in particular, has been proposed to outperform or standardize human reading, and a growing number of studies have applied these models to panoramic radiographs for osteoporosis detection.(19,20) Earlier systematic reviews have begun to summarize this literature and have generally reported favorable accuracy, but they differ in scope, some focus on plain radiographs of other skeletal sites, and the most recent reviews are themselves heterogeneous in the studies they include.(21-23)

Broader reviews of artificial intelligence in osteoporosis management and in radiological diagnosis have likewise surveyed this work and reached compatible conclusions about both its promise and its methodological weaknesses.(24-26) What remains unsettled is not whether the models can detect osteoporosis, but how accurate they are under rigorous evaluation, which inputs and architectures perform best, whether they improve on the conventional mandibular indices and on human observers, and how much confidence the current evidence can support.

The objective of this review was to synthesize the diagnostic accuracy of artificial-intelligence models that screen for osteoporosis on dental panoramic radiographs and to characterize their inputs, their comparators, and the risk of bias of the evidence, in order to judge whether they are ready to support opportunistic screening in dental practice. To that end we posed four research questions. P1: What is the diagnostic accuracy, expressed as sensitivity, specificity, accuracy, and area under the receiver operating characteristic curve, of these models? P2: Which model architectures and image inputs are used, and how do end-to-end deep-learning approaches compare with methods built on hand-crafted radiomic or cortical features? P3: How do these models compare with conventional mandibular cortical indices and with human observers? P4: What is the risk of bias and the certainty of the evidence across the included studies?

METHODS

This review was conducted and reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020). The protocol was not registered prospectively in a review registry, and this deviation is stated explicitly. The eligibility criteria, the search, the extraction, and the risk-of-bias appraisal were each designed to answer one or more of the four research questions, and the correspondence is indicated as each is described.

Eligibility criteria (P1-P3)

We included studies that developed or evaluated an artificial-intelligence model applied to dental panoramic radiographs for the detection, classification, or screening of osteoporosis or low bone mineral density. The population comprised adults with a bone-mineral-density measurement or a mandibular cortical classification available as a label. The index test was any artificial-intelligence or machine-learning model, including deep convolutional networks, radiomics pipelines, and classical machine-learning classifiers, applied to panoramic images.

The reference standard was either DXA at the lumbar spine, hip, or femoral neck, or a radiologist-assigned mandibular cortical index where DXA was not available; this distinction is carried through the synthesis because a model trained to reproduce a radiologist's cortical grading is not answering the same question as a model trained against densitometry. The primary outcome was diagnostic accuracy. We excluded narrative and systematic reviews, commentaries, conference abstracts without retrievable full reports, and studies that used radiographs of sites other than the dental panoramic view, such as chest, knee, spine, hip, or ankle radiographs, or that measured cortical indices by hand without a machine-learning model. Studies published outside the 2013-2025 window or in a language other than English were also excluded.

Eligible designs were primary diagnostic accuracy studies and model-development studies that reported a quantitative performance measure against a reference standard; purely methodological papers without an evaluation, and studies that reported only a correlation between an automated measurement and densitometry without a classification result, were treated as ineligible because they do not answer the diagnostic question. The distinction between a model trained against DXA and a model trained against a radiologist's cortical grade was recorded for every included study, because the two tasks, although related, have different clinical meanings and different expected accuracy.

Information sources and search strategy (P1-P3)

PubMed, OpenAlex, and Crossref were searched from database inception up to 2025 for the nucleus of eligible studies. The search combined terms for artificial intelligence ("artificial intelligence", "deep learning", "machine learning", "convolutional neural network") with terms for the target condition ("osteoporosis", "bone mineral density", "bone density") and the imaging modality ("panoramic radiograph", "panoramic radiography", "orthopantomogram"). A representative search string is: ("artificial intelligence" OR "deep learning" OR "machine learning" OR "convolutional neural network") AND ("osteoporosis" OR "bone mineral density") AND ("panoramic radiograph" OR "panoramic radiography"). The full strings are reproducible from the record of searches; the complete strategy is available in the supplementary search log. A separate set of supporting searches retrieved prior systematic reviews and studies of conventional mandibular indices, used only for contextual comparison in the discussion and not for the synthesis of included studies. The last search was executed during the study period, and no date limit below 2013 was applied, so the window matches the 2013-2025 scope of the review.

Study selection (P1-P4)

Records retrieved from the three sources were pooled and deduplicated. Screening was performed in two stages. First, titles and abstracts were screened against the eligibility criteria. Second, the full text of potentially eligible records was retrieved where a legitimate open-access version was available, and assessed against the same criteria.

Data extraction (P1-P4)

For each included study we extracted the first author and year, country, study design, sample size and image source, population, the model architecture and input, the reference standard, and the principal accuracy result with its point estimate. Extraction was performed from the full-text article where available and from the abstract where full text was not openly accessible. No field

was completed without evidence, and where a study did not report a quantity it was left blank rather than imputed. When a study reported accuracy for several models, the principal result was taken to be the model the authors identified as best, and when a study reported several reference sites, the value for each site was retained separately. The extracted data were tabulated for the characteristics and accuracy tables.

Risk of bias (P4)

Risk of bias was appraised with the QUADAS-2 domains: patient selection, index test, reference standard, and flow and timing. Patient selection was judged high risk when the study used a case-control design with deliberately separated or non-consecutive groups; the index test was judged high or unclear when the model was evaluated on the same data used for development without an independent held-out set; the reference standard was judged low risk when DXA was applied to all participants; and flow and timing was judged on whether the interval between radiograph and densitometry was reported and reasonable.

Because name- and design-level reporting in this literature is frequently incomplete, the category “unclear” was used liberally, and a judgment of low risk was never assigned on the basis of missing information. Studies in low-tier or non-indexed journals, studies with small single-centre samples, and studies that selected cases and controls in a way that could inflate performance were rated as high risk where the design justified it.

Synthesis (P1-P3)

A meta-analysis was planned where studies reported a common metric against a common reference standard. In practice the studies were heterogeneous in reference standard (DXA versus mandibular cortical index), in outcome metric (area under the curve, sensitivity, specificity, or accuracy), and in population, and the number reporting any single metric against DXA was small. A pooled estimate was therefore judged uninformative and was not performed; instead, results were synthesized narratively by research question, with the range and distribution of reported accuracy reported alongside each study's individual estimate. Heterogeneity across studies is described qualitatively rather than quantified by a single statistic, and no summary receiver operating characteristic curve was estimated.

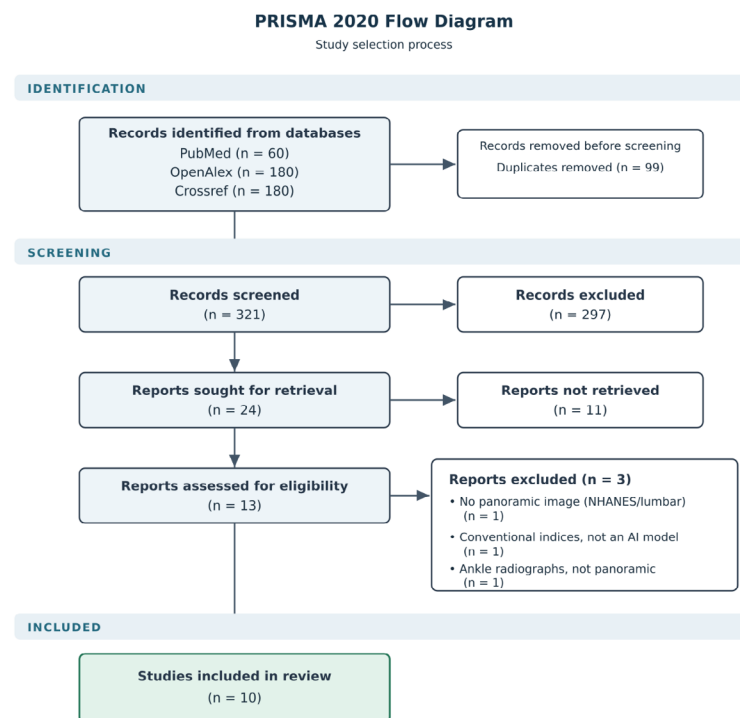
RESULTS

Study selection

The searches identified 420 records, of which 99 were duplicates, leaving 321 records for screening (Figure 1). After title and abstract screening, 297 records were excluded and 24 reports were sought for retrieval. Of these, 11 could not be retrieved because no legitimate open-access full text was available, and 13 reports were assessed for eligibility. At full-text assessment, 3 reports were excluded because the intervention was not an artificial-intelligence model on dental panoramic radiographs (a conventional mandibular-index study, an ankle-radiograph model, and a model trained on non-panoramic bone-density images), and 10 studies were included in the review. Most records excluded at title and abstract screening were reviews, editorials, and studies of conditions other than osteoporosis, or studies using radiographs of other skeletal sites, which reflects the breadth of the original search rather than a narrow first pass. The three full-text exclusions illustrate the boundary of the eligibility criteria, and the decisions were deliberately conservative, erring toward a smaller set of clearly relevant studies.

Characteristics of the included studies

The 10 included studies were published between 2013 and 2025 and originated from 9 countries, with two studies each from Japan and Austria and one each from Turkey, South Korea, Italy, Syria, Indonesia, and the United Kingdom (Table 1). Eight studies used an end-to-end deep convolutional network, and two used classical machine-learning classifiers built on extracted features: a multiclass support vector machine on cortical width and grayscale texture, and a random forest on co-occurrence and fractal texture features.(27,28) Nine studies used DXA as the reference standard, and one trained against a radiologist-assigned mandibular cortical index.(29)

Figure 1. PRISMA 2020 flow diagram of study selection.

Source: PRISMA 2020 flow diagram generated from the review's search and screening records.

The samples were predominantly female, and four studies enrolled only women of postmenopausal or older age, in line with the epidemiology of osteoporosis. Sample sizes ranged from 123 to 1500 images, and every study used retrospective, single-centre or convenience data. Only one study incorporated an external validation set from a second institution.(30)

The input region varied: some models analyzed the full panoramic image, while several cropped the mandibular inferior cortex or the mental foramen region, a choice that concentrates the model on the cortical signal but removes surrounding context. The studies span the full period of the review, with the earliest, from 2013, already using a machine-learning classifier on texture features, and the most recent, from 2025, testing radiomics and modern convolutional networks, so the literature reflects a decade of methodological change rather than a single moment.

Diagnostic accuracy of the models (P1)

The reported accuracy was consistently high but varied with the difficulty of the comparison. Among the six studies that reported an area under the curve against DXA, values ranged from 0.76 to 0.89, with the highest estimates coming from the random-forest texture classifier of Roberts et al., which reached an area under the curve of 0.872 for femoral-neck osteoporosis, and from the ensemble model of Sukegawa et al., which reached 0.893.(28,31)

Sensitivity and specificity were reported less uniformly. Where a DXA reference was used, sensitivity ranged from 78.3% to 90.5%, with Nakamoto et al. reporting 78.3%-82.6% for the lumbar spine and 80.0%-86.7% for the femoral neck, and Fanelli et al. reporting 90.5% for their best external-validation model.(30,32) Accuracy ranged from 74% to 98%, but the upper end is inflated by comparisons against a young, easily separable control group: Gaudin et al. achieved 98% accuracy against young controls but only an F1 of 0.74 and an area under the curve of 0.840 against age-matched controls, and the authors themselves emphasized that including young controls falsely inflates performance.(33)

The one study that trained against a radiologist's cortical grades rather than against densitometry is not directly comparable to the rest, because it answers a different question, namely whether a model can reproduce an expert's reading rather than whether it can detect low bone density.(29)The per-study estimates are tabulated in (Table 2).

Table 1. Characteristics of the included studies.

Study (year)	Country	Sample	Population	Model and input	Reference standard
Roberts et al. (2013)	United Kingdom	663 panoramic images	Female patients	RF on texture and cortical width	DXA femoral neck, hip, spine
Lee et al. (2020)	South Korea	680 images	Adults with DXA	VGG-16 transfer learning	DXA
Nakamoto et al. (2022)	Japan	1500 training + 100 test	Women ≥50 years	AlexNet, VGG-16, GoogLeNet	DXA lumbar spine, femoral neck
Tassoker et al. (2022)	Turkey	744 images	Women >50 years	Five CNNs on mandibular cortical ROI	MCI (radiologist)
Sukegawa et al. (2022)	Japan	778 images	Adults with DXA	EfficientNet/ResNet ensemble with covariates	DXA hip or spine
Astuti et al. (2022)	Indonesia	123 images	Postmenopausal women	SVM on cortical width and texture	DXA
Gaudin et al. (2024)	Austria	500 + 224 images	Adults with DXA	CNN on mental foramen ROI	DXA
Gaudin et al. (2024)	Austria	348 + 58 + 51 images	Adults with DXA	YOLOv8 detection + EfficientNet	DXA
Brangakgi et al. (2024)	Syria	400 images	Mostly women	MobileNet CNN	DXA hip and lumbar spine
Fanelli et al. (2025)	Italy	301 images (56 external)	Postmenopausal women	Radiomics and ResNet-50 CNN	DXA

Note: CNN = convolutional neural network; DXA = dual-energy X-ray absorptiometry; MCI = mandibular cortical index; ROI = region of interest; RF = random forest; SVM = support vector machine. Source: Data extracted from the included studies.

Model architectures and inputs (P2)

Table 2. Reported diagnostic accuracy of the included studies.

Study (year)	Metric	Reported value
Roberts et al. (2013)	AUC (femoral neck, width + texture)	0.872
Lee et al. (2020)	AUC	0.858
Nakamoto et al. (2022)	Se/Sp/Acc (lumbar spine)	78.3-82.6% / 71.4-79.2% / 74.0-79.0%
Tassoker et al. (2022)	Acc (three-class MCI)	81.14%
Sukegawa et al. (2022)	AUC / Acc (ensemble)	0.893 / 82.4%
Astuti et al. (2022)	Acc	86.5%
Gaudin et al. (2024, Med Sci)	AUC / F1 (age-matched)	0.840 / 0.74
Gaudin et al. (2024, Diagnostics)	Se / AUC	0.83 / 0.76
Brangakgi et al. (2024)	Acc	85%
Fanelli et al. (2025)	AUC / Se (external)	0.786 / 90.5%

Note: AUC = area under the receiver operating characteristic curve; Se = sensitivity; Sp = specificity; Acc = accuracy; — = not reported in the extractable text. Values are the principal result reported by each study. Source: Data extracted from the included studies.

Deep convolutional networks dominated the literature, appearing in 8 of the 10 studies, and the architectures were drawn almost entirely from standard computer-vision models, including AlexNet, VGG-16, GoogLeNet, ResNet, EfficientNet, MobileNet, and YOLOv8, rather than purpose-built dental networks.(31,32,34,35)The two studies that used classical machine-learning classifiers, a support vector machine and a random forest, built on hand-crafted features of cortical width and texture, and their accuracy was in the same range as the deep models, which suggests that the cortical signal itself, rather than any particular network, carries most of the discriminative information.(27,28)

The choice of input region emerged as a recurring design decision: studies that cropped the mental foramen or the mandibular inferior cortex reported that the models attended to the cortical border, and one study found that the mental foramen region gave higher sensitivity than the angulus region, while gradient-based visualization repeatedly pointed to the lower mandibular cortex as the informative region.(34,35) Adding clinical covariates such as age, sex, and body mass

index to the image features improved discrimination in the one study that tested it formally, with the ensemble model outperforming the image-only models on the area under the curve.(31)

A comparison across paradigms is only suggestive, because the studies differ in reference standard, sample, and input region, but the literature does not show a clear advantage of deeper or more modern architectures over the earlier ones, a pattern that is typical of small single-centre datasets in which differences in sample composition swamp differences in model capacity. The three transfer-trained networks in one study, for example, agreed closely with one another and with the radiologist's grading, with accuracy between 86% and 91% and no network clearly dominant, which again points to the cortical label rather than the network as the source of the signal.(32)

Where a comparison was made across five architectures on the same cortical task, the ranking depended on the specific class split rather than on a single best model, so no architecture can be recommended over the others on the current evidence.(29)

Comparison with conventional indices and with human observers (P3)

Direct comparison against conventional mandibular indices and against human observers was the least developed aspect of the included evidence. Only one study trained a model to reproduce a radiologist's mandibular cortical index grades, reaching 81.14% three-class accuracy, a performance in line with expert agreement but not clearly superior to it.(29)The remaining studies evaluated their models against densitometry rather than against the manual cortical measurements that a dentist would actually use in practice, so the review cannot quantify the added value of a model over a clinician who measures the cortical width by eye.

The automated system of Astuti et al. sits on the boundary between the two paradigms, because it computes the cortical width and grayscale texture that a radiologist would measure by hand and then feeds them to a support vector machine, and its 86.5% accuracy is a reminder that automating the conventional indices is a distinct route from end-to-end learning.(27)None of the included studies reported a rigorous, blinded, multi-reader comparison against dentists or radiologists as a primary analysis, and one study acknowledged that it remained unclear whether the model was superior to clinicians and deferred that comparison to future work.(31)The claims of clinical readiness in this literature therefore rest on accuracy against densitometry, not on demonstrated superiority to the current standard of care.

Risk of bias and certainty of the evidence (P4)

Risk of bias was high or unclear across most studies, and the certainty of the evidence is correspondingly low (Table 3). Patient selection carried the most concern, because convenience and case-control designs with artificially separable groups were recurrent, and the clearest example is the study that achieved 98% accuracy against a young control cohort, a comparison that is not clinically meaningful.(33) The reference standard was generally well chosen, with nine of the 10 studies using DXA, but the index test was in most reports developed and evaluated on the same single-centre dataset with internal splitting rather than a held-out external cohort, which risks overestimating performance. Only one study reported external validation, and its area under the curve fell from internal to external testing, a result that shows the optimism internal evaluation can produce.(30)

Table 3. Risk of bias across the included studies (QUADAS-2 domains).

Study (year)	Patient selection	Index test	Reference standard	Flow and timing
Roberts et al. (2013)	U	U	L	U
Lee et al. (2020)	U	U	L	U
Nakamoto et al. (2022)	U	U	L	U
Tassoker et al. (2022)	U	U	U	U
Sukegawa et al. (2022)	U	U	L	U
Astuti et al. (2022)	H	U	L	U
Gaudin et al. (2024, Med Sci)	H	U	L	U
Gaudin et al. (2024, Diagnostics)	U	U	L	U
Brangakgi et al. (2024)	H	U	L	U
Fanelli et al. (2025)	L	L	L	L

Note: H = high risk; U = unclear risk; L = low risk. A study rated H in any domain is considered at high risk overall for that domain. Source: Assessment of the included studies.

Flow and timing were incompletely described in most reports, and the interval between the radiograph and densitometry was not stated in several of the included studies. Two studies appeared in venues with limited editorial or peer-review infrastructure, and their accuracy estimates should be treated with particular caution.⁽³⁶⁾ No study reported a low risk of bias in all domains, and the aggregate evidence was judged to be at serious risk of bias overall. Across the four domains, the reference standard was the least problematic, with nine of the ten studies using densitometry, whereas patient selection and the index test carried most of the concern, a pattern that reflects how these studies were assembled rather than any difficulty with the reference test.

DISCUSSION

The main finding, in answer to P1, is that artificial-intelligence models can detect osteoporosis on dental panoramic radiographs with encouraging accuracy, with area-under-the-curve values between 0.76 and 0.893 in the studies that reported it, but that this accuracy has not yet been demonstrated under conditions that would support clinical deployment. The estimates are broadly consistent with the prior systematic reviews of this specific application, which have reported pooled or median sensitivities and specificities in a comparable range, and with reviews of deep learning for osteoporosis on plain radiographs more generally, which likewise find high reported accuracy alongside a pervasive risk of bias.^(19,21,22)

The convergence between the present synthesis and these reviews is notable because the reviews were conducted with different search strategies, different inclusion decisions, and different years of coverage, yet they reach the same qualitative verdict, which is itself a form of robustness for the direction of the finding even as the exact pooled estimates differ. The agreement across reviews is reassuring for the direction of the effect, but it should not be mistaken for confirmation, because the reviews draw on the same underlying studies and therefore share their limitations.

The most important qualification is that reported accuracy is inflated by designs in which cases and controls are separated by age and disease status in ways that would not occur in a real screening population, and the highest figures in the present review, including the 98% accuracy against young controls, are precisely the ones that are least clinically meaningful.⁽³³⁾ When the comparison is made hard, by matching controls for age and sex, the performance falls to a level that is still useful but no longer impressive, and this pattern, rather than the headline numbers, is the honest summary of P1. A further caveat on the quantitative picture is that the area under the curve is a relative measure of discrimination rather than a statement about how many patients would be correctly referred, and the clinically relevant operating characteristics depend on the chosen threshold, which the studies rarely justified.

A model with an area under the curve of 0.85 will behave very differently in a low-prevalence screening setting, where the positive predictive value is dominated by the false-positive rate, than in the enriched case-control samples in which most of these models were trained. None of the included studies reported calibration or a prospective estimate of the number of patients who would be referred and the number of fractures that would be prevented, and that is the information a clinician actually needs before adopting a screening instrument, so the accuracy figures, however favorable, answer only part of the question.

On P2, the architectures and inputs, the literature is technologically conventional: standard convolutional networks applied to either the full radiograph or a cropped cortical region, with no evidence that newer or larger architectures outperform the older ones once the sample and comparison are held constant.

The finding that classical feature-based classifiers reach accuracy comparable to deep networks is consistent with the wider radiomics literature, in which hand-crafted features and end-to-end learning often converge on similar performance for moderate-sized medical datasets, and it cautions against attributing the field's progress to deep learning per se rather than to the cortical signal that has been known since the earliest mandibular-index studies.^(20,28)

The choice of input region matters more than the choice of network: models concentrated on the mandibular inferior cortex and the mental foramen, which is exactly where the conventional indices measure, and the visualization results that point to the lower cortical border confirm that the models are, at least in part, rediscovering the anatomy that radiologists have used for decades.^(34,37) Adding clinical covariates improved discrimination in the single study that tested it, and this is a promising direction because age and sex are known to carry independent information about fracture risk, but it also complicates the claim that the radiograph alone is sufficient.⁽³¹⁾

What is missing from P2 is any systematic comparison of input regions, preprocessing, and covariate strategies within a single dataset, and until such a comparison exists the design choices

in this literature remain a matter of habit rather than evidence. The radiomics approach tested in one included study is instructive here, because it showed that the choice of feature extraction and transfer-learning strategy changed the area under the curve by more than a quarter of a point across otherwise comparable models, and that some configurations achieved high sensitivity only by sacrificing almost all specificity.(30)

This means the published headline for any given study is one point in a distribution of possible results, not a stable property of the method, and the field would benefit from reporting the full range of configurations tested rather than only the best. The apparent plateau in accuracy across a decade of architectures, from the early random forest to the recent ensembles, reinforces the point that the signal is cortical and that the gains now available lie in the clinical covariates, the reference standard, and the study population rather than in more elaborate networks.

On P3, the comparison with conventional indices and with human observers, the evidence is the weakest part of the field, and this is itself an important finding. The conventional mandibular cortical width and index have a long validation history and a known, modest sensitivity, and they have been proposed as a low-cost referral tool precisely because they trade sensitivity for simplicity.(4,6,7)

The artificial-intelligence models report numbers in a similar or somewhat higher range, but almost none of the included studies benchmarked the model directly against the manual index or against a clinician reading the same images, so the headline claim, that the model automates and improves the existing screening pathway, is asserted rather than demonstrated. The one study that trained a model to reproduce the radiologist's cortical grades reached three-class accuracy of about 81%, which suggests the model learns the index but does not obviously exceed the human who defined it.(29)

This gap matters for translation: a dentist who already measures the cortical width by eye needs evidence that the model is more accurate, faster, or more consistent than that simple, free procedure, and the present literature does not yet supply it. The comparison with human observers has been made more thoroughly in adjacent fields, and the absence of a rigorous multi-reader comparison here is a specific, addressable weakness rather than a fundamental limitation of the method.(37,38) There is also a wider problem of who would act on the model's output, because surveys of practicing dentists find that knowledge of osteoporosis and of the radiographic signs of low bone density is uneven, and a referral tool is only useful if the clinician using it recognizes both its purpose and its limits.(39) A model embedded in a dental workflow would therefore need to be paired with clear guidance and with a pathway to densitometry, and the literature so far stops at the diagnostic test without examining that downstream chain.

On P4, the risk of bias is high and the certainty of the evidence is low, which tempers every quantitative claim in this review. The dominant problems are small single-centre samples, internal rather than external evaluation, and case-control designs with artificially separable groups, all of which bias accuracy upward; the single externally validated study is the only one that can be said to estimate generalization, and it showed a fall in performance from internal to external testing that quantifies the optimism of the rest.(30)

These are not unique to this application but reproduce the known weaknesses of diagnostic artificial-intelligence studies across medicine, which have been documented in systematic reviews of the field.(23,40)

A further recurring problem in this literature is the mismatch between the index test and the target condition: models trained to reproduce a radiologist's cortical grades are not being tested against osteoporosis itself, and models tested against densitometry often omit the interval between the radiograph and the scan, which can be long enough to matter in a progressive disease.

The certainty of the evidence, judged informally, is very low, and the practical implication is that the published accuracy figures should be read as upper bounds on what a deployed model would achieve rather than as predictions of real-world performance.

The geographic distribution of the evidence also bears on its generalizability. The included studies come from 9 countries across East Asia, Europe, Southeast Asia, and the Middle East, but the samples are concentrated in Japanese, Korean, and European populations, with single studies from Indonesia and Syria and none from Africa or Latin America.

This matters because cortical morphology, and the mandibular cortical index in particular, varies with ancestry, body size, and dental status, so a model or threshold developed in one population may not transfer to another without recalibration.(17)

The absence of African and Latin American data is a specific gap, because those regions carry a growing osteoporosis burden and could benefit most from a low-cost screening pathway embedded in routine dental imaging, and a model that has been validated only on East Asian and

European women cannot be assumed to serve them.

Generalizability across health systems is likewise untested: the models assume access to a panoramic machine and to densitometry for confirmation, and the referral pathway that a model implies may not exist in settings where densitometry is scarce, which inverts the value proposition of opportunistic screening and would make a model's output an unreachable recommendation rather than a step toward care.

The implications for practice and research follow from the answers to the four questions. For practice, the models are not ready to replace or even clearly to augment existing opportunistic screening, because their accuracy has been measured mostly in conditions that flatter them and because they have not been shown to beat the simple, established act of measuring the cortical width.

For research, the field would be advanced most by external validation on consecutive, unselected patients; by head-to-head comparison against the mandibular cortical index and against human readers; by reporting a common set of metrics against a common densitometric reference; and by testing whether adding clinical covariates, the one design choice with direct evidence of benefit, generalizes beyond a single institution.

A dental artificial-intelligence model that screens for osteoporosis is an attractive idea precisely because the images already exist and the disease is important, but the idea is only as good as the evidence that the model works outside the dataset on which it was built, and that evidence is the next thing this literature must produce.

The research agenda is therefore specific: prospective studies that enroll consecutive, unselected dental patients, link each radiograph to a contemporaneous densitometric result, and report sensitivity, specificity, and calibration at a predefined threshold would provide the numbers that the current case-control literature cannot; direct comparison against the mandibular cortical index measured by the same clinicians who would use the model would establish whether automation adds value; and a defined referral pathway, with its downstream fracture outcomes, would show whether the whole opportunistic-screening strategy, and not merely the classifier, improves care. Each of these three pieces is measurable and, unlike the model itself, does not require a technical breakthrough to produce.

Limitations

- The search was restricted to three databases and to English, and although the three sources have broad coverage, relevant studies indexed elsewhere or published in other languages, including a cluster of Korean-language studies on this exact question, were not included, so the result may understate work from East Asia.
- Full-text retrieval was limited to legitimate open access, which left 11 potentially eligible reports unassessed and excluded several otherwise relevant studies published behind paywalls; this is a known constraint of reviews that do not have subscription access and is a source of potential selection bias.
- No meta-analysis was performed because the studies were too heterogeneous, so the synthesis is qualitative and does not yield a pooled accuracy with a confidence interval. Each of these limitations points to a concrete remedy: a broader multilingual search, retrieval of paywalled full texts through institutional access, dual screening, and a meta-analysis restricted to studies that report a common metric against DXA.
- The exclusion of non-English studies is the most consequential of these choices, because it is plausible that a substantial body of Korean and Japanese work, countries with strong dental-imaging research programs, remains outside this synthesis, and a future update should treat that literature as a priority rather than a convenience.

CONCLUSIONS

Artificial-intelligence models can detect osteoporosis on dental panoramic radiographs with reported accuracy that is promising, with area-under-the-curve values between 0.76 and 0.893, but the evidence is limited by small single-centre datasets, internal evaluation, case-control designs with artificially separable groups, and a high or unclear risk of bias in most studies, and only one study reported external validation.

Deep convolutional networks dominate the field and attend to the mandibular cortical regions that the conventional indices already target, yet the models have not been shown to outperform those indices or human readers in direct comparison. The reported accuracy should therefore

be treated as an upper bound on clinical performance rather than a prediction of it.

The next step for the field is external validation on consecutive, unselected patients together with direct benchmarking against the established mandibular cortical index and against human observers, and until that evidence exists, the use of these models as opportunistic screening instruments remains premature.

The question that remains open is therefore not whether artificial intelligence can read the mandibular cortex, which the included studies show it can, but whether it can do so well enough, consistently enough, and in populations diverse enough to justify changing how osteoporosis is found, and that question has yet to be answered. The evidence gathered here is best read as the first stage of that answer rather than as the answer itself.

REFERENCES

1. Compston JE, McClung MR, Leslie WD. Osteoporosis. *The Lancet*. 2019;393(10169):364-376. doi: 10.1016/s0140-6736(18)32112-3
2. Das S, Kotwal V, Thakur A, T L, Kumar N. Bone Health and Osteoporosis Management in the Elderly: A Review of Guidelines and Orthopedic Implications. *Cureus*. 2025;17(9):e93229. doi: 10.7759/cureus.93229
3. Mirzaei A, Zabihyeganeh M, Tabrizian P, Heydari P, Farajimoghadam F. Osteoporotic fractures in adults under 65 without prior indications for BMD testing: a call to reevaluate screening guidelines. *BMC Musculoskelet Disord*. 2025;26(1):719. doi: 10.1186/s12891-025-08928-x
4. Japanese Society for Oral and Maxillofacial Radiology, Taguchi A, Tanaka R, Kakimoto N, Morimoto Y, Arai Y, et al. Clinical guidelines for the application of panoramic radiographs in screening for osteoporosis. *Oral Radiol*. 2021;37(2):189-208. doi: 10.1007/s11282-021-00518-6
5. Mupparapu M, Akintoye SO. Application of Panoramic Radiography in the Detection of Osteopenia and Osteoporosis—Current State of the Art. *Curr Osteoporos Rep*. 2023;21(4):354-359. doi: 10.1007/s11914-023-00807-5
6. Munhoz L, Morita L, Nagai AY, Moreira J, Arita ES. Mandibular cortical index in the screening of postmenopausal at low mineral density risk: a systematic review. *Dentomaxillofacial Radiology*. 2021;50(4):20200514. doi: 10.1259/dmfr.20200514
7. Teterina A, Niratisairak S, Morseth B, Bolstad N. Diagnostic efficacy of radiomorphometric indices for predicting osteoporosis in a Norwegian population in the Tromsø Study: Tromsø7. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*. 2023;135(3):444-455. doi: 10.1016/j.oooo.2022.10.039
8. Hwang JJ, Lee JH, Han SS, Kim YH, Jeong HG, Choi YJ, et al. Strut analysis for osteoporosis detection model using dental panoramic radiography. *Dentomaxillofacial Radiology*. 2017;46(7):20170006. doi: 10.1259/dmfr.20170006
9. Kinalski MA, Boscatto N, Damian MF. The accuracy of panoramic radiography as a screening of bone mineral density in women: a systematic review. *Dentomaxillofacial Radiology*. 2020;49(2):20190149. doi: 10.1259/dmfr.20190149
10. Sangeetha, Kailasam S, Selvamuthukumar S, Kumar S. Comparison of Accuracy of Prediction of Osteoporosis in Digital Panoramic Radiography with DEXA Scan in Postmenopausal Women: A Pilot Study. *J Indian Acad Oral Med Radiol*. 2013;25(4):281. doi: 10.4103/0972-1363.161051
11. Munhoz L, Gil Choi I, Miura D, Watanabe PA, Arita E. Bone mineral density and mandibular osteoporotic alterations in panoramic radiographs: Correlation by peripheral bone densitometry in men. *Indian J Dent Res*. 2020;31(3):457. doi: 10.4103/ijdr.ijdr_440_18
12. Alapati S, Reddy R, Tatapudi R, Kotha R, Bodu N, Chennoju S. Identifying risk groups for osteoporosis by digital panoramic radiography. *Contemp Clin Dent*. 2015;6(6):253. doi: 10.4103/0976-237x.166833
13. Panoramic radiographs to find osteoporosis. *Dental Abstracts*. 2015;60(4):206-207. doi: 10.1016/j.dentabs.2015.05.022
14. Gupta P, Khaitan T, Shukla A, Naik S. Is panoramic mandibular index a reliable marker in the evaluation of bone mineral density and sexual dimorphism?. *J Forensic Dent Sci*. 2019;11(3):133. doi: 10.4103/jfo.jfds_92_19
15. Mitsis VI, Lambrou GI. Diagnostic methods of osteopenia and osteoporosis with the use of dental panoramic radiograph. *JRPMS*. 2023;7(3):96-100. doi: 10.22540/jrpms-07-096
16. Audhi MR, Salsabila S. The Analysis Study of Diagnostic Performance and Accuracy of Panoramic Radiography to Screen Bone Mineral Density in Woman: A Comprehensive Systematic Review. *the Indonesian j gen med*. 2024;2(2):1-20. doi: 10.70070/88g24546
17. Prijatelj V, Grgic-Chavez O, van der Tas J, Andaur Navarro CL, Uitterlinden AG, Rivadeneira F, et al. Genetic and epidemiologic assessment of mandibular cortical indices and bone mineral density in peripubertal children: the Generation R study. *Clin Oral Invest*. 2025;29(12):585. doi: 10.1007/s00784-025-06653-2
18. Barra SG, Arruda JAAD, Souza AF, Brasileiro CB, Camargos BM, Lima EMD, et al. Indices in dental image exams for bone mineral density evaluation of aromatase inhibitor users. *Braz oral res*. 2022;36:e138. doi: 10.1590/1807-3107bor-2022.vol36.0138
19. Yen TY, Ho CS, Chen YP, Pei YC. 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. doi: 10.3390/diagnostics14020207
20. Liu RW, Ong W, Makmur A, Kumar N, Low XZ, Shuliang G, et al. Application of Artificial Intelligence Methods on Osteoporosis Classification with Radiographs—A Systematic Review. *Bioengineering*. 2024;11(5):484. doi: 10.3390/bioengineering11050484
21. Ghasemi N, Rokhshad R, Zare Q, Shobeiri P, Schwendicke F. Artificial intelligence for osteoporosis detection on panoramic radiography: A systematic review and meta analysis. *Journal of Dentistry*. 2025;156:105650. doi: 10.1016/j.jdent.2025.105650
22. Tarighatnia A, Amanzadeh M, Hamedan M, Mohammadnia A, Nader ND. Deep learning-based evaluation of panoramic radiographs for osteoporosis screening: a systematic review and meta-analysis. *BMC Med Imaging*. 2025;25(1):86. doi: 10.1186/s12880-025-01626-z
23. Khadivi G, Akhtari A, Sharifi F, Zargarian N, Esmaeili S, Ahsaie MG, et al. Diagnostic accuracy of artificial intelligence models in detecting osteoporosis using dental images: a systematic review and meta-analysis. *Osteoporos Int*. 2025;36(1):1-19. doi: 10.1007/s00198-024-07229-8
24. Dimai HP. New Horizons: Artificial Intelligence Tools for Managing Osteoporosis. *The Journal of Clinical Endocrinology & Metabolism*. 2023;108(4):775-783. doi: 10.1210/clinem/dgac702

25. He Y, Lin J, Zhu S, Zhu J, Xu Z. Deep learning in the radiologic diagnosis of osteoporosis: a literature review. *J Int Med Res.* 2024;52(4):03000605241244754. doi: 10.1177/03000605241244754
26. Zhang W, Tai D, Kang S, Li K. The application progress of artificial intelligence in osteoporosis diagnosis. *Front Artif Intell.* 2025;8:1699762. doi: 10.3389/frai.2025.1699762
27. Astuti ER, Arifin AZ, Indraswari R, Putra RH, Ramadhani NF, Pramatika B. Computer-Aided System of the Mandibular Cortical Bone Porosity Assessment on Digital Panoramic Radiographs. *Eur J Dent.* 2023;17(02):464-471. doi: 10.1055/s-0042-1749158
28. Roberts MG, Graham J, Devlin H. Image Texture in Dental Panoramic Radiographs as a Potential Biomarker of Osteoporosis. *IEEE Trans Biomed Eng.* 2013;60(9):2384-2392. doi: 10.1109/tbme.2013.2256908
29. Tassoker M, Öziç MÜ, Yuce F. Comparison of five convolutional neural networks for predicting osteoporosis based on mandibular cortical index on panoramic radiographs. *Dentomaxillofacial Radiology.* 2022;51(6):20220108. doi: 10.1259/dmfr.20220108
30. Fanelli F, Guglielmi G, Troiano G, Rivara F, Passeri G, Prencipe G, et al. Development of AI-Based Predictive Models for Osteoporosis Diagnosis in Postmenopausal Women from Panoramic Radiographs. *JCM.* 2025;14(13):4462. doi: 10.3390/jcm14134462
31. 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. doi: 10.1038/s41598-022-10150-x
32. Nakamoto T, Taguchi A, Kakimoto N. Osteoporosis screening support system from panoramic radiographs using deep learning by convolutional neural network. *Dentomaxillofacial Radiology.* 2022;51(6):20220135. doi: 10.1259/dmfr.20220135
33. 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. *Medical Sciences.* 2024;12(3):49. doi: 10.3390/medsci12030049
34. 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. *JCM.* 2020;9(2):392. doi: 10.3390/jcm9020392
35. Gaudin R, Vinayahalingam S, van Nistelrooij N, Ghanad I, Otto W, Kewenig S, et al. AI-Powered Identification of Osteoporosis in Dental Panoramic Radiographs: Addressing Methodological Flaws in Current Research. *Diagnostics.* 2024;14(20):2298. doi: 10.3390/diagnostics14202298
36. Brangakgi FH, Khadra Y. Building a Deep Learning Model to Detect Osteoporosis from Dental Panoramic X-Ray Image. *OJAppS.* 2024;14(12):3480-3489. doi: 10.4236/ojapps.2024.1412227
37. Turosz N, Chęcińska K, Chęciński M, Brzozowska A, Nowak Z, Sikora M. Applications of artificial intelligence in the analysis of dental panoramic radiographs: an overview of systematic reviews. *Dentomaxillofac Radiol.* 2023;52(7):20230284. doi: 10.1259/dmfr.20230284
38. Yeung AWK, Mozos I. The Innovative and Sustainable Use of Dental Panoramic Radiographs for the Detection of Osteoporosis. *IJERPH.* 2020;17(7):2449. doi: 10.3390/ijerph17072449
39. Ferreira RV, Roizenblatt S, Szejnfeld VL. Dentists' knowledge about osteoporosis and their ability to identify the disease. *Adv Rheumatol.* 2021;61(1):49. doi: 10.1186/s42358-021-00206-3
40. Gatineau G, Shevroja E, Vendrami C, Gonzalez-Rodriguez E, Leslie WD, Lamy O, et al. Development and reporting of artificial intelligence in osteoporosis management. *Journal of Bone and Mineral Research.* 2024;39(11):1553-1573. doi: 10.1093/jbmr/zjae131

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Neisy Pérez Ramos.

Writing–review and editing: Neisy Pérez Ramos.