Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Diagnostic accuracy meta-analysis of artificial intelligence for mandibular third molar and inferior alveolar nerve assessment

Pietro Montagna
University of Verona
Riccardo Nocini
University of Verona
Paolo Faccioni
University of Verona
Siwei Wang
Zunyi Medical University
Giulio Conti
giulio.conti@univr.it
University of Verona
Federica Melloni
University of Verona
Giulio Montagna
University of Trieste
Alice Sammarco
University of Verona
Daniele De Santis
University of Verona

Systematic Review

Keywords: Molar, Third, Artificial Intelligence, Radiomics, Radiography, Panoramic, Oral Surgery

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11365834/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Diagnostic accuracy meta-analysis of artificial intelligence for mandibular third molar and inferior alveolar nerve assessment

Pietro Montagna [1†], Riccardo Nocini [2†], Paolo Faccioni [1], Siwei Wang [3], Giulio Conti [1*], Federica Melloni [1], Giulio Montagna [4], Alice Sammarco [1], and Daniele De Santis [1]

[1] Head and Neck Department, Department of Surgery, Dentistry, Pediatrics and Gynecology, University of Verona, Verona, Italy.
[2] Department of Otolaryngology Head and Neck Surgery, University of Verona, Piazzale L.A. Scuro 10, Verona, 37134, Italy.
[3] Department of Dental Implantology, The Affiliated Stomatological Hospital, Zunyi Medical University, Zunyi, Guizhou, China.
[4] Department of Medicine, Surgery and Health Sciences, University of Trieste, Trieste, Italy.
* Correspondence: Dr. Giulio Conti; Head and Neck Department, Department of Surgery, Dentistry, Pediatrics and Gynecology, University of Verona, Verona, Italy. E-mail: giulio.conti@univr.it.

Abstract.

Artificial intelligence applied to panoramic radiographs has been proposed to triage which mandibular third molars require cone-beam computed tomography before extraction. We searched seven databases and registers to June 2026 for studies of any machine- or deep-learning model applied to panoramic radiographs against a three-dimensional reference standard, appraised them with QUADAS-3 and PROBAST+AI, and rated certainty with GRADE. Twenty-one studies were included. Across seven pooled studies, a Bayesian bivariate model gave a summary sensitivity of 0.81 (95% CrI 0.75–0.87) and specificity of 0.87 (95% CrI 0.70–0.95) for detecting mandibular third molar–inferior alveolar nerve contact, yet every study was at high risk of bias, only one was externally validated, and certainty was very low. In nine within-study comparisons no study reported readers as more accurate than the model, though each was internal and unblinded. No architecture was consistently superior and prognostic models for nerve injury remained preliminary. Registered in PROSPERO (CRD420251000827).

Keywords: Molar, Third; Artificial Intelligence; Radiomics; Radiography, Panoramic; Oral Surgery.

1. Introduction

Removal of impacted mandibular third molars (M3M) is among the most frequently performed oral surgical interventions, and is widely regarded as a safe procedure [1]. Nevertheless, as with any surgery, it carries a certain risk, mostly related to injury to the inferior alveolar nerve (IAN). Such occurrence causes temporary or, less often, permanent, altered sensation of the lip, chin and ipsilateral teeth [2] and, although the overall incidence of IAN injury is variably reported to be low, ranging from 0.35% [3] up to 10% [4], the risk rises markedly when the roots of the third molar are in close contact with the mandibular canal or other specific radiographic features are detected [4]. A careful preoperative assessment of the relationship between the mandibular canal (MC) and the M3M apices is therefore central to surgical planning, allowing the clinician to estimate risk, inform the patient and weigh therapeutic alternatives. This is all the more relevant now that coronectomy has proven a viable option in high-risk cases, both clinically and economically [1,5,6].

Radiographic assessment rests on a set of features associated with a higher likelihood of close root–canal contact, namely (i) root darkening adjacent to the MC, (ii) root deflection in the vicinity of the MC, (iii) root narrowing as the roots approach the MC, (iv) a dark line at the root apex, (v) interruption of the cortical line

† These authors contributed equally to this work and should be regarded as co-first authors.

forming the roof of the MC, (vi) narrowing of the MC, and (vii) diversion of the MC. For many decades, clinicians have relied on these radiographic warning signs on panoramic images (PR) to flag teeth at higher risk [7], as the presence of one or more of them suggests a higher probability of direct root–MC contact and, consequently, an increased risk of IAN injury during extraction [8].

Panoramic radiography (PR) is the routine first-line examination before third molar surgery, being inexpensive, widely available and low in radiation dose; as a two-dimensional projection, however, it cannot resolve the true three-dimensional relationship between the tooth and the MC [9,10]. Cone-beam computed tomography (CBCT) supplies this three-dimensional information and serves as the reference standard for classifying true contact, but at greater cost and radiation exposure [11]. Moreover, randomised comparisons have shown that CBCT-based planning does not consistently reduce postoperative complications relative to PR-based planning [12], so routine CBCT is neither feasible nor clearly justified for every patient. This creates a genuine tension: IAN injury is a recurrent source of medicolegal claims after third molar surgery [2], and the absence of CBCT may be construed as a shortcoming should injury occur, yet from a clinical and ethical standpoint its use as a first-line examination is difficult to justify, not least because it has little demonstrable effect on the incidence of IAN injury. Artificial intelligence (AI) has been proposed to bridge this gap. In current literature, AI is being trained to flag M3M–IAN contact directly on the panoramic image and could act as a triage test that selects which patients should proceed to CBCT. A parallel line of work moves beyond the anatomical relationship to the clinical outcome, building models that predict postoperative IAN injury or paraesthesia from imaging and patient factors. Individual studies report accuracies above 90%, and AI is already being proposed as a chairside triage tool, yet most are small, single-centre and retrospective and validate their models only internally. A single diagnostic-accuracy meta-analysis has so far pooled this evidence, combining seven deep-learning studies to a summary sensitivity of 80.2% and specificity of 85.8% with substantial heterogeneity (I^2>90%) [13]; a companion narrative review summarised deep learning for predicting nerve injury and extraction difficulty [14].

Important questions nonetheless remain open. The evidence base has expanded rapidly through 2025 and 2026 and now extends to machine-learning models beyond convolutional networks, but these studies have not been appraised with the instruments developed specifically for diagnostic and prognostic AI research, namely QUADAS-3, PROBAST+AI [15,16], nor has the certainty of their evidence been graded [17]. We therefore undertook an updated systematic review that deliberately keeps apart the two questions the field tends to merge. For the diagnostic question (FQ1) we evaluated the accuracy of AI on panoramic radiographs against a three-dimensional reference standard, broadening the eligible index tests to include machine learning as well as deep learning, appraising every study with QUADAS-3 [15] and rating certainty with GRADE for diagnostic test accuracy [17]. For the prognostic question (FQ2) we summarised and appraised, with PROBAST+AI [16], the models that predict IAN injury after extraction. Our aim was to establish how accurate, and how certain, the current evidence truly is, and to define what remains to be demonstrated before AI can be trusted at the chairside as a gatekeeper for CBCT referral.

2. Materials and Methods

2.1. Review protocol and registration

The review was registered in PROSPERO (CRD420251000827) and is reported in accordance with PRISMA 2020 and the PRISMA-DTA extension [18,19]; the completed PRISMA-DTA checklist is provided as a supplementary file.

2.2. Eligibility criteria

Two focused questions were addressed: FQ1 on diagnostic accuracy and FQ2 on prognosis. Eligibility was defined against the following elements; where an element differed between the two questions it is specified for each.

- *Types of studies*. Clinical or validation studies of AI models were eligible, of any machine-learning or deep-learning class. Editorials, commentaries, narrative reviews and case reports were excluded.
- *Participants*. Patients with a mandibular third molar assessed for its relationship to the mandibular canal (FQ1) or undergoing third-molar extraction (FQ2).
- *Index test*. For FQ1, an AI model applied to a preoperative panoramic radiography; 3D imaging and segmentation-only models that issued no contact decision were excluded. For FQ2, an AI model predicting postoperative nerve injury from imaging (both 2D or 3D) with or without clinical predictors.
- *Target condition (FQ1) and outcome (FQ2)*. For FQ1, true contact between the third-molar roots and the mandibular canal. For FQ2, postoperative inferior alveolar nerve injury.
- *Reference standard*. For FQ1, three-dimensional imaging (cone-beam or conventional CT), with or without intraoperative confirmation; studies whose reference standard was not three-dimensional were excluded. For FQ2, the clinical neurosensory outcome.
- *Data requirement (FQ1).* A study contributed to the meta-analysis only if a contact 2×2 table was reported or uniquely reconstructable; otherwise it was synthesised narratively.

Among the FQ1 studies meeting these criteria, entry into the primary meta-analytic pool followed prespecified rules, applied in order: (i) only studies with a contact 2×2 were pooled, those without a poolable 2×2 being synthesised narratively; (ii) each patient cohort contributed a single 2×2 table, so that the pooled observations remained independent; (iii) where a study reported more than one model, its primary (proposed) model was used; (iv) where a study reported both an internal test set and an external validation set, the internal test set was used, because most studies reported an internal test set only and this maximised comparability; (v) a study that did not report its accuracy on a single fixed set of observations, each classified by a model that had not seen it, did not enter the primary pool: this excludes both accuracy aggregated over k-fold cross-validation spanning the entire cohort and accuracy averaged over repeated random re-splitting, because in neither case does a single held-out contingency table exist; and (vi) where two reports drew on the same patient cohort, the more complete report entered the primary pool and its companion was reserved for sensitivity analysis. The two shared-cohort pairs, and the report retained from each, are identified in the Results.

2.3. Information sources and search strategy

PubMed, Embase (Ovid), Scopus, Web of Science, IEEE Xplore, CENTRAL and the WHO ICTRP were searched for relevant papers, ranging from 2014 to the 26th of June 2026, without a diagnostic methodological filter. The electronic search was supplemented with backward and forward citation searching and manual

searching in Google Scholar. The full electronic search strategy for each database is provided in Supplementary Note 1.

2.4. Study selection

Study selection followed the PRISMA 2020 and PRISMA-DTA guidelines [18,19]. After manual deduplication, using Microsoft Excel (Microsoft 365 Edition), titles and abstracts were screened independently and in duplicate by two reviewers (P.M. and A.S.); full-text eligibility, through to the final inclusion decisions of the PRISMA flow, was then assessed by the same two reviewers, again independently, with disagreements resolved by a third reviewer (D.D.S.); inter-reviewer agreement at the full-text stage was quantified with Cohen's κ.

2.5. Data extraction

Data were extracted from the included reports independently and in duplicate by two further reviewers (F.M. and G.M.), with disagreements resolved by P.M., on a piloted form covering study and dataset characteristics, the AI model, the reference standard, and accuracy or prognostic metrics. For FQ1, 2×2 tables were read directly from reported confusion matrices, or reconstructed at one of two further levels of directness, recorded per study in Supplementary Table 1b.

2.6. Risk of bias and applicability

Risk of bias and applicability were assessed for FQ1 with QUADAS-3 [15,20], as the corresponding QUADAS-AI is not yet finalized [21], and for FQ2 with the model-evaluation component of PROBAST+AI [16]. Both appraisals were performed independently and in duplicate by P.M. and A.S., with disagreements resolved by D.D.S. Judgements were summarised as traffic-light and summary plots with Risk-of-bias VISualization (RoBVIS) [22].

2.7. Data synthesis and statistical analysis

For FQ1, the studies selected for the primary pool were combined with the bivariate random-effects model of Reitsma [23]. Sensitivity and specificity were not summarised as independent univariate estimates, because across studies they move together with the implicit positivity threshold each model adopts and are, in consequence, negatively correlated; the model estimates the two jointly with their between-study correlation and yields a summary operating point with its 95% credible and posterior predictive regions. The summary curve was drawn on the hierarchical summary ROC (HSROC) parameterisation [24], and the positive and negative likelihood ratios and the diagnostic odds ratio were derived from the summary operating point rather than pooled separately. The bivariate model handles zero cells directly through its exact likelihood, so no continuity correction was applied.

A study contributed to the AI-versus-reader comparison only where the human reading was performed inside that study, on its own test images, and the readers' performance was measured there. Accuracy figures quoted from a different study were not treated as an in-study comparator. Two reports that name human observers were excluded on this basis, Jeon et al., 2023 [9]; Sukegawa et al., 2022 [25], as in both the reader accuracy is a figure reported by Choi et al., 2022 [26].

Heterogeneity was examined with prespecified sensitivity analyses, excluding the reconstructed 2×2 tables; within each shared-cohort pair, substituting the companion report for the primary study; and, post hoc, halving the cells of the one study whose reported table counts the same test images twice. Exploratory meta-regressions were fitted on two study-level covariates, namely publication year and whether the contact 2×2 was read from a published confusion matrix or reconstructed, each entered singly within the bivariate model

rather than fitted separately to sensitivity and specificity. A third covariate, contact prevalence in the test set, was specified but not fitted, because across the pooled studies the observed prevalence spans only 0.32 to 0.56 and a coefficient estimated over so narrow a range of case-mix would carry no information. FQ2 was synthesised without meta-analysis [27], because the few, clinically and methodologically diverse prognostic models could not be pooled validly. The bivariate model was first fitted by maximum likelihood in the MetaDTA application [28], but on the seven pooled studies it does not converge, returning a singular fit at the boundary of the parameter space and no summary estimate. The primary analysis, the sensitivity analyses and the meta-regressions were all fitted instead in MetaBayesDTA [29], with the application's default weakly informative priors: normal(0, 1) on the logit-sensitivity and logit-specificity means, half-normal(0, 1) truncated at zero on the between-study standard deviations, and LKJ(2) on the between-study correlation. Each model was run with four chains of 5000 iterations, the first 1000 discarded as warm-up. All reported fits had zero divergent transitions. Summary estimates are posterior medians and the intervals are 95% credible intervals throughout.

2.8. Assessment of certainty and reporting bias

Certainty of evidence for FQ1 was rated with GRADE for diagnostic test accuracy [17], separately for sensitivity and specificity, and set out as an evidence profile giving the expected true and false positives and negatives per 1000 patients tested. Because accuracy datasets in this field are typically enriched for contact, these consequences were computed at plausible external pre-test probabilities (10%, 30% and 50%) rather than at the study-level prevalence, with 30% taken as the base case. Publication bias was assessed narratively rather than with a funnel-plot asymmetry test [30]. Certainty for FQ2 was described narratively.

3. Results

The database and register searches yielded 288 records; after removal of 186 duplicates, 102 records were screened, 55 were excluded at title and abstract, and 47 were sought for retrieval. Citation searching identified a further three records, of which two were sought, the third having been excluded at title and abstract, so that 49 reports (47 from databases and registers and two from citation searching) were assessed at full text. The full PRISMA Flow Diagram is reported (Fig. 1). Twenty-eight full-text reports were excluded with reasons (listed in Supplementary Table 2), most often a reference standard based on annotation of the panoramic radiograph rather than three-dimensional imaging, leaving 21 studies in the qualitative synthesis: 15 for FQ1 and 6 for FQ2. Inter-reviewer agreement at the full-text stage was almost perfect (raw agreement 46 of 49 reports; Cohen's $\kappa = 0.87$); the three disagreements were adjudicated in favour of inclusion (Sukegawa et al. [25], Takebe et al. [31] and Fernandes et al. [32]). Seven panoramic-index studies were pooled in the primary FQ1 meta-analysis; three further panoramic-index studies with a reconstructable 2×2 were excluded from the primary pool under a prespecified rule, two of them entering a sensitivity analysis in place of the study with which they share a cohort and the third entering no analysis at all. Kumbasar et al. [33], an overlapping cohort with Güller et al. [34], reports one confusion matrix aggregated over ten cross-validation folds spanning its whole cohort. Ulusoy et al. [35] reports, as the caption of its results table states, the mean of thirty Monte-Carlo experiments, each splitting the dataset at random and, in the source's own words, irrespective of contact status, so that no single held-out table and no fixed class composition underlie the published figures; consistent with this, the accuracy, positive and negative predictive values it reports imply mutually incompatible ratios of contact to non-contact observations (3.88, 4.25 and 3.61 against a dataset ratio of 2.91), whereas the same check on every pooled study agrees to within 2%. Sharing no cohort with any other included study, it could not enter a substitution analysis

either, and so contributes to no quantitative synthesis. Fukuda et al. 2026 [36] is an overlapping cohort with Fukuda et al. 2024 [37].

The 15 FQ1 studies were all retrospective and came mostly from East Asia, Turkey and Iran, with single contributions from Norway (Table 1). All used a CBCT or 3D reference standard, in two studies combined with intraoperative findings [9,31]. The AI approaches included detection networks [9,31], classification CNNs [26,33,34,37–39], classical machine learning on extracted features [35,36] and segmentation-plus-classifier or geometry-aware pipelines [40,41].

Figure 1. PRISMA Flow Diagram.

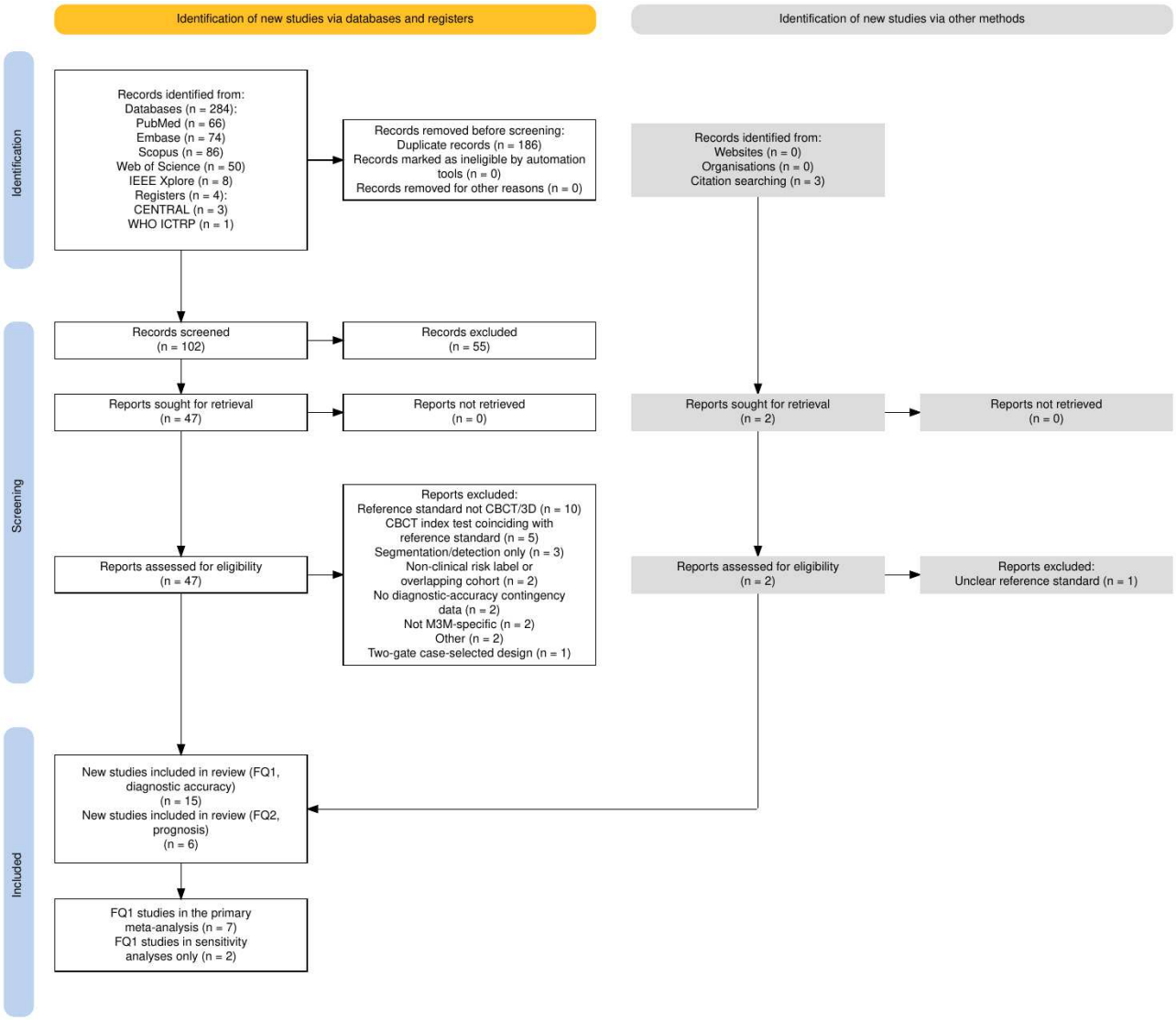


Table 1. Characteristics of the 15 studies addressing FQ1. Studies are ordered by their role in the synthesis: primary meta-analysis (n = 7); excluded from the primary pool by a prespecified rule (n = 3), of which two contributed to a sensitivity analysis and one to none; narrative synthesis (n = 5); within each group they follow the year of publication, and this order is used for the FQ1 studies throughout the paper and the supplementary material. For each study the table reports first author and year, country, study design, the reference standard, the dataset as patients / M3M units / test-set size, the artificial-intelligence architecture, the in-study human comparator where one exists, the reported sensitivity (Se), specificity (Sp) and area under the curve (AUC), and whether the study entered FQ1 meta-analysis. A dash marks a value not reported or not applicable; footnotes: [a] The reader accuracy quoted is a figure cited from Choi et al., 2022 and is not an in-study comparison. [b] The model localised 94 of the 96 internal test images and accuracy was computed on those, so the pooled 2×2 rests on 94; plus an independent external set of 58 images from a second hospital. [c] The reported 2×2 aggregates two training/testing runs over the same 100 test patches (200 observations). [d] Metrics are means over five model initialisations of this fixed test set. [e] Monte-Carlo cross-validation; metrics averaged over thirty random 80:20 splits drawn irrespective of contact status, so no single held-out table exists. [f] No independent held-out test set; performance from cross-validation only. [g] Best-split operating point; the source states in its text that 100 images were analysed, but the reader and model metrics it reports are realisable only on a test set of 112 (65 contact / 47 non-contact). [h] The reported accuracy, recall and specificity do not reconcile to a single 2×2 table.

Study	Country	Study Design	Ground Truth	Dataset	AI architecture	Human comparator	Se / Sp / AUC	In MA (?)
Jeon et al., 2023 [9]	South Korea	Retrospective	CBCT + intraoperative	518 / 901 / 178	EfficientDet-D4	None[a]	Se 0.82; Sp 0.75	Yes (primary)

Yes (primary)	Yes (primary)	Yes (primary)	Yes (primary)	Yes (primary)	Yes (primary)	Yes (sensitivity only)	No
Se 0.92; Sp 0.85	Se 0.79; Sp 0.80; AUC 0.85	Se 0.75; Sp 0.96; AUC 0.95	Se 0.72; Sp 1.00	Se 0.86; Sp 0.62; AUC 0.81	Se 0.85; Sp 0.89	Se 0.95; Sp 0.93; AUC 0.97	Se 0.86; Sp 0.82
5 oral surgeons; AI higher (accuracy 0.89 vs 0.63)	2 radiologists + 2 residents; AI ≈ radiologists, > residents	2 OMF surgeons (consensus); AI higher	None	5 expert dentists; AI higher (AUROC	3 OMF surgery special	None	2 radiologists; AI higher
YOLO detection	GoogLeNet	MM3-IACnet (segmentation + classifier, YOLOv5 backbone)	Inception-v3	ResNet50	ResNet-50	Majority voting (AlexNet + VGG16 + VGG19)	ANN on radiographic features
384 / 579 / 96 (94 analysed)[b]	508 / 800 / 100[c]	1570 / 2543 / 243	290 / 546 / 164	279 / 441 / 89[d]	— / — / 82	290 / 546 / —[f]	1010 / 1478 / 296[e]
CBCT + intraoperative	CT or CBCT	CBCT	CBCT	CBCT	CBCT	CBCT	CBCT
Retrospective	Retrospective	Retrospective	Retrospective	Retrospective	Retrospective	Retrospective	Retrospective
Japan	Japan	China	Turkey	China	Iran	Turkey	Turkey
Takebe et al., 2023 [31]	Fukuda et al., 2024 [37]	Jing et al., 2024 [40]	Güller et al., 2025 [34]	Jiang et al., 2025 [38]	Yari et al., 2025 [39]	Kumbasar et al., 2025 [33]	Ulusoy et al., 2025 [35]

Yes (sensitivity only)	No	No	No	No	No	No
Se 0.80; Sp 0.96; AUC 0.88	Se 0.83; Sp not reported	Se 0.85; Sp 0.55[g]	Se 0.79; Sp not reported; AUC 0.89	Se 0.85; Sp 0.93; AUROC [illegible]	Se not reported; Sp not [illegible]	
2 radiologists + 1 oral surgery	AI comparable (not 5 dentists;	6 OMF specialists (mean 15.3 y); AI	None[a]	None	None	
AutoML on radiographic features	MM3-IANnet (YOLOv4)	ResNet-50	ResNet50v2 + SAM	Faster R-CNN + ResNeXt	CNN, ViT and gMLP (multi-task)	
1372 / 2103 / 100	503 / 915 / 188	394 / 571 / 112	— / 1279 / —[f]	387 / 619 / 58	— / 450 / 43	
CT or CBCT	CBCT	CBCT	CT	CBCT	CBCT	
Retrospective	Retrospective	Retrospective	Retrospective	Retrospective	Retrospective	
Japan	China	South Korea	Japan	Iran	Norway	
Fukuda et al., 2026 [36]	Zhu et al., 2021 [42]	Choi et al., 2022 [26]	Sukegawa et al., 2022 [25]	Soltani et al., 2024 [41]	Fernandes et al., 2025 [32]	

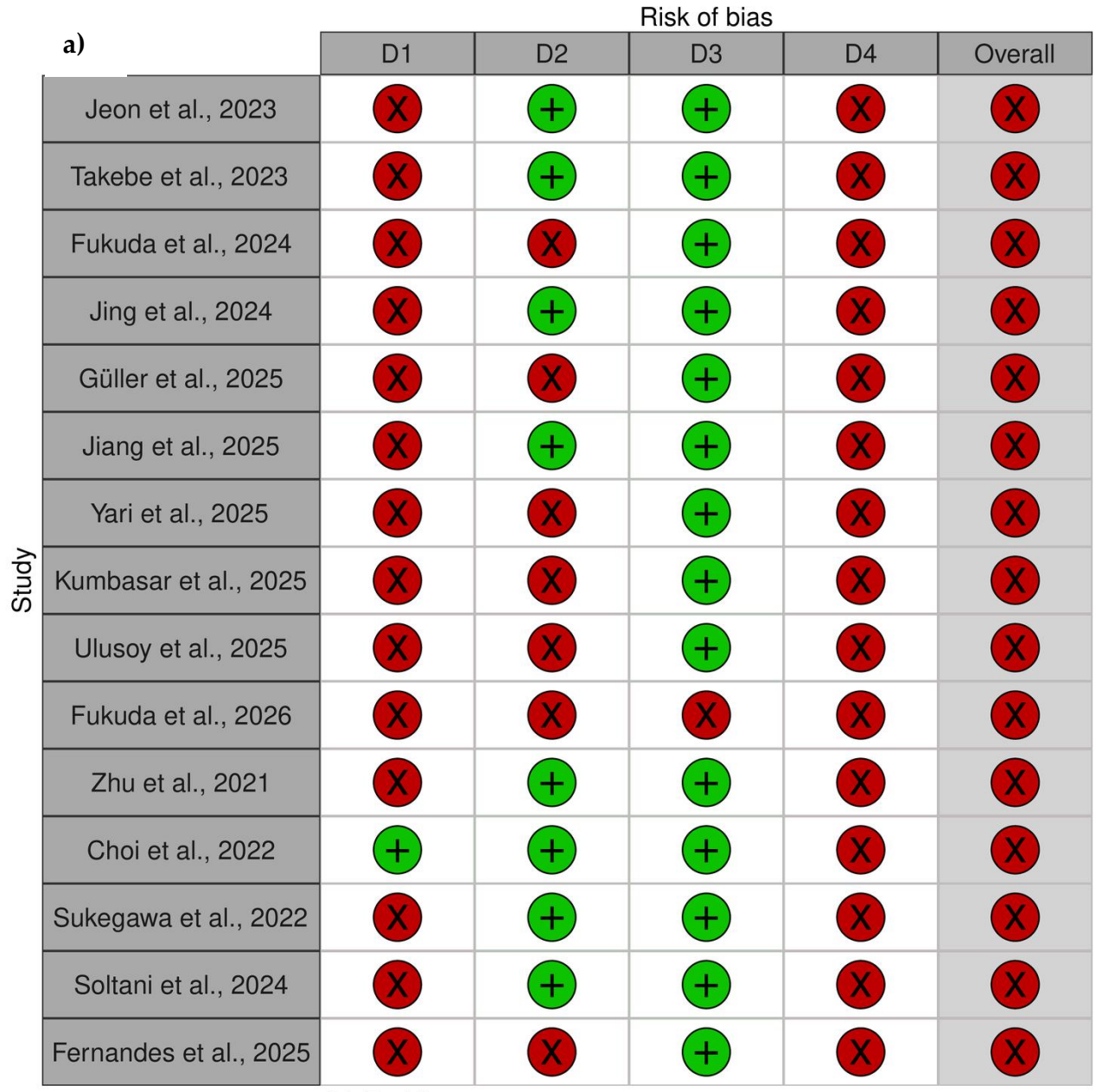
3.1 Diagnostic accuracy of AI on panoramic radiographs for FQ1

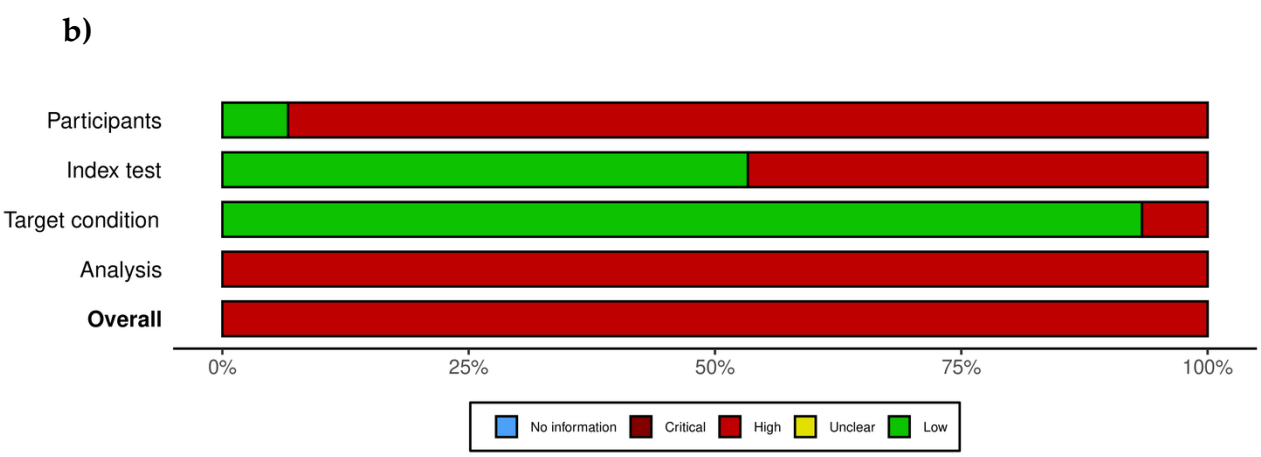
All 15 FQ1 studies were at high overall risk of bias (Fig. 2; per-study judgements are given in Supplementary Table 3). The judgements were driven by the participants domain, where sampling was retrospective and non-consecutive in all but one study and the test set was balanced by construction in several, and by the analysis domain, per-tooth analysis uncorrected for clustering, internal or cross-validation without an external test set in all but one study, post-hoc threshold selection, and, in some studies, reconstructed 2×2 tables. Concerns regarding applicability were high overall in all 15 studies, and again the participants domain drove the judgement: QUADAS-3 anchors applicability to the ideal accuracy study defined in phase 2, whose participants component requires prospective, consecutive or random enrolment of a sample representative of the intended-use population at the prevalence of clinical practice, and no study met it. Concern was also high in the index-test domain in 13 studies, the ideal study requiring a threshold fixed in advance and an image handled as it would be at the chairside, and in the target-condition domain in the two reports of the Fukuda et al. cohort [36,37].

Seven studies were pooled for FQ1, contributing 1050 units[9,31,34,37–40]. The bivariate model gave a summary sensitivity of 0.81 (95% CrI 0.75–0.87) and specificity of 0.87 (95% CrI 0.70–0.95) (Fig. 3 and Fig. 4), with a positive likelihood ratio of 5.98, a negative likelihood ratio of 0.22 and a diagnostic odds ratio (DOR) of 27.62 at the summary operating point. Per-study sensitivity was consistent across the pool (0.72–0.92); specificity ranged more widely, from 0.62 in Jiang et al. [38] to 0.96 in Jing et al. [40] and 1.00 in Güller et al. [34], giving a moderately wide 95% prediction region.

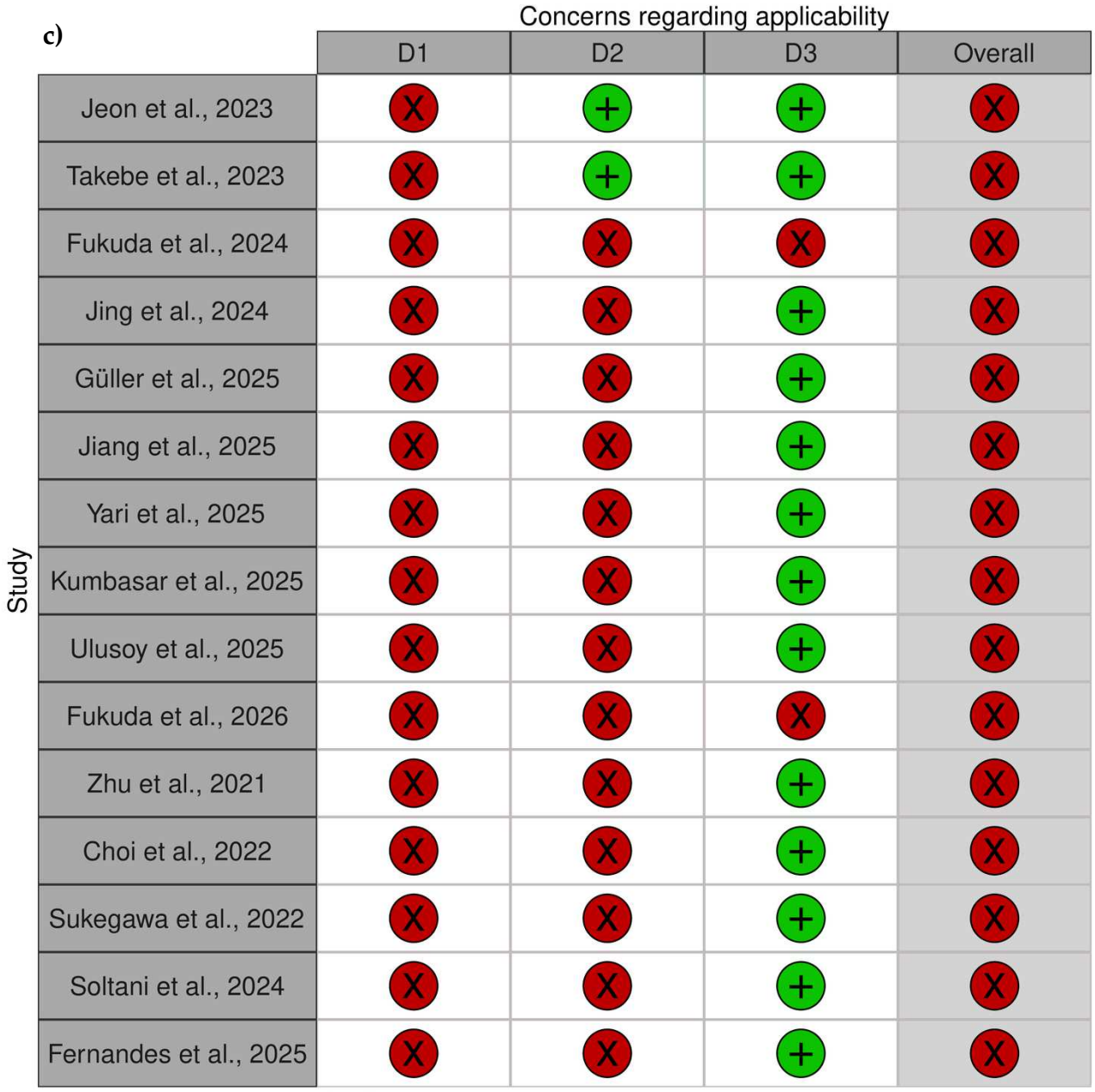
Every sensitivity analysis could be fitted, and across them the summary sensitivity ranged from 0.79 to 0.86 and the specificity from 0.85 to 0.89. Restricting the pool to the four studies whose 2×2 was read directly from a published confusion matrix [9,34,37,39] gave a sensitivity of 0.79 (95% CrI 0.69–0.86) and specificity of 0.87 (95% CrI 0.60–0.96); substituting Kumbasar et al. [33] for Güller et al. [34] in their shared cohort gave 0.86 (0.75–0.92) and 0.85 (0.70–0.92); substituting Fukuda et al., 2026 [36] for Fukuda et al., 2024 [37] gave 0.82 (0.75–0.88) and 0.89 (0.73–0.96). Halving the cells of Fukuda et al., 2024 [37], whose reported table counts the same 100 test images twice, gave 0.81 (0.75–0.87) and 0.87 (0.70–0.95), within a fraction of a point of the primary fit. None of these displaces the primary estimate, so the pool is not driven by any single study or by any of the inclusion decisions. That the same model fitted by maximum likelihood does not converge on these seven studies at all, returning a singular fit at the boundary of the parameter space and no summary estimate, is the plainest measure of how thinly a bivariate model is identified at this number of studies; the Bayesian fit recovers a proper covariance, with a between-study correlation of −0.40 (−0.89 to 0.39), but the width of the resulting prediction region is the honest expression of that same sparsity, and it is why we present the prediction region alongside the summary point. The exploratory bivariate meta-regression showed no association of accuracy with publication year (coefficient on logit sensitivity −0.19, 95% credible interval −0.70 to 0.34; on logit specificity +0.31, −0.76 to 1.37) or with whether the 2×2 table was reported or reconstructed (difference in sensitivity +0.05, −0.08 to 0.17; in specificity −0.05, −0.40 to 0.21). The full sensitivity analyses and meta-regression are given in Supplementary Note 2. Five panoramic studies were synthesised narratively, none providing a poolable contact 2×2: two reported no contact table [32,42]; Choi et al. [26] reported a single best-split sensitivity–specificity pair (0.85/0.55) discordant with its averaged contact accuracy (0.63); Sukegawa et al. [25] reported ten-fold cross-validated metrics averaged over folds (recall 0.79, precision 0.82) without a specificity value or the contact/non-contact counts, so no coherent 2×2 could be reconstructed; and Soltani et al. [41] reported test-set accuracy, recall and specificity (0.85, 0.85 and 0.93) that do not reconcile to one. Certainty was very low for both sensitivity and specificity (Fig. 5; full evidence profile in Supplementary Table 4), downgraded for high risk of bias, indirectness and suspected publication bias, with specificity further downgraded for inconsistency. At a 30% pre-test probability of contact, the summary estimates correspond to about 57 missed contacts and 93 unnecessary CBCT referrals per 1000 patients tested.

Figure 2. QUADAS-3 risk-of-bias and applicability assessment of the 15 FQ1 diagnostic-accuracy studies. (a) Per-study risk-of-bias traffic-light plot and (b) risk-of-bias summary bar plot; (c) per-study applicability traffic-light plot and (d) applicability summary bar plot. Studies follow the order of Table 1. Risk-of-bias domains: D1 participants, D2 index test, D3 target condition, D4 analysis; applicability domains: participants, index test, target condition. Green = low, red = high concern/risk; no domain was rated unclear.

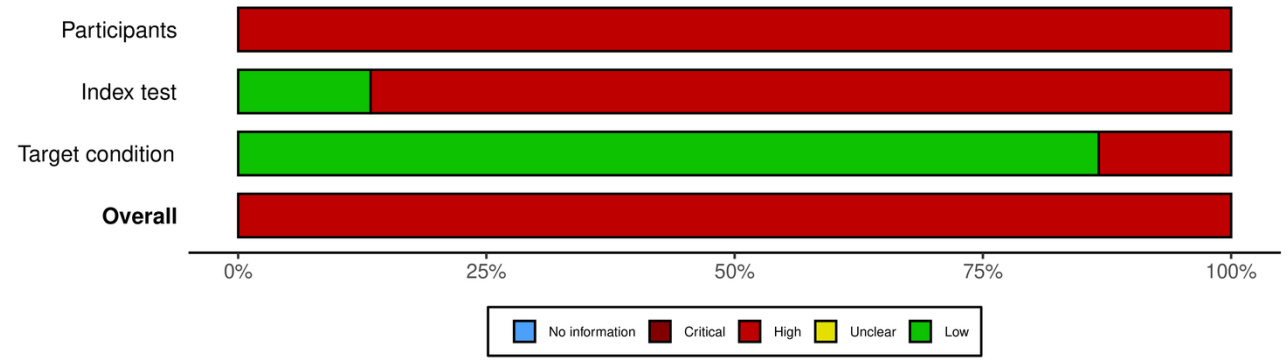

a)
Risk of bias
D1
D2
D3
D4
Overall
Study
Jeon et al., 2023
Takebe et al., 2023
Fukuda et al., 2024
Jing et al., 2024
Güller et al., 2025
Jiang et al., 2025
Yari et al., 2025
Kumbasar et al., 2025
Ulusoy et al., 2025
Fukuda et al., 2026
Zhu et al., 2021
Choi et al., 2022
Sukegawa et al., 2022
Soltani et al., 2024
Fernandes et al., 2025
D1: Participants
D2: Index test
D3: Target condition
D4: Analysis
Judgement
High
Low


b)
Participants
Index test
Target condition
Analysis
Overall
0%
25%
50%
75%
100%
No information
Critical
High
Unclear
Low

c)


Concerns regarding applicability
D1
D2
D3
Overall
Study
Jeon et al., 2023
Takebe et al., 2023
Fukuda et al., 2024
Jing et al., 2024
Güller et al., 2025
Jiang et al., 2025
Yari et al., 2025
Kumbasar et al., 2025
Ulusoy et al., 2025
Fukuda et al., 2026
Zhu et al., 2021
Choi et al., 2022
Sukegawa et al., 2022
Soltani et al., 2024
Fernandes et al., 2025
D1: Participants
D2: Index test
D3: Target condition
Judgement
High
Low

d)


Participants
Index test
Target condition
Overall
0%
25%
50%
75%
100%
No information
Critical
High
Unclear
Low

277

Figure 3. Paired forest plot of the observed study-level sensitivity and specificity, each with an exact 95% confidence interval, for artificial intelligence on panoramic radiographs detecting M3M–IAN contact. The values are the reported data and do not depend on the synthesis model. Studies follow the order of Table 1.

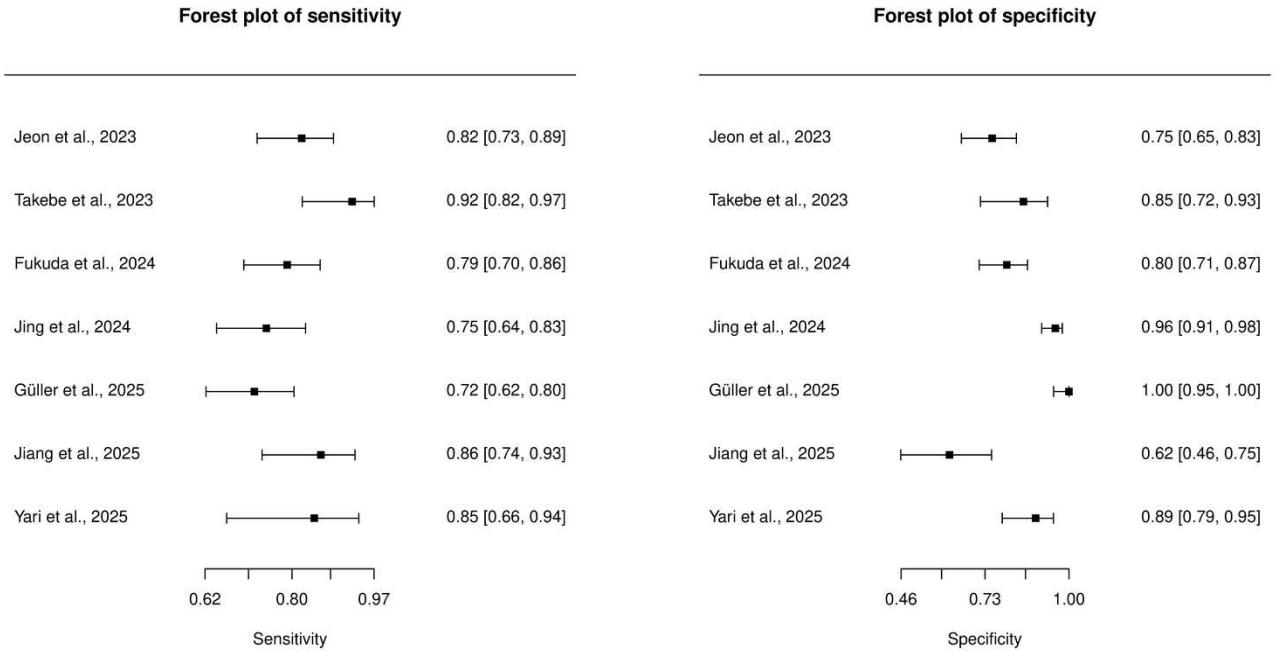


Figure 4. Bivariate/HSROC summary receiver operating characteristic (SROC) curve for FQ1 from the Bayesian bivariate model, showing the seven study operating points, the summary operating point, the 95% credible region (shaded) and the 95% posterior predictive region (dashed), produced with MetaBayesDTA.

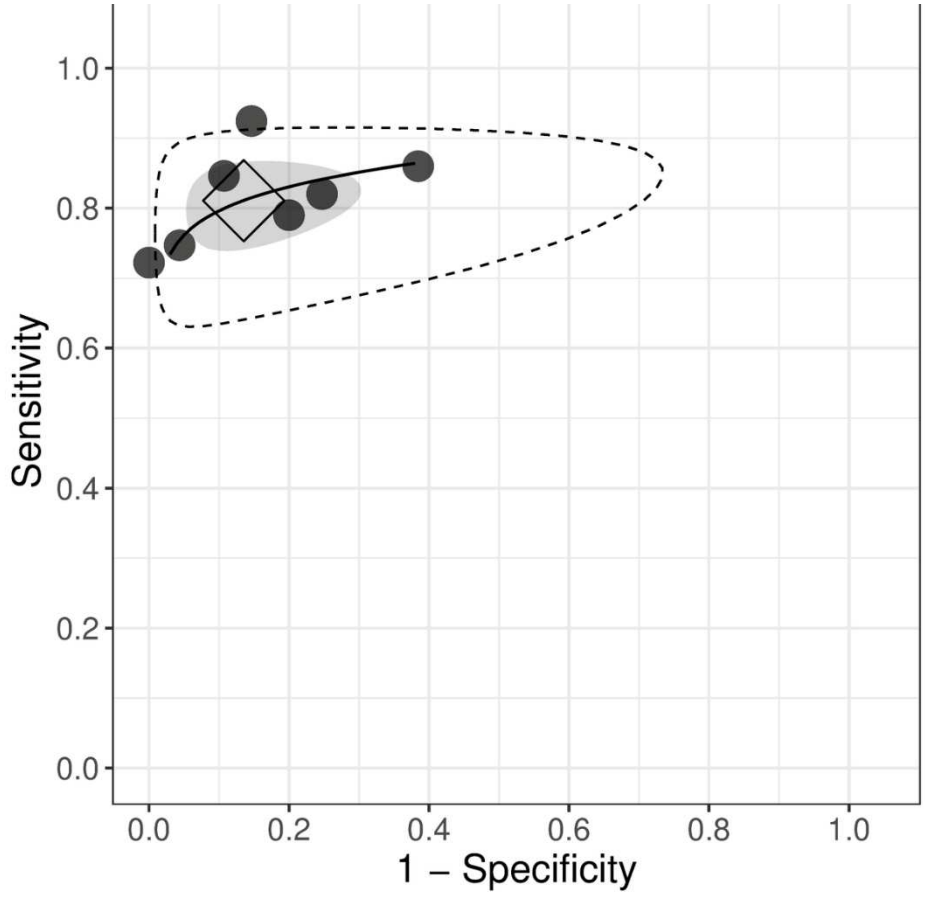


Figure 5. Summary of findings (diagnostic test accuracy) for FQ1, reporting the summary sensitivity and specificity and the expected numbers of true positives, false negatives, true negatives and false positives per 1000 patients tested at three plausible pre-test probabilities of contact (10%, unselected screening population; 30%, referred pre-extraction population, base case; 50%, selected referral with panoramic risk signs), with the GRADE certainty of evidence rated separately for sensitivity and specificity (GRADEpro GDT). The summary estimates are posterior medians from the Bayesian bivariate model and the intervals, labelled CI by the software, are 95% credible intervals. GRADEpro truncates the interval bounds to two decimals, so the figure shows 0.74 to 0.87 for sensitivity and 0.69 to 0.94 for specificity where the exact posterior intervals of the primary model are 0.746 to 0.871 and 0.695 to 0.945.

Should AI on panoramic radiography (OPG) be used to diagnose direct M3M-IAN contact in adults undergoing mandibular third molar assessment?

Patient or population: adults undergoing mandibular third molar assessment
Setting: Pre-extraction assessment of the mandibular third molar (M3M-IAN relationship)
New test: AI on panoramic radiography | **Cut-off value**:
Reference test: CBCT / 3D imaging | **Threshold**:
Pooled sensitivity:0.81 (95% CI: 0.74 to 0.87)| **Pooled specificity**:0.87 (95% CI: 0.69 to 0.94)

Test result	Number of results per 1000 patients tested (95% CI)			Number of participants (studies)	Certainty of the Evidence (GRADE)
	Prevalence 10% Broad, unselected population of impacted mandibular third molars (screening setting)	**Prevalence 30%** Referred pre-extraction population with panoramic M3M-IAC overlap (base case)	**Prevalence 50%** Selected referral with panoramic risk signs (plausible upper bound)		
True positives	**81** (75 to 87)	**243** (224 to 261)	**406** (373 to 436)	1050 (7)	⨁◯◯◯ **Very low**[a,b,c]
False negatives	**19** (13 to 25)	**57** (39 to 76)	**94** (64 to 127)		
True negatives	**780** (625 to 850)	**607** (486 to 662)	**434** (347 to 473)	1050 (7)	⨁◯◯◯ **Very low**[a,b,c,d]
False positives	**120** (50 to 275)	**93** (38 to 214)	**66** (27 to 153)		

CI: confidence interval

Explanations

a. All 7 studies at high risk of bias (QUADAS-3): retrospective, non-consecutive sampling, with the test set balanced by construction in several (Participants domain); uncorrected per-tooth clustering, internal/cross-validation without an external test set, post-hoc threshold optimisation, and some 2x2 tables reconstructed by inference (Analysis domain). Bias tends to overestimate accuracy.
b. Heterogeneous AI architectures (detection CNN, classification CNN, segmentation+classifier); predominantly single-centre and retrospective datasets (some public/overlapping); no model evaluated as a triage tool actually deployed in the clinical pathway.
c. Strong potential for selective reporting / positive-result bias typical of the diagnostic AI literature; too few studies for a reliable funnel-plot test.
d. Marked, unexplained heterogeneity of specificity (per-study 0.62-1.00; very wide 95% prediction region on the SROC).

3.2 Performance of AI models vs human readers of FQ1

Nine panoramic-index studies compared the AI with human readers of the same radiographs [26,31,35–40,42]. Because the comparison was reported in different metrics (accuracy, AUROC, or paired sensitivity and specificity), the paired AI-versus-human data are tabulated in Table 2 and shown in Fig. 6; no study reported the readers as more accurate overall than the AI.

Where the metric was accuracy or discrimination, the AI was ahead. Choi et al. [26] was more accurate than six oral and maxillofacial specialists at calling true contact (72.3% versus 52.7–69.6%), though its overall test-set accuracy for contact was lower (0.63); in Jiang et al. [38] the AUROC of all five experienced dentists fell below the 95% CI of the deep-learning model (0.688 versus 0.811); and Fukuda et al. [36] placed its best convolutional network (ResNet18, AUC 0.83) above human observers (0.67–0.80). The companion study Fukuda et al. [37] showed the same pattern, its model reaching an AUROC of 0.85 that matched two experienced radiologists (0.81) and exceeded two residents (0.55).

The studies reporting paired reader sensitivity and specificity located the difference in specificity. In Ulusoy et al. [35] the two radiologists were as sensitive as the model (0.86 and 0.93) but far less specific (about 0.50), over-calling contact, whereas the model held a similar sensitivity (0.86) at a specificity of 0.82; this comparison is not paired, however, because the readers assessed the whole dataset of 1478 images and the model a 20% partition of it, so it is not plotted alongside the others. Takebe et al. [31] shows the same pattern more starkly: on the same 94 images its five oral surgeons averaged a sensitivity of 0.83 at a specificity of 0.37, with individual specificities from 0.15 to 0.56, against 0.92 and 0.85 for the model. Fukuda et al., 2024 [37] sits between the two, its radiologists reaching a sensitivity of 0.80 at a specificity of 0.69 against 0.79 and 0.80 for the model, and its residents falling well below both (0.55 and 0.59). Yari et al. is the exception: its three oral and maxillofacial specialists matched the model almost exactly (sensitivity 0.85 versus 0.85, specificity 0.85 versus 0.89), while three newly graduated dentists fell below both (0.74 and 0.74). Jing et al. [40] is the one study whose readers erred in the opposite direction: the consensus reading of two oral and maxillofacial surgeons was less sensitive

than the model (0.72 versus 0.75) and also less specific (0.89 versus 0.96), so the readers under-called rather than over-called contact.

On the ROC plane (Fig. 6) the readers therefore lie to the high-false-positive-rate side of their own study's AI operating point in Takebe et al.[31] and Fukuda et al., 2024[37], coincide with it in Yari et al. [39], and fall below it on both axes in Jing et al. [40]. The same study also compared its model with dentists on a separate cohort of 100 high-risk third molars, where the model again had the higher area under the curve, and in Zhu et al. [42] the model was numerically but not statistically superior to five dentists. All comparisons were internal to single studies and used small reader panels of uneven experience, with the clinicians confined to the panoramic image.

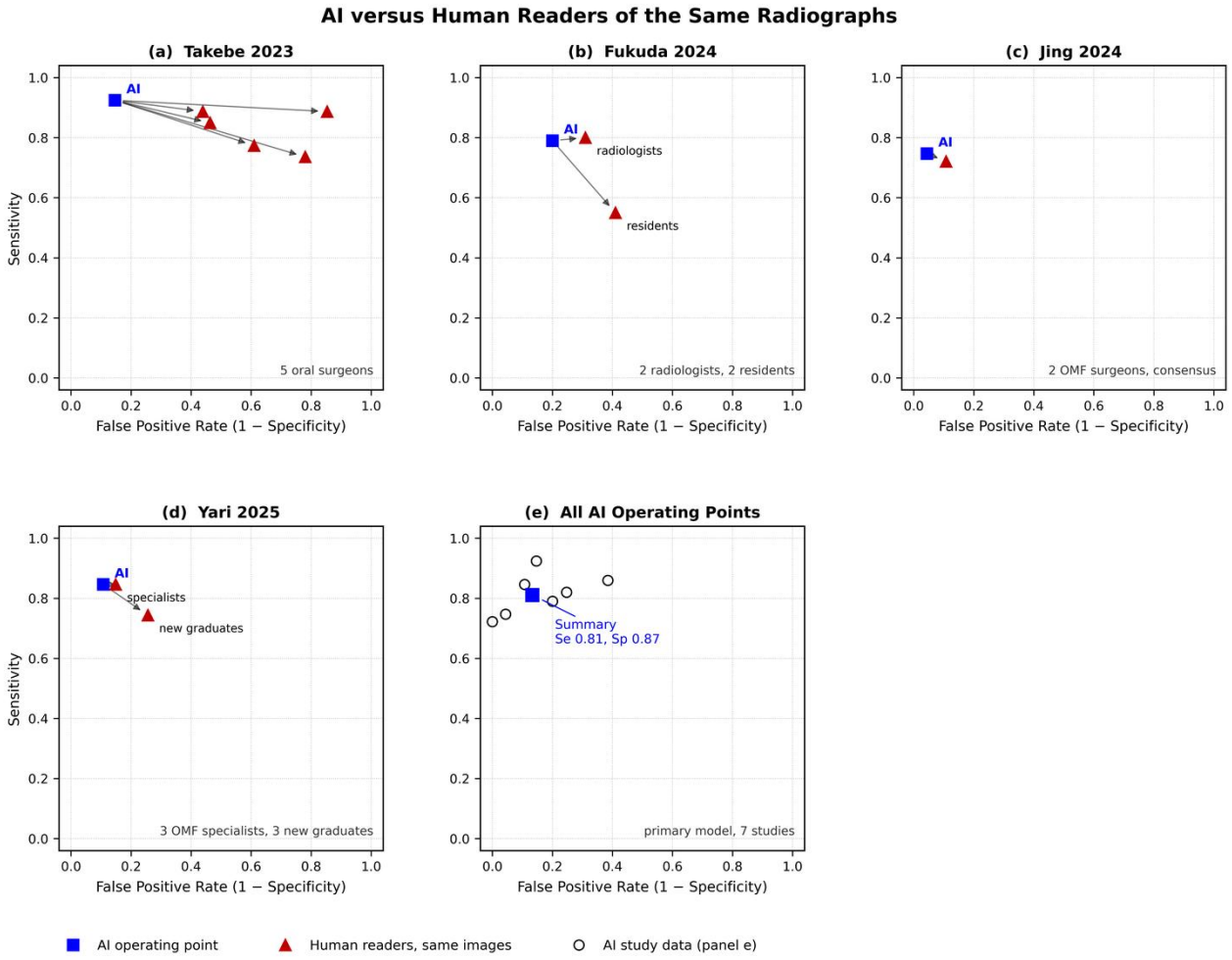
Table 2. Within-study comparison of artificial intelligence against human readers of the same panoramic radiographs (FQ1). For each of the nine head-to-head studies the table reports the human reader panel (number and expertise), the comparison metric, the human and AI results on the same test data, and the direction of the difference. Studies follow the order of Table 1. In Ulusoy et al. the readers assessed all 1478 images while the model was evaluated on a 20% partition, so that comparison is not paired on the same test data and the study is not plotted in Fig. 6.

Study	Human reader (n; expertise)	Metric	Human result	AI result (same test set)	Direction
Takebe et al., 2023 [31]	5 oral surgeons (≥1 y in M3 surgery)	Accuracy; Se / Sp	Acc 0.63; Se 0.83 / Sp 0.37 (means; per-reader Sp 0.15–0.56)	Acc 0.89; Se 0.92 / Sp 0.85	AI higher on all three
Fukuda et al., 2024 [37]	2 radiologists + 2 residents	AUROC; Se / Sp	Radiologists AUROC 0.81, Se 0.80 / Sp 0.69; residents AUROC 0.55, Se 0.55 / Sp 0.59	AUROC 0.85; Se 0.79 / Sp 0.80	AI ≈ radiologists; AI > residents
Jing et al., 2024 [40]	2 OMF surgeons (consensus, >20-y expert adjudication)	Accuracy; Se / Sp	Acc 0.75; Se 0.72 / Sp 0.89	Acc 0.89; Se 0.75 / Sp 0.96	AI higher on all three

Jiang et al., 2025 [38]	5 expert dentists	AUROC	0.688 (all below AI 95% CI)	0.811	AI higher
Yari et al., 2025 [39]	3 OMF specialists; 3 newly graduated dentists	Accuracy; Se / Sp	Specialists acc 0.85, Se 0.85 / Sp 0.85; novices acc 0.74, Se 0.74 / Sp 0.74	Acc 0.88; Se 0.85 / Sp 0.89	AI ≈ specialists; AI > novices
Ulusoy et al., 2025 [35]	2 radiologists	Se / Sp	Se 0.86 & 0.93 / Sp ~0.50	Se 0.86 / Sp 0.82 (best ANN)	Comparable Se; AI markedly more specific
Fukuda et al., 2026 [36]	2 radiologists + 1 oral surgery resident	AUC	0.67–0.80	0.83	AI (best CNN) ≈ radiologists; > resident
Zhu et al., 2021 [42]	5 dentists (1–3 y)	Average precision	AP 0.76	Numerically higher	Comparable (not significant)
Choi et al., 2022 [26]	6 OMF specialists (mean 15.3 y)	Accuracy (true contact)	0.53–0.70	0.72 (reader subset); 0.63 (test set)	AI higher

Figure 6. Summary receiver operating characteristic (SROC) plot comparing artificial intelligence with human readers for detecting M3M–IAN contact on panoramic radiographs (FQ1), plotted as sensitivity against the false-positive rate (1 − specificity). Panels (a) to (d) hold one head-to-head study each, in the order of Table 1, so that every AI-to-reader comparison is read within a single panel: the blue square is that study's AI operating

point, the red triangles are its human readers on the same images. The four studies are those that report, or allow the exact derivation of, a paired sensitivity and specificity on a test set common to model and readers. Ulusoy et al. is not shown as a paired comparison, because its readers assessed all 1478 images while the model was evaluated on a 20% partition, so the two are not measured on the same data. Panel (e) carries the seven AI operating points of the primary panoramic-index meta-analysis (open circles) and their bivariate summary (blue square). The paired AI-versus-human data are tabulated in Table 2.



In all four studies the advantage is in specificity (+0.06 to +0.48), not sensitivity (+0.03 to +0.11); gaps are AI minus the mean of the reader groups within each study

3.3 Comparison between AI architectures used in FQ1

Five studies trained more than one architecture on the same dataset, so architectures could be compared within studies [9,34–36,38]. No backbone was uniformly best. A ResNet backbone was the strongest convolutional choice in two: ResNet50 in Jiang et al. [38] (accuracy 0.75, AUROC 0.811), ahead of the VGG, ResNet101 and EfficientNet models, and ResNet18 in Fukuda et al. [36] (AUC 0.83), ahead of EfficientNet-B0 (0.77) and Inception-v3 (0.70). In Jeon et al. [9], by contrast, an EfficientNet-backbone detector (EfficientDet-D4) gave the best accuracy (0.79) ahead of YOLOv3 (0.73), while the ResNet-backbone detector (RetinaNet) had the lowest accuracy (0.60) at the highest specificity (0.89); and in Güller et al. [34] Inception-v3 led GoogLeNet and LeNet (accuracy 0.85, 0.77 and 0.70). Feature-based machine learning was competitive with or better than the convolutional networks wherever the two were compared directly: random-forest and gradient-boosted models reached AUCs of 0.85 and 0.84 in Fukuda et al. [36], at or above the best network, and in Ulusoy et al. [35] the convolutional networks

were the most sensitive but least specific (as low as 0.23), whereas the feature-based neural networks were better balanced. Within a study, the ranking also shifted with the operating point, the same model being the most sensitive or the most specific according to its threshold.

3.4 Prediction of inferior alveolar nerve injury by AI on radiographic images for FQ2

Six studies addressed prediction of postoperative IAN injury or paraesthesia (Table 3), and all six were appraised with PROBAST+AI (per-model judgements in Supplementary Table 5; risk-of-bias traffic-light plot in Fig. 7) [16]. Four predicted a specific neurosensory outcome: three targeted inferior alveolar nerve injury or paraesthesia specifically, while Feher et al. [43] predicted post-traumatic trigeminal neuropathy, combining inferior alveolar, lingual and buccal nerve involvement rather than being IAN-specific [43–46]; two predicted a broader composite of postoperative complications in which paraesthesia was one component but without accuracy metrics specific to nerve injury [47,48]. The models split into CNNs applied to images and machine learning on clinical or radiographic features. Reported discrimination varied widely and was not comparable across studies; it ranged from an AUC of 0.92 with sensitivity 0.84 and specificity 0.82 in Kim et al. [45] to an AUC of 0.63 in the small within-patient study of Picoli et al. [44]. Validation was internal in every study, none reported external validation, calibration or net-benefit analysis, and event counts were often small. All six studies were at high overall risk of bias: the analysis domain was the common driver, and the outcome domain was high in three studies: the two composite-outcome models, whose follow-up is too short to separate transient from permanent injury, and the within-patient study of Picoli et al., whose selection fixes the event rate by design. In Jain et al. [48] the predictor domain was high as well. Certainty was correspondingly very low.

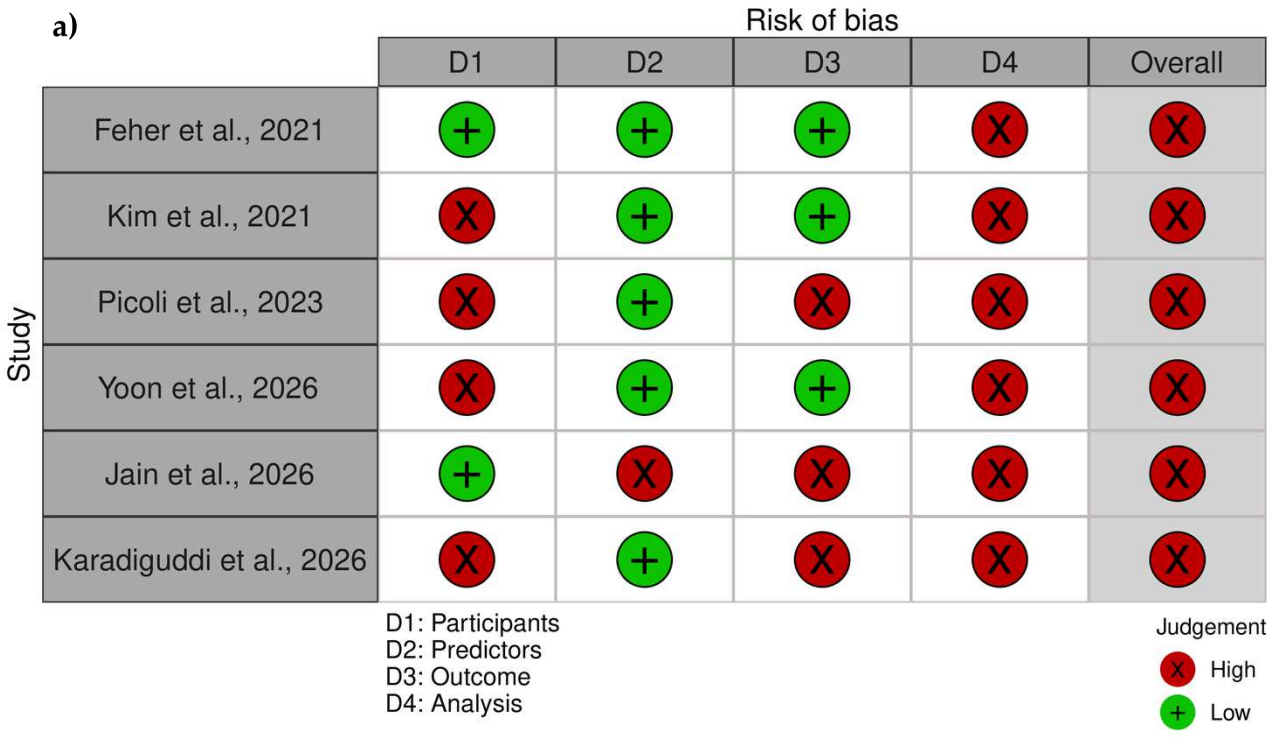
Table 3. Characteristics of the six studies addressing FQ2. For each study the table reports first author and year, country, study design, the sample as patients (M3M units) with the number of events and their prevalence, the outcome definition, the model and its predictors, the validation strategy, discrimination (AUC or c-statistic), and sensitivity and specificity where reported. Studies are ordered by outcome definition, the four with a specific neurosensory outcome first, then by year. Validation is internal in every study; a dash marks a value not reported. Footnotes: [a] The report describes a within-patient controlled design nested in a prospective cohort, but cases were selected based on a known postoperative injury, so data collection for this analysis was retrospective. [b] 50:50 hold-out with nested five-fold cross-validation for tuning. [c] The outcome combines inferior alveolar, lingual and buccal nerve involvement. [d] Pain, swelling and paraesthesia combined, assessed at days 7–14. [e] Assessed at days 3 and 7; calibration and decision-curve analysis not reported. [f] The report describes its design both as cross-sectional and as prospective observational, and the sample is purposive rather than consecutive.

Study	Country	Design	N (events; prevalence)	Outcome	Model	Validation	AUC / c-stat	Se / Sp

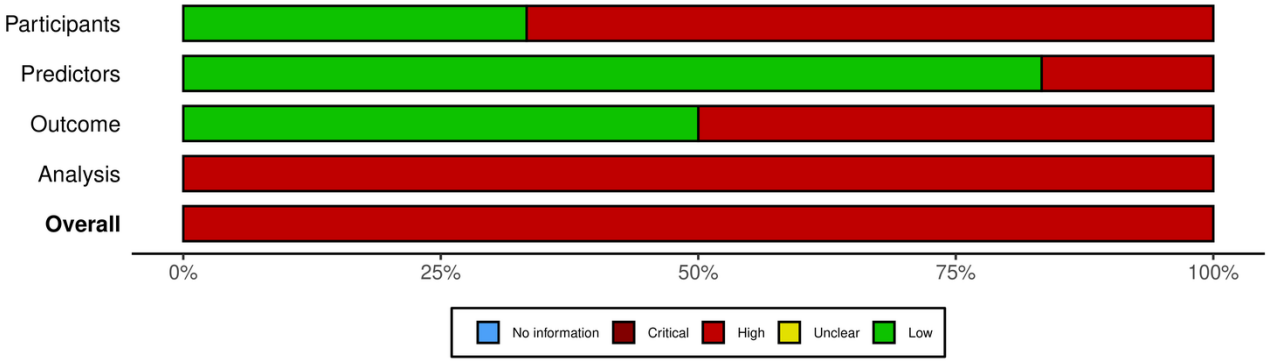
—	Se 0.84 / Sp 0.82	Se 0.87 / Sp 0.39	—
0.75	0.92	0.63 (3D-AI); 0.58 (CBCT)	Not reported; F1-score primary (0.92 CBCT, 0.67 PR)
Internal (k-fold CV)	Internal (hold-out + k-fold CV)[b]	Internal (hold-out)	Internal (hold-out)
LASSO on clinical and radiographic features	CNN on imaging (PR)	CNN on imaging (3D-AI, PR, CBCT)	CNN on imaging (PR and CBCT)
Post-traumatic trigeminal neuropathy, not IAN-specific[c]	IAN paraesthesia after M3M extraction	Postoperative unilateral IAN injury, transient (≤ 10 days)	Postoperative IAN injury
648 patients (812 M3M); 14 events (2%)	300 patients (300 images); 100 events (33%)	25 patients (50 M3M); 25 events (50%)	902 PR and 1500 CBCT images; 162 events (18%)
Retrospective	Retrospective	Retrospective[a]	Retrospective
Austria	South Korea	Belgium	South Korea
Feher et al., 2021 [43]	Kim et al., 2021 [45]	Picoli et al., 2023 [44]	Yoon et al., 2026 [46]

Jain et al., 2026 [48]	India	Retrospective	472 patients; more than 200 events	Eight postoperative complications including paraesthesia, each	Five ML models (logistic, RF, XGBoost, SVM, naïve Bayes) with SHAP	0.91 (RF); 0.89 (XGBoost)	Se 0.84 / Sp 0.87 (RF); Se 0.82 / Sp 0.85 (XGBoost)
Karadiguddi et al., 2026 [47]	India	Prospective[f]	40 patients; events not reported	Composite postoperative morbidity including paraesthesia, not IAN-specific[d]	Random forest and CNN on CBCT and clinical features	0.88 (CNN, composite morbidity)	Se 0.81 / Sp 0.84

Figure 7. PROBAST+AI risk-of-bias and applicability assessment of the six FQ2 prognostic studies (prediction of postoperative inferior alveolar nerve injury). (a) Per-study risk-of-bias traffic-light plot and (b) risk-of-bias summary bar plot; (c) per-study applicability traffic-light plot and (d) applicability summary bar plot. Studies follow the order of Table 3. Risk-of-bias domains: D1 participants, D2 predictors, D3 outcome, D4 analysis; applicability domains: participants, predictors, outcome. Green = low, red = high concern/risk.



b)



392

c)

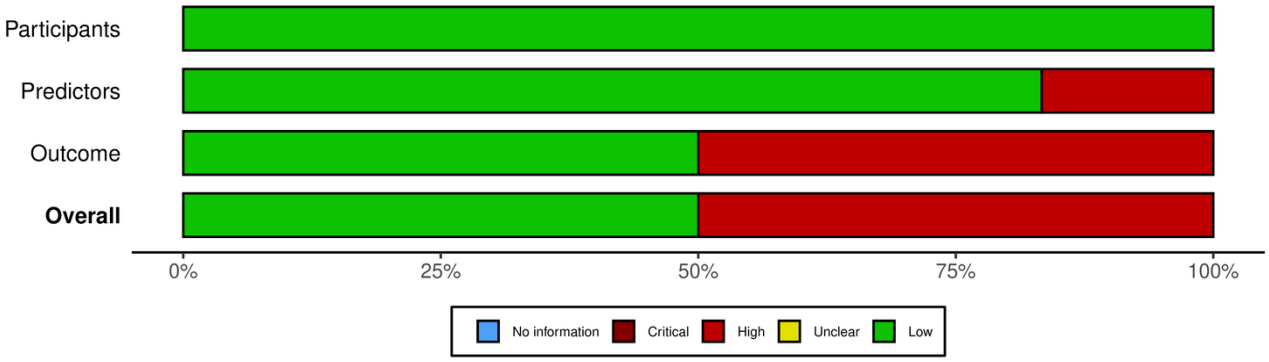
Concerns regarding applicability

	D1	D2	D3	Overall
Feher et al., 2021	+	+	+	+
Kim et al., 2021	+	+	+	+
Picoli et al., 2023	+	+	X	X
Yoon et al., 2026	+	+	+	+
Jain et al., 2026	+	X	X	X
Karadiguddi et al., 2026	+	+	X	X

D1: Participants
D2: Predictors
D3: Outcome

Judgement
X High
\+ Low

d)



4. Discussion

AI applied to panoramic radiographs detected M3M–IAN contact with a pooled sensitivity of 0.81 (95% CrI 0.75–0.87) and a comparable specificity of 0.87 (95% CrI 0.70–0.95) that ranged from 0.62 to 1.00 across studies. Both estimates were stable across prespecified sensitivity analyses and showed no relationship with publication year, and specificity remained the larger source of heterogeneity. The prognostic models that predict nerve injury after extraction were few, methodologically diverse and weaker. The certainty of evidence was very low for every outcome.

Our summary accuracy is close to the only previous meta-analysis, which pooled seven deep-learning studies to a sensitivity of 80.2% and specificity of 85.8% [13]. While the number of included studies may seem similar, the present review considered different papers, adopted diverse methodology and reversed the quality judgements due to a different appraisal tool. That review, and the companion narrative review of nerve-injury and difficulty prediction by the same group [14], rated the overall risk of bias as low using QUADAS-2 and did not grade the certainty of the evidence. Applying QUADAS-3, we instead judged every one of the 15 studies to be at high risk of bias, driven by the participants and analysis domains reported above. For the six studies common to both reviews, the verdict flips because QUADAS-3 adds a dedicated analysis domain and a case-mix signal that QUADAS-2 does not contain, not merely because the two teams disagreed. These are the conditions under which internally validated, single-centre models report their best numbers, so a low-risk verdict on this literature is more likely to reflect a less balanced tool than a genuinely low risk.

Across the seven studies in the primary model, architecture family did not explain the heterogeneity. Classification CNNs spanned a specificity range of 0.62 to 1.00 and detection networks 0.75 to 0.85, so no family was reliably more accurate across studies. The within-study bake-offs provide the only unconfounded read: a ResNet backbone was strongest in some studies but not others (an EfficientNet-backbone detector won in Jeon et al. [9]), feature-engineered models matched or beat the convolutional networks, and the winner shifted with the operating metric, so the evidence does not single out an architecture to build on; with seven studies we cannot exclude an architecture effect, but neither the between-study nor the within-study signal identifies one, and the binding constraints are data quality, external validation and calibration rather than the network. The heterogeneity tracked the data and the operating point instead. Sensitivity and specificity traded off against each other, the expected consequence of different decision thresholds (Spearman $\rho = -0.61$ between the two across the seven studies, exact permutation $p = 0.17$, so the pattern is descriptive rather than established). The two ends of the specificity range make the mechanism concrete. Güller et al. [34], the only study reaching a specificity of 1.00, called contact in 40% of its test images against a true contact prevalence of 55%, and bought the absence of false positives at the cost of the lowest sensitivity in the pool (0.72); Jiang et al. [38] did the reverse, calling contact in 65% of a test set of 56% prevalence to reach a sensitivity of 0.86 at a specificity of 0.62. Neither is the more accurate model: they are one trade-off read at two operating points, and a specificity estimated on 74 negatives remains compatible with values as low as 0.95 even when no false positive is observed. Since no study reported its full operating characteristic, these positions cannot be placed on a common curve, and the headline specificity of the most specific model is a property of where its threshold was set as much as of what it learned. The variation in specificity could not be attributed to any measured covariate. Architecture family and publication year both failed to track it, and contact prevalence could not be tested at all: across the pooled studies it spans only 0.32 to 0.56, too narrow a range of case-mix for a coefficient to carry information. Headline specificity therefore appears shaped by each study's design and thresholding rather than by any single identifiable factor. Some of this variation is a spectrum effect: each estimate comes from a narrow, artificially

enriched case-mix that need not transfer to the consecutive clinical population, and none of the studies reported the stratified estimates that would expose it [49]. These observations are exploratory: they rest on seven studies, are sensitive to a few datasets at the edges of the specificity range, and none would survive a formal test, so we present them as hypotheses for primary studies to test rather than as established effects. Validation design could not be separated from case-mix in this sample; what all but one study shared, whether it used a hold-out set or cross-validation, was the absence of an external test set, and that, not the choice of resampling, is the relevant limitation. The present review had too few studies to model temporal trends reliably, so the flat publication-year covariate is uninformative rather than evidence that accuracy has not improved.

These problems are not specific to dentistry [50]; they are the familiar failure modes of medical-AI accuracy studies that the current appraisal frameworks were written to expose, with their emphasis on external validation, calibration and clinical utility [16,51]. We excluded studies that applied AI directly to CBCT: because their index test is the same 3D modality used as the reference standard, they automate the reading of CBCT rather than decide who needs it, a different clinical question from the panoramic triage this review addresses.

For clinical practice, the intended role of AI on panoramic radiographs is triage before CBCT, and the following calculation sets the bar. At a contact prevalence of 30%, the summary estimates correspond to about 57 missed contacts and 93 unnecessary CBCT referrals per 1000 patients tested, although these figures are illustrative: the predictive values depend on a prevalence that varies widely with case-mix, and the summary specificity is itself drawn from artificially enriched datasets, so the false-positive count in particular is approximate. The two errors are not equivalent. A missed contact removes the warning that would have prompted 3D imaging, a change of surgical plan, or the offer of a coronectomy, and can end in a permanent and litigated nerve injury, whereas an unnecessary CBCT is a low-dose scan and a modest cost. For triage the decision threshold should therefore favour sensitivity, accepting more unnecessary CBCT as the price of fewer misses, so a tool that still misses roughly one in five true contacts at this operating point is not yet adequate. The deployment model matters as much as the accuracy: an AI allowed to withhold CBCT when it calls "no contact" makes every false negative dangerous, whereas an AI restricted to escalating borderline cases, never to veto imaging the clinician already wants, carries almost none of that risk and is the defensible near-term use. AI must also earn its place against the current practice, because the established radiographic warning signs [7] already provide a free, high-sensitivity triage rule that sends sign-positive teeth to CBCT; a sensitivity of 0.81 is not self-evidently better, and only a head-to-head comparison would show whether AI adds to those signs or merely restates them.

Where studies did pit AI against clinicians reading the same images, the AI was as accurate or better, experienced specialists included, mostly by being more specific: in two of the four studies that place readers and model on the same ROC plane, the clinicians over-called contact, reaching a specificity between 0.37 and 0.69 against 0.80 to 0.85 for the model at broadly comparable sensitivity (Fig. 6). In Yari et al. [39] the three oral and maxillofacial specialists matched the model on both axes and only the newly graduated dentists fell behind, and in Fukuda et al., 2024 [37] the gap likewise opened against residents rather than radiologists, so the margin narrows as reader experience rises and part of what these comparisons measure is inexperience rather than a ceiling on expert reading. On current evidence, AI is a potential decision aid used additively, not a gatekeeper, and should not change practice yet.

Two deeper limits bound what any of these accuracy figures can mean, and both point to the same missing evidence. The reference standard is itself imperfect: contact was defined by CBCT or CT in every study, with two adding intraoperative confirmation [9,31], yet three-dimensional imaging infers rather than observes the true cortical contact and nerve exposure that only surgery reveals, so the reported sensitivity and specificity are benchmarked against an imperfect label that both lowers the ceiling on measurable accuracy and biases the

estimates by an uncertain amount and direction [52]. The other limit is more basic: accuracy is only the second of six ascending levels of diagnostic efficacy (technical quality, accuracy, diagnostic thinking, therapeutic choice, patient outcome and societal cost), where efficacy at one level is necessary but not sufficient for the level above [53]. This literature sits entirely at level two: no study tested whether an AI reading changed the clinician's decision, the operation performed, or the rate of nerve injury, and the one randomised trial that links the imaging step to outcomes found that CBCT-guided planning did not consistently reduce postoperative deficits over panoramic-based planning [12]. A triage tool must ultimately be judged by whether the imaging it directs changes management and prevents injury, and validating the index test against those downstream events rather than against another image is the more informative next step.

The path to useful evidence is concrete. Diagnostic studies should enrol consecutive patients at the prevalence seen in practice rather than artificially balanced datasets, analyse at the patient level with correction for multiple teeth, fix the decision threshold in advance, and validate externally on data from other centres and devices; and, building on the within-study reader comparisons already available, test AI prospectively against clinicians of defined experience and against the formal radiographic warning signs, to show whether it adds to rather than restates them. Prognostic models need larger cohorts with enough nerve-injury events, a clearly defined and clinically meaningful outcome that separates transient from permanent deficit, and reporting of calibration and net benefit rather than discrimination alone.

Inferior alveolar nerve injury is among the most feared and most litigated complications of third-molar surgery, and the appeal of reading panoramic radiographs with AI is real: panoramic imaging is inexpensive and near-universal, whereas CBCT costs more, is less available and delivers a higher radiation dose [11], so a dependable triage step could concentrate three-dimensional imaging on the patients who need it and extend structured risk assessment to settings that lack CBCT altogether. That promise is also where the difficulties concentrate. The models were trained and tested almost entirely on single-centre datasets from a few countries, so their accuracy may not transfer across populations, scanners and acquisition protocols. A tool with imperfect specificity over-refers, turning a software output into added radiation, cost and opportunity cost that press hardest on resource-limited services. The sensitivity–specificity trade-off is not a tuning detail but an encoded value judgement about how many unnecessary scans are acceptable to prevent one nerve injury, a choice that belongs to patients and professional bodies rather than to a default threshold. Human factors compound this: an AI that returns "no contact" invites automation bias and, over time, the erosion of the clinician's own skill at reading the established warning signs, most acutely in trainees.

Finally, one should note that the present systematic review and meta-analysis has some limitations. Although the major databases were searched and performed forward and backward citation searching, some studies may have been missed. Several FQ1 contingency tables were reconstructed rather than read directly. Excluding them did not change the conclusions, and entering the distinction as a covariate gave a difference in summary sensitivity of 0.05 (95% credible interval −0.08 to 0.17) and in specificity of −0.05 (−0.40 to 0.21); the reconstructed tables are therefore not demonstrably less reliable, though a difference of up to about 17 percentage points in sensitivity cannot be excluded. Two of the pooled tables deserve naming, because neither is a single observed confusion matrix: Jiang et al.[38] reports metrics averaged over five model initialisations of a fixed test set, and Fukuda et al., 2024[37] a table that aggregates two training and testing runs over the same 100 test patches, so its nominal 200 observations are 100 images counted twice. We therefore added a post-hoc sensitivity analysis halving its cells, which moved the summary by less than half a point on either axis, but the study nonetheless enters the primary pool at twice its independent information content. Rule (v) also decided the panoramic studies kept out of the primary pool. Kumbasar et al. [33] aggregates one confusion matrix over

ten cross-validation folds spanning its entire cohort of 546 units; Ulusoy et al. [35] averages its metrics over thirty random re-splits drawn, as its Methods state, irrespective of contact status. In neither case does a single held-out table with a known class composition exist, and in the latter the published accuracy, positive and negative predictive values imply mutually incompatible class ratios, which is the expected arithmetic consequence of averaging non-linear ratios across repeated splits rather than any defect of the source. Kumbasar et al. still enters a sensitivity analysis, in place of the report with which it shares a cohort; Ulusoy et al. shares a cohort with no other included study and therefore enters none. The estimates themselves are stable across every prespecified sensitivity analysis, but the model is weakly identified: fitted by maximum likelihood it does not converge on these seven studies, and only the Bayesian fit, with weakly informative priors on the between-study variances, returns a proper covariance (Supplementary Note 2). What seven studies cannot pin down is how widely accuracy varies from one dataset to the next, and that is why we report the prediction region alongside the summary point. Because the primary studies analysed at the level of the individual tooth or side without correcting for the clustering of multiple teeth within a patient, the pooled credible region may be narrower than the true between-patient uncertainty, and the summary point itself may be biased where multiple teeth within a patient share difficulty. The pattern analysis rests on seven studies and is exploratory, so the associations we describe are hypotheses for primary studies to test, not established effects. The prognostic evidence was too sparse and diverse for quantitative synthesis, and was reported narratively.

Overall, AI applied to panoramic radiographs detects the M3M–IAN relationship with encouraging pooled accuracy, but the evidence is uniformly at high risk of bias, the specificity varies across datasets, only one model has been validated externally and none prospectively, and the certainty is very low; prognostic models for nerve injury are more preliminary still. AI is therefore a potential decision aid rather than a gatekeeper, and is not yet ready for the chairside. What the field needs next is not another accuracy test but studies designed to show clinical usefulness, along the lines set out above.

Data availability

All data supporting this paper are aggregated data extracted from the included primary studies and are provided within the paper itself and its supplementary materials; no new individual-participant data were generated. The review protocol is registered in PROSPERO (CRD420251000827).

Code availability

The bivariate model, the sensitivity analyses and the meta-regressions were fitted in MetaBayesDTA; the maximum-likelihood fit attempted in MetaDTA did not converge. Both are web applications requiring no custom code. The Python scripts used to generate Fig. 6 and to compute the threshold trade-off analyses reported in the Discussion are not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author.

Competing interests

All authors declare no financial or non-financial competing interests.

Authors' contributions

P.M. (Pietro Montagna): Conceptualization, Methodology, Investigation (literature search, study screening, risk-of-bias assessment and adjudication of data-extraction disagreements), Validation, Data curation, writing original draft, writing review & editing, Project administration. R.N. (Riccardo Nocini): writing original draft, Supervision. P.F. (Paolo Faccioni): Validation, Data curation, writing original draft, writing review & editing. S.W. (Siwei Wang): Investigation. A.S. (Alice Sammarco): Investigation (study screening and risk-of-bias

assessment), writing original draft. F.M. (Federica Melloni): Formal analysis, Investigation (data extraction), Writing original draft. G.M. (Giulio Montagna): Software, Investigation (data extraction), Data curation, Visualization. G.C. (Giulio Conti): Software, Visualization, Writing review & editing. D.D.S. (Daniele De Santis): Conceptualization, Methodology, Investigation (adjudication of screening disagreements), Supervision, Project administration. All authors read and approved the final version of the manuscript.

Ethics declarations

Not applicable. This systematic review and meta-analysis used only previously published, aggregated data and did not involve new human participants, animals or identifiable personal data.

Founding

This research received no external funding.

Clinical trial registration

Clinical trial number: not applicable.

5. References

1. Pitros, P., O'Connor, N., Tryfonos, A. & Lopes, V. A systematic review of the complications of high-risk third molar removal and coronectomy: development of a decision tree model and preliminary health economic analysis to assist in treatment planning. *British Journal of Oral and Maxillofacial Surgery* **58**, e16–e24 (2020).
2. Putrino, A. *et al.* Inferior Alveolar Nerve Impairment Following Third-Molar Extraction: Management of Complications and Medicolegal Considerations. *JCM* **14**, 2349 (2025).
3. Cheung, L. K. *et al.* Incidence of neurosensory deficits and recovery after lower third molar surgery: a prospective clinical study of 4338 cases. *International Journal of Oral and Maxillofacial Surgery* **39**, 320–326 (2010).
4. Thereza-Bussolaro, C., Fagundes, N., Stefani, C. M., Flores-Mir, C. & Borba, A. M. Risk of Postsurgical Inferior Alveolar Nerve Injury in At-Risk Patients Based on Tomographic Imaging Features: Systematic Review and Meta-Analysis. *Journal of Oral and Maxillofacial Surgery* S0278239126002478 (2026) doi:10.1016/j.joms.2026.03.008.
5. Lamiae, H., Samir, M. C., Marouane, B. & Bouchra, T. Coronectomy of mandibular wisdom teeth: A case series. *International Journal of Surgery Case Reports* **90**, (2022).
6. Montagna, P. Minimally Invasive Techniques and Emerging Diagnostics in Mandibular Third Molar Surgery: A Narrative Review. *Journal of Applied Cosmetology* (2025).
7. Rood, J. P. & Nooraldeen Shehab, B. A. A. The radiological prediction of inferior alveolar nerve injury during third molar surgery. *British Journal of Oral and Maxillofacial Surgery* **28**, 20–25 (1990).
8. Kim, H. J. *et al.* Anatomical Risk Factors of Inferior Alveolar Nerve Injury Association with Surgical Extraction of Mandibular Third Molar in Korean Population. *Applied Sciences* **11**, 816 (2021).
9. Jeon, K. J., Choi, H., Lee, C. & Han, S.-S. Automatic diagnosis of true proximity between the mandibular canal and the third molar on panoramic radiographs using deep learning. *Scientific Reports* **13**, (2023).
10. Hasani, A., Ahmadi Moshtaghin, F., Roohi, P. & Rakhshan, V. Diagnostic value of cone beam computed tomography and panoramic radiography in predicting mandibular nerve exposure during third molar surgery. *International Journal of Oral and Maxillofacial Surgery* **46**, 230–235 (2017).

11. Petersen, L. B., Olsen, K. R., Christensen, J. & Wenzel, A. Image and surgery-related costs comparing cone beam CT and panoramic imaging before removal of impacted mandibular third molars. *Dentomaxillofacial Radiology* **43**, 20140001 (2014).
12. Guerrero, M. E., Botetano, R., Beltran, J., Horner, K. & Jacobs, R. Can preoperative imaging help to predict postoperative outcome after wisdom tooth removal? A randomized controlled trial using panoramic radiography versus cone-beam CT. *Clinical Oral Investigations* **18**, 335–342 (2013).
13. Al Salieti, H., Al Sliti, H. & Alkadi, S. Diagnostic performance of deep learning models in classifying mandibular third molar and mandibular canal contact status on panoramic radiographs: A systematic review and meta-analysis. *Imaging Science in Dentistry* **55**, 139 (2025).
14. Al Salieti, H. *et al.* Predicting alveolar nerve injury and the difficulty level of extraction impacted third molars: a systematic review of deep learning approaches. *Front. Dent. Med.* **6**, 1534406 (2025).
15. Whiting, P. F. *et al.* QUADAS-3: A Revised Tool for the Quality Assessment of Diagnostic Test Accuracy Studies. *Annals of Internal Medicine* **179**, 548–555 (2026).
16. Moons, K. G. M. *et al.* PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* **388**, e082505 (2025).
17. Schünemann, H. J. *et al.* Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* **336**, 1106–1110 (2008).
18. McInnes, M. D. F. *et al.* Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies. *JAMA* **319**, 388 (2018).
19. Page, M. J. *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **372**, n71 (2021).
20. Tomlinson, E. *et al.* Piloting QUADAS-3: a revised tool for the quality assessment of diagnostic accuracy studies. *Journal of Clinical Epidemiology* **188**, 111983 (2025).
21. Sounderajah, V. *et al.* A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nature Medicine* **27**, 1663–1665 (2021).
22. McGuinness, L. A. & Higgins, J. P. T. Risk-of-bias VISualization (robvis): An R package and Shiny web app for visualizing risk-of-bias assessments. *Research Synthesis Methods* **12**, 55–61 (2021).
23. Reitsma, J. B. *et al.* Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology* **58**, 982–990 (2005).

24. Rutter, C. M. & Gatsonis, C. A. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine* **20**, 2865–2884 (2001).
25. Sukegawa, S. *et al.* Deep learning model for analyzing the relationship between mandibular third molar and inferior alveolar nerve in panoramic radiography. *Sci Rep* **12**, 16925 (2022).
26. Choi, E. *et al.* Artificial intelligence in positioning between mandibular third molar and inferior alveolar nerve on panoramic radiography. *Sci Rep* **12**, 2456 (2022).
27. Campbell, M. *et al.* Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *BMJ* l6890 (2020) doi:10.1136/bmj.l6890.
28. Freeman, S. C. *et al.* Development of an interactive web-based tool to conduct and interrogate meta-analysis of diagnostic test accuracy studies: MetaDTA. *BMC Medical Research Methodology* **19**, (2019).
29. Cerullo, E. *et al.* MetaBayesDTA: codeless Bayesian meta-analysis of test accuracy, with or without a gold standard. *BMC Med Res Methodol* **23**, 127 (2023).
30. Deeks, J. J., Macaskill, P. & Irwig, L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal of Clinical Epidemiology* **58**, 882–893 (2005).
31. Takebe, K. *et al.* Deep learning model for the automated evaluation of contact between the lower third molar and inferior alveolar nerve on panoramic radiography. *Journal of Dental Sciences* **18**, 991–996 (2023).
32. Fernandes, F. A., Ge, M., Chaltikyan, G., Gerdes, M. W. & Omlin, C. W. Preparing for downstream tasks in artificial intelligence for dental radiology: a baseline performance comparison of deep learning models. *Dentomaxillofacial Radiology* **54**, 149–162 (2025).
33. Kumbasar, N., Güller, M. T., Miloğlu, Ö., Oral, E. A. & Ozbek, I. Y. Deep-learning based fusion of spatial relationship classification between mandibular third molar and inferior alveolar nerve using panoramic radiograph images. *Biomedical Signal Processing and Control* **100**, 107059 (2025).
34. Güller, M. T., Kumbasar, N. & Miloğlu, Ö. Evaluation of the effectiveness of panoramic radiography in impacted mandibular third molars on deep learning models developed with findings obtained with cone beam computed tomography. *Oral Radiol* **41**, 260–275 (2025).
35. Ulusoy, A. C., Toprak, T., Selver, M. A., Güneri, P. & İlhan, B. Panoramic radiographic features for machine learning based detection of mandibular third molar root and inferior alveolar canal contact. *Sci Rep* **15**, 4178 (2025).

36. Fukuda, M. *et al.* The optimal diagnostic assistance system for predicting three-dimensional contact between mandibular third molars and the mandibular canal on panoramic radiographs. *Oral Radiol* **42**, 322–332 (2026).

37. Fukuda, M. *et al.* Deep learning system to predict the three-dimensional contact status between the mandibular third molar and mandibular canal using panoramic radiographs. *Oral Science International* **21**, 46–53 (2024).

38. Jiang, Y., Sun, H.-T., Luo, Z., Wang, J. & Xiao, R.-P. Efficacy of a deep learning system for automatic analysis of the comprehensive spatial relationship between the mandibular third molar and inferior alveolar canal on panoramic radiographs. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology* **139**, 612–622 (2025).

39. Yari, A., Fasih, P., Nouralishahi, A., Mohammadikhah, M. & Nikeghbal, D. Performance of artificial intelligence versus clinicians on the detection of contact between mandibular third molar and inferior alveolar nerve. *Oral Science International* **22**, e1256 (2025).

40. Jing, Q. *et al.* Fully automated deep learning model for detecting proximity of mandibular third molar root to inferior alveolar canal using panoramic radiographs. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology* **137**, 671–678 (2024).

41. Soltani, P. *et al.* A two-stage deep-learning model for determination of the contact of mandibular third molars with the mandibular canal on panoramic radiographs. *BMC Oral Health* **24**, 1373 (2024).

42. Zhu, T., Chen, D., Wu, F., Zhu, F. & Zhu, H. Artificial Intelligence Model to Detect Real Contact Relationship between Mandibular Third Molars and Inferior Alveolar Nerve Based on Panoramic Radiographs. *Diagnostics (Basel)* **11**, 1664 (2021).

43. Feher, B. *et al.* Prediction of post-traumatic neuropathy following impacted mandibular third molar removal. *Journal of Dentistry* **115**, 103838 (2021).

44. Picoli, F. F. *et al.* Risk assessment of inferior alveolar nerve injury after wisdom tooth removal using 3D AI-driven models: A within-patient study. *Journal of Dentistry* **139**, 104765 (2023).

45. Kim, B. S. *et al.* Deep Learning-Based Prediction of Paresthesia after Third Molar Extraction: A Preliminary Study. *Diagnostics* **11**, 1572 (2021).

46. Yoon, K. *et al.* Single-step prediction of inferior alveolar nerve injury after mandibular third-molar extraction using contrastive learning and Bayesian auto-tuned deep learning model. *Dentomaxillofacial Radiology* **55**, 150–157 (2026).
47. Karadiguddi, P. *et al.* Artificial Intelligence in Predicting Surgical Problems and Postoperative Morbidity in Mandibular Third Molar Extractions. *Annals of African Medicine* **25**, 801–804 (2026).
48. Jain, A. & Jadhav, P. Machine learning models for predicting postoperative complications following mandibular third molar surgery: Development, validation, and explainable AI insights. *Journal of Cranio-Maxillofacial Surgery* **54**, 104449 (2026).
49. Mulherin, S. A. & Miller, W. C. Spectrum Bias or Spectrum Effect? Subgroup Variation in Diagnostic Test Evaluation. *Annals of Internal Medicine* **137**, 598–602 (2002).
50. Nocini, R. *et al.* Image-based Artificial Intelligence models in the diagnosis and classification of vascular anomalies of the soft tissue in the head and neck. *Computer Methods and Programs in Biomedicine* **259**, 108525 (2025).
51. Collins, G. S. *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* e078378 (2024) doi:10.1136/bmj-2023-078378.
52. Reitsma, J. B., Rutjes, A. W. S., Khan, K. S., Coomarasamy, A. & Bossuyt, P. M. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *Journal of Clinical Epidemiology* **62**, 797–806 (2009).
53. Fryback, D. G. & Thornbury, J. R. The Efficacy of Diagnostic Imaging. *Medical Decision Making* **11**, 88–94 (1991).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf