Supplementary Note 1

Electronic search strategy

Sources, dates and limits

Databases and registers searched (seven): MEDLINE (via PubMed), Embase (via Ovid), Scopus, Web of Science Core Collection, IEEE Xplore, the Cochrane Central Register of Controlled Trials (CENTRAL) and the WHO ICTRP trial registry. Google Scholar and manual reference checking were used as supplementary methods. Backward and forward citation searching was performed on all included studies.

Coverage window: 1 January 2014 to 26th June 2026. Language: no restriction. No study-design or reference-standard filter was applied to the strings, to maximise sensitivity as recommended for DTA reviews.

Concept blocks (PICO):

(1) third molar: third molar, wisdom tooth/teeth, mandibular/lower third molar, MeSH “Molar, Third”;

(2) IAN / mandibular canal: inferior alveolar nerve, alveolar nerve, mandibular nerve, mandibular canal, inferior alveolar canal, IAN, IAC, nerve injury, paraesthesia, neurosensory, MeSH “Mandibular Nerve”, “Trigeminal Nerve Injuries”, “Paresthesia”;

(3) artificial intelligence: artificial intelligence, machine learning, deep learning, neural network(s), convolutional, radiomics, automated detection, computer-aided, MeSH “Artificial Intelligence”.

Database strings

MEDLINE via PubMed

(("Molar, Third"[MeSH Terms] OR "third molar"[Title/Abstract] OR "wisdom tooth"[Title/Abstract] OR "wisdom teeth"[Title/Abstract] OR "mandibular third molar"[Title/Abstract] OR "lower third molar"[Title/Abstract]) AND ("Mandibular Nerve"[MeSH Terms] OR "Trigeminal Nerve Injuries"[MeSH Terms] OR "Paresthesia"[MeSH Terms] OR "inferior alveolar nerve"[Title/Abstract] OR "alveolar nerve"[Title/Abstract] OR "mandibular nerve"[Title/Abstract] OR "mandibular canal"[Title/Abstract] OR "inferior alveolar canal"[Title/Abstract] OR "IAN"[Title/Abstract] OR "IAC"[Title/Abstract] OR "nerve injury"[Title/Abstract] OR "paresthesia"[Title/Abstract] OR "neurosensory"[Title/Abstract]) AND ("Artificial Intelligence"[MeSH Terms] OR "artificial intelligence"[Title/Abstract] OR "machine learning"[Title/Abstract] OR "deep learning"[Title/Abstract] OR "neural network"[Title/Abstract] OR "neural networks"[Title/Abstract] OR "convolutional"[Title/Abstract] OR "radiomics"[Title/Abstract] OR "automated detection"[Title/Abstract] OR "computer-aided"[Title/Abstract])) NOT ("animals"[MeSH Terms] NOT "humans"[MeSH Terms])

Scopus

(TITLE-ABS-KEY ("third molar" OR "wisdom tooth" OR "wisdom teeth" OR "mandibular third molar" OR "lower third molar") AND TITLE-ABS-KEY ("inferior alveolar nerve" OR "alveolar nerve" OR "mandibular nerve" OR "mandibular canal" OR "inferior alveolar canal" OR "IAN" OR "IAC" OR "nerve injury" OR "paresthesia" OR "neurosensory") AND TITLE-ABS-KEY ("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network" OR "convolutional" OR "radiomics" OR "automated detection" OR "computer-aided") AND PUBYEAR > 2013 AND PUBYEAR < 2027)

Web of Science (Core Collection)

title AND abstract search (("third molar" OR "wisdom tooth" OR "wisdom teeth" OR "mandibular third molar" OR "lower third molar") AND ("inferior alveolar nerve" OR "mandibular canal" OR "inferior alveolar canal" OR "nerve injury" OR paresthesia) AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network*" OR radiomics))

IEEE Xplore (Command Search, All Metadata)

(("third molar" OR "wisdom tooth" OR "wisdom teeth" OR "mandibular third molar" OR "lower third molar") AND ("inferior alveolar nerve" OR "alveolar nerve" OR "mandibular nerve" OR "mandibular canal" OR "inferior alveolar canal" OR "IAN" OR "IAC" OR "nerve injury" OR "paresthesia" OR "neurosensory") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network" OR "convolutional" OR "radiomics" OR "automated detection" OR "computer-aided"))

CENTRAL (Cochrane Library), Title/Abstract/Keyword search

("third molar" OR "wisdom tooth" OR "wisdom teeth" OR "mandibular third molar" OR "lower third molar") AND ("inferior alveolar nerve" OR "alveolar nerve" OR "mandibular nerve" OR "mandibular canal" OR "inferior alveolar canal" OR "IAN" OR "IAC" OR "nerve injury" OR "paresthesia" OR "neurosensory") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network" OR "convolutional" OR "radiomics" OR "automated detection" OR "computer-aided")

Embase (via Ovid)

(third molar or wisdom tooth or wisdom teeth or mandibular third molar or lower third molar).mp. AND (inferior alveolar nerve or alveolar nerve or mandibular nerve or mandibular canal or inferior alveolar canal or IAN or IAC or nerve injury or paresthesia or neurosensory).mp. AND (artificial intelligence or machine learning or deep learning or neural network* or convolutional or radiomics or automated detection or computer-aided).mp. [Publication Year 2014–2026]

WHO ICTRP and Google Scholar (supplementary)

ICTRP simple searches: third molar AND artificial intelligence; wisdom tooth AND (deep learning OR machine learning); inferior alveolar nerve AND artificial intelligence (2014–2026). Google Scholar (<=256-character limit, no field tags): ("third molar" OR "wisdom tooth") AND ("inferior alveolar nerve" OR "mandibular canal") AND ("artificial intelligence" OR "deep learning" OR "machine learning"), custom range 2014–2026, first 100 records screened.

Supplementary Note 2

Meta-regression on study-level covariates, sensitivity analyses, the maximum-likelihood fit and SWiM reporting

Exploratory meta-regression on study-level covariates (FQ1)

Two covariates were fitted, each chosen for an explicit mechanism and each exploratory: publication year (a proxy for technological progress) and whether the contact 2×2 was read from a published confusion matrix or reconstructed (a proxy for extraction reliability). A third covariate, contact prevalence in the test set, was specified as a proxy for case-mix enrichment but was not fitted: across the seven 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. Each covariate was entered singly within the bivariate model rather than fitted separately to sensitivity and specificity. Priors are those of the meta-regression section of the application, which differ from those of the meta-analysis section: normal(0, 1.5) on the logit-sensitivity and logit-specificity intercepts, normal(0, 1) on the covariate coefficients, half-normal(0, 1) truncated at zero on the between-study standard deviations, and LKJ(2) on the between-study correlation. For the categorical covariate the application places normal(0, 1) on the group-specific logit sensitivities and specificities.

Covariate and outcome	Coefficient or difference	Posterior SD	95% credible interval	Interpretation
Publication year, sensitivity	−0.194	0.253	−0.697 to 0.336	No association
Publication year, specificity	+0.312	0.539	−0.759 to 1.366	No association
Reconstructed vs reported 2×2, difference in sensitivity	+0.047	0.064	−0.080 to 0.171	No difference
Reconstructed vs reported 2×2, difference in specificity	−0.051	0.148	−0.401 to 0.208	No difference

Every credible interval includes zero: there is no evidence that accuracy varies with publication year or with whether the 2×2 table was reported or reconstructed. Coefficients for the continuous covariate are on the logit scale, per calendar year; for the categorical covariate the entries are differences in sensitivity and specificity on the probability scale, reconstructed minus reported. Pooled estimates at the mean covariate value were sensitivity 0.809 (0.743–0.871) and specificity 0.866 (0.690–0.947), with prediction intervals of 0.607 to 0.930 and 0.254 to 0.992.

Sensitivity analyses (FQ1)

Analysis	Studies	Sensitivity (95% CrI)	Specificity (95% CrI)
Primary pool	7	0.811 (0.746–0.871)	0.867 (0.695–0.945)
Direct confusion matrices only	4	0.787 (0.692–0.862)	0.866 (0.600–0.959)
Replacing Güller with Kumbasar (overlapping cohort)	7	0.855 (0.746–0.916)	0.845 (0.696–0.920)
Replacing Fukuda 2024 with Fukuda 2026 (overlapping cohort)	7	0.815 (0.748–0.876)	0.891 (0.730–0.957)
Fukuda 2024 with halved cells (post hoc)	7	0.814 (0.748–0.874)	0.867 (0.695–0.945)

All five fits converged, with no divergent transitions, no iteration exceeding the maximum tree depth and all R-hat values $\leq$ 1.001. Across the analyses summary sensitivity ranged 0.79–0.86 and specificity 0.85–0.89, so no single study and no inclusion decision drives the pooled estimate. Summary estimates are those of the reference fit of the primary pool, from which the GRADE evidence profile (Supplementary Table 4) and Fig. 5 derive the expected consequences per 1000 patients tested. The halved-cell analysis for Fukuda et al., 2024 is post hoc and addresses the fact that its published table counts the same 100 test images twice.

Maximum-likelihood fit (FQ1)

The bivariate model was first fitted by maximum likelihood in MetaDTA v2.1.5. On the seven studies of the primary pool it does not converge: the application returns a singular fit at the boundary of the parameter

space, warns that most output cannot be displayed, and gives no summary sensitivity, specificity, likelihood ratio or diagnostic odds ratio. No maximum-likelihood summary is therefore available for these studies and none is reported. That the model is inestimable by maximum likelihood at this number of studies is why the Bayesian fit is the primary analysis, and why the posterior predictive region is reported alongside the summary operating point.

SWiM reporting items (FQ2 synthesis without meta-analysis)

#	SWiM item	How addressed
1	Grouping studies for synthesis	Grouped by model family (CNN on images vs machine learning on clinical/radiographic features) and by outcome (IAN injury/paraesthesia vs composite complications)
2	Standardised metric	None pooled; discrimination (AUC, sensitivity/specificity) reported per study as available
3	Synthesis method	Structured narrative synthesis; no meta-analysis; no vote-counting on significance
4	Criteria for prioritising results	Prioritised studies appraised with PROBAST+AI and reporting extractable model metrics
5	Investigation of heterogeneity	Described across model family, validation strategy, sample/event size, calibration, outcome definition (Supplementary Table 5)
6	Certainty of evidence	Narrative; very low certainty (all six appraised studies at high risk of bias; no external validation, calibration or net benefit)
7	Data presentation	Table 3 and the FQ2 results paragraph; per-model risk of bias in Supplementary Table 5

Supplementary Table 1

Per-study 2×2 contingency tables for FQ1

Per-study 2×2 counts for the FQ1 panoramic-index studies with a reported or uniquely reconstructable contact table. Seven studies entered the primary meta-analysis; Fukuda et al. (2026) and Kumbasar et al. (2025) contributed to a sensitivity analysis only, in place of the report with which each shares a cohort; Ulusoy et al. (2025) was excluded from the primary pool under a prespecified rule and, sharing a cohort with no other included study, contributed to no analysis; and Choi et al. (2022) was synthesised narratively. TP true positive, FN false negative, FP false positive, TN true negative. The 'Source of 2×2' column states whether counts were taken directly from a reported confusion matrix or uniquely reconstructed from reported sensitivity, specificity, the disease-positive count and the total sample size; reconstruction was performed only when no alternative integer-valued table was consistent with the reported rounding. No continuity correction was applied at any point: the bivariate model handles the single zero cell (Güller et al., FP = 0) directly through its exact binomial likelihood.

Study	Architecture	TP	FN	FP	TN	Sens	Spec	Source of 2×2
Jeon et al., 2023	Detection CNN	73	16	22	67	0.82	0.75	Reported confusion matrix
Takebe et al., 2023	Detection CNN	49	4	6	35	0.92	0.85	Reconstructed (recall+prec+N; internal Osaka test set)
Fukuda et al., 2024	Classification CNN	79	21	20	80	0.79	0.80	Reported confusion matrix
Jing et al., 2024	Segmentation+classifier	62	21	7	153	0.75	0.96	Reconstructed (sens+spec+N strat)
Güller et al., 2025	Classification CNN	65	25	0	74	0.72	1.00	Reported confusion matrix
Jiang et al., 2025	Classification CNN	43	7	15	24	0.86	0.62	Reconstructed (sens+spec+prev)
Yari et al., 2025	Classification CNN (ResNet-50)	22	4	6	50	0.85	0.89	Reported confusion matrix
Kumbasar et al., 2025	Classification CNN (ens.)	284	15	17	230	0.95	0.93	Reported confusion matrix
Ulusoy et al., 2025	Classical ML (features)	189	31	14	62	0.86	0.82	Not pooled in any analysis (mean of 30 Monte-Carlo splits; no single held-out table)
Fukuda et al., 2026	Classical ML (features)	40	10	2	48	0.80	0.96	Reconstructed (sens+spec+N)
Choi et al., 2022	Classification CNN	55	10	21	26	0.85	0.55	Not pooled; narrative synthesis (single best split, discordant)

The per-study derivation of each 2×2 — source of the counts, reported values used, formula applied, and verification of the derived table against every accuracy metric published by the source — is given in full in Supplementary Table 1b. Where the class split of the test set was not reported and had to be solved, Supplementary Table 1b additionally lists every integer-valued table compatible with the published metrics within rounding, and the rule used to select among them. The completed extraction form is deposited with the analysis datasets.

Supplementary Table 1b

Derivation record for the FQ1 contact 2×2 tables

For every FQ1 panoramic-index study with a poolable contact 2×2 this table records the source of the counts, the reported values used, the formula applied where a derivation was needed, and the verification of the derived table against every accuracy metric published by the source. Tables are classified at one of three levels of directness. Level 1, read directly from a published confusion matrix. Level 2, reconstructed from a reported operating point together with a class split reported by the source. Level 3, reconstructed with a class split solved from the reported rates and the total, in which case the integer table is not unique and every compatible alternative is listed below.

Before pooling, each candidate table was checked for internal consistency. Sensitivity, specificity, accuracy and the two predictive values are all functions of the same four counts, so accuracy, positive predictive value and negative predictive value each imply the ratio of diseased to non-diseased observations in the test set, and those implied ratios must agree if a single table underlies them. Every table below satisfies this check to within 2%.

Part A. Derivation, study by study

Study	Level	Source of the 2×2	Test set (n; class split)	Reported values used	2×2 (TP/FN/FP/TN)	Verification against the source
Jeon 2023	1	Published confusion matrix	178 (89 contact / 89 non-contact)	Counts read directly; accuracy 78.65, Se 82.02, Sp 75.28 (Table 2, EfficientDet-D4)	**73 / 16 / 22 / 67**	Se 73/89 = 0.8202, Sp 67/89 = 0.7528, accuracy 140/178 = 0.7865. All three reproduce exactly.
Takebe 2023	2	Reconstructed; class split reported	94 (53 contact / 41 intact); internal Osaka set	Table 2, final model: accuracy 0.894, recall 0.925, precision 0.891, F1 0.908	49 / 4 / 6 / 35	Se 0.925, Sp 0.854, precision 49/55 = 0.891, F1 0.908. All four reproduce exactly. The external Ikeda set (58 images) is not the pooled arm.
Fukuda 2024	1	Published confusion matrix	200 observations (100 contact / 100 non-contact)	Counts read from Table 1; Table 2 gives Se 0.79, Sp 0.80	**79 / 21 / 20 / 80**	Se 79/100 = 0.79, Sp 80/100 = 0.80. Exact. Caveat: the 200 observations are 100 test patches evaluated in two runs, so the study enters the pool at twice its independent information content. A post-hoc sensitivity analysis halving the cells is reported in Supplementary Note 2.
Jing 2024	3	Reconstructed; class split solved	243; split solved as 83 contact / 160 non-contact	Table II, MM3-IACnet: accuracy 0.885, Se 0.747, Sp 0.956, precision 0.899	62 / 21 / 7 / 153	Se 62/83 = 0.747, Sp 153/160 = 0.956, accuracy 215/243 = 0.885, precision 62/69 = 0.899. All four reproduce exactly, and no alternative split does so on all four (Part B).
Güller 2025	1	Published confusion matrix	164 (90 contact / 74 non-contact); 70:30 hold-out	Fig. 4c, right panel, horizontal RoI Inception-v3: 65 / 0 / 25 /	**65 / 25 / 0 / 74**	Se 65/90 = 0.722, Sp 74/74 = 1.00, accuracy 139/164 = 0.848, matching the accuracy of Table 5. Table 5 also reports recall 86.11 and precision 87.37, which are macro-averages across the

Study	Level	Source of the 2×2	Test set (n; class split)	Reported values used	2×2 (TP/FN/FP/TN)	Verification against the source
				74		two classes and do not describe the contact class alone; the confusion matrix was preferred.
Jiang 2025	3	Reconstructed; class split solved	89; split solved as 50 contact / 39 non-contact; fixed 80:20 hold-out	Table I, ResNet50 contact row, means over seeds 40–44: accuracy 0.748, recall 0.858, specificity 0.608	43 / 7 / 15 / 24	Se 0.860 against 0.858 reported, Sp 0.615 against 0.608, accuracy 0.753 against 0.748. Only three metrics are reported, so no candidate reproduces them exactly and the choice among candidates is not unique (Part B). Caveat: the reported metrics are means of ratios over five model initialisations, not one observed matrix.
Yari 2025	1	Published confusion matrix	82 (26 contact / 56 non-contact)	Table 1, row "Artificial intelligence"; gold standard stated as 26 positive and 56 negative	**22 / 4 / 6 / 50**	Se 22/26 = 0.846, Sp 50/56 = 0.893, accuracy 72/82 = 0.878. Reproduce the reported 84.61, 89.28 and 87.80 exactly.
Kumbasar 2025	1	Published confusion matrix	546 (299 related / 247 unrelated); ten-fold cross-validation over the whole cohort	Majority voting, related versus unrelated: accuracy 94.1%	284 / 15 / 17 / 230	Se 284/299 = 0.950, Sp 230/247 = 0.931, accuracy 514/546 = 0.9414, reproducing the published 94.1% exactly. Reserved for sensitivity analysis under rule (v): no single fixed set of observations was withheld.
Ulusoy 2025	—	Not derivable	1478 in total (1100 contact / 378 non-contact); thirty Monte-Carlo 80:20 re-splits	Table 2, ANN6 (η = 0.65), stated in the table caption as the mean of thirty experiments: accuracy 85.45, F1 90.76, SE 86.36, SP 81.92, PPV 95.30, NPV 62.44	**not derivable**	The six reported metrics are not jointly realisable by any single 2×2. Accuracy, PPV and NPV imply ratios of contact to non-contact observations of 3.88, 4.25 and 3.61 respectively, against a dataset ratio of 2.91; a single table would give one ratio by all three routes. This is the expected arithmetic consequence of averaging non-linear ratios across repeated splits, not a defect of the source. Excluded from the primary pool under rule (v) and, sharing a cohort with no other included study, from the substitution analyses as well, so no table was pooled for it. A class-stratified 20% partition would give 189 / 31 / 14 / 62, but the source does not claim to have used one.
Fukuda 2026	2	Reconstructed; class split reported	100 (50 high-risk / 50 low-risk); balanced by design	Prediction One: Se 0.80, Sp 0.96; test set stated as 50 + 50 teeth	40 / 10 / 2 / 48	Se 40/50 = 0.80, Sp 48/50 = 0.96. Exact on an explicitly balanced test set. Reserved for sensitivity analysis as the companion report of Fukuda 2024.

Part B. Alternative integer tables for the two studies whose class split was solved

Where the class split of the test set is not reported, more than one integer table can be compatible with the published metrics within rounding. Both sets are listed in full. The adopted table is shown in bold.

Jing et al., 2024 — n = 243; targets accuracy 0.885, Se 0.747, Sp 0.956, PPV 0.899; tolerance ±0.005 on each

Split (P / N)	TP / FN / FP / TN	Se	Sp	Accuracy	PPV	Note
80 / 163	60 / 20 / 7 / 156	0.750	0.957	0.889	0.896	Accuracy off by 0.004
82 / 161	61 / 21 / 7 / 154	0.744	0.957	0.885	0.897	Se off by 0.003
83 / 160	**62 / 21 / 7 / 153**	**0.747**	**0.956**	**0.885**	**0.899**	**Adopted: the only candidate reproducing all four reported metrics exactly**
84 / 159	63 / 21 / 7 / 152	0.750	0.956	0.885	0.900	Se off by 0.003
86 / 157	64 / 22 / 7 / 150	0.744	0.955	0.881	0.901	Accuracy off by 0.004
87 / 156	65 / 22 / 7 / 149	0.747	0.955	0.881	0.903	Accuracy off by 0.004

Jiang et al., 2025 — n = 89; targets accuracy 0.748, Se 0.858, Sp 0.608; tolerance ±0.010 on each

Split (P / N)	TP / FN / FP / TN	Se	Sp	Accuracy	Note
45 / 44	39 / 6 / 17 / 27	0.867	0.614	0.742	Total absolute deviation 0.021
48 / 41	41 / 7 / 16 / 25	0.854	0.610	0.742	Total absolute deviation 0.012, the smallest of the set
49 / 40	42 / 7 / 16 / 24	0.857	0.600	0.742	Closest on sensitivity (0.001)
50 / 39	**43 / 7 / 15 / 24**	**0.860**	**0.615**	**0.753**	**Adopted; total absolute deviation 0.014**
51 / 38	44 / 7 / 15 / 23	0.863	0.605	0.753	Total absolute deviation 0.012
53 / 36	45 / 8 / 14 / 22	0.849	0.611	0.753	Total absolute deviation 0.017
54 / 35	46 / 8 / 14 / 21	0.852	0.600	0.753	Total absolute deviation 0.019

With only three reported metrics, no single criterion identifies the Jiang et al. table uniquely: the candidate closest overall is 48 / 41 and the candidate closest on sensitivity is 49 / 40, while the adopted 50 / 39 lies between them. The choice is immaterial to the synthesis, because across all seven candidates sensitivity spans 0.849 to 0.867 and specificity 0.600 to 0.615, a range narrower than the width of this study's own exact binomial confidence intervals. We nonetheless record the ambiguity rather than present the adopted table as uniquely determined.

Supplementary Table 2

Full-text records excluded, with reasons

Twenty-eight reports were excluded at full-text eligibility. The dominant reason was a reference standard based on annotation of the panoramic radiograph rather than CBCT/3D imaging, which cannot answer the FQ1 question. Studies that addressed a broader composite complication outcome (Karadiguddi 2026; Jain 2026) were retained for the narrative FQ2 synthesis, not excluded.

Author (year)	Title (short)	DOI	Reason for exclusion
Carvalho JS (2023)	Preinterventional Third-Molar Assessment Using Robust Machin…	10.1177/00220345231200786	Reference standard not CBCT/3D (annotation on panoramic radiograph)
Fukuda M (2020)	Comparison of 3 deep learning neural networks for classifyin…	10.1016/j.oooo.2020.04.005	Reference standard not CBCT/3D (annotation on panoramic radiograph)
Buyuk C (2022)	A Fused Deep Learning Architecture for the Detection of the …	10.3390/diagnostics12082018	Reference standard not CBCT/3D (annotation on panoramic radiograph)
Lee J. (2022)	Automated Prediction of Extraction Difficulty and Inferior A…	10.3390/app12010475	Non-clinical risk label / overlapping cohort
Kempers S (2023)	Positional assessment of lower third molar and mandibular ca…	10.1016/j.jdent.2023.104519	Reference standard not CBCT/3D (annotation on panoramic radiograph)
Lo Casto A (2023)	Artificial Intelligence for Classifying the Relationship bet…	10.3390/life13071441	Reference standard not CBCT/3D (annotation on panoramic radiograph)
Papasratorn D (2023)	Investigation of the best effective fold of data augmentatio…	10.1007/s00784-023-04992-6	CBCT reference, but a two-gate case-selected design (900 clearly-contact vs 900 clearly-no-contact, with a separate 'false-contact' Group III of panoramic–CBCT discordant cases) rather than a consecutive series; this case selection precludes a valid contact-vs-no-contact 2×2 for triage
Kim JY (2023)	The efficacy of supervised learning and semi-supervised lear…	10.1259/dmfr.20230030	Reference standard not CBCT/3D (annotation on panoramic radiograph)
Joo Y. (2023)	Classification of the Relationship between Mandibular Third …	10.1109/access.2023.3302271	Reference standard not CBCT/3D (annotation on panoramic radiograph)
Fang X (2024)	Automatic detection of the third molar and mandibular canal …	10.1016/j.jormas.2024.101946	No diagnostic-accuracy contingency data
Gong Z (2024)	System for automatically assessing the likelihood of inferio…	10.1016/j.compbiomed.2024.107923	Other: no reconstructable contact 2×2 (injury-likelihood score with segmentation on the panoramic radiograph, no contact classification)
Huang C (2024)	Automatic detection and proximity quantification of inferior…	10.1007/s00784-024-05967-x	Segmentation/detection only; no reconstructable contact 2×2
Kazimierczak W. (2024)	Reliability of the AI-Assisted Assessment of the Proximity o…	10.3390/jcm13123605	Not M3M-specific / relationship to M3M unclear
Chen S.-L. (2024)	Classification of the Relative Position between the Third Mo…	10.3390/electronics13040702	Reference standard not CBCT/3D (annotation on panoramic radiograph)
P. Manikandaprab	Combined Fire Fly -Support Vector Machine Digital	10.47738/jads.v5i3.356	Unclear reference standard

Author (year)	Title (short)	DOI	Reason for exclusion
hu (2024)	Radiograph…		
Klaib AF (2025)	Automated assessment and detection of third molar and inferi…	10.1038/s41598-025-17753-0	Reference standard not CBCT/3D (annotation on panoramic radiograph)
Khiyavi A.A. (2025)	Effectiveness of lightweight convolutional neural networks f…	10.22038/jdmt.2025.86587.1801	Reference standard not CBCT/3D (annotation on panoramic radiograph)
Lee J (2025)	Automated diagnosis for extraction difficulty of maxillary a…	10.1038/s41598-025-00236-7	Non-clinical risk label / overlapping cohort
Lyu W (2025)	Assessing the spatial relationship between mandibular third …	10.1186/s12903-025-06539-5	Segmentation/detection only; no reconstructable contact 2×2
Ünal SY (2025)	Evaluation of the mandibular canal and the third mandibular …	10.1007/s11282-024-00793-z	Segmentation/detection only; no reconstructable contact 2×2
Lin YJ (2025)	Precision Oral Medicine: A DPR Segmentation and Transfer Lea…	10.1109/jtehm.2025.3568922	Other: no reconstructable contact 2×2 (panoramic segmentation with a transfer-learned likelihood score, no contact classification)
Ko J (2026)	Quantitative Assessment of Third Molar Extraction Difficulty…	10.1007/s10439-026-04114-9	No diagnostic-accuracy contingency data
R. C H (2026)	Vision Transformers for Accurate Inferior Alveolar Nerve Cla…	10.1109/aide69088.2026.11544588	Not M3M-specific / relationship to M3M unclear
Kim H (2026)	Automatic multi-class classification of 3D M3M–canal relationships (CBCT)	10.1093/dmfr/twag010	CBCT index test coinciding with reference standard
Yasin ET (2025)	Automatic mandibular third molar and mandibular canal relationship detection (CBCT)	10.1007/s00784-025-06285-6	CBCT index test coinciding with reference standard
Barnes NA (2024)	Automated classification of mandibular canal in relation to third molar (CBCT)	10.12688/f1000research.154985.1	CBCT index test coinciding with reference standard
Chun S-Y (2023)	Automatic classification of 3D positional relationship (CBCT)	10.1186/s12903-023-03496-9	CBCT index test coinciding with reference standard
Liu M-Q (2022)	Deep learning-based evaluation of the M3M–canal relationship (CBCT)	10.1007/s00784-021-04082-5	CBCT index test coinciding with reference standard

Supplementary Table 3

QUADAS-3 risk-of-bias and applicability judgements (FQ1)

Per-study domain judgements for the 15 FQ1 diagnostic-accuracy studies.

Tool: QUADAS-3.

Index test as used in each study (OPG = panoramic radiograph; CBCT = cone-beam CT).

RoB domains: D1 participants, D2 index test, D3 target condition (reference standard), D4 analysis.

Applicability concern: A1 participants, A2 index test, A3 target condition.

L = low; H = high. No domain judgement in this appraisal was rated insufficient information, so no cell is plotted as 'unclear' in Fig. 2.

Risk of bias and applicability summary

Study	Index	D1	D2	D3	D4	A1	A2	A3	Overall RoB	Overall applicability
Jeon 2023	OPG	H	L	L	H	H	L	L	High	High
Takebe 2023	OPG	H	L	L	H	H	L	L	High	High
Fukuda 2024	OPG	H	H	L	H	H	H	H	High	High
Jing 2024	OPG	H	L	L	H	H	H	L	High	High
Güller 2025	OPG	H	H	L	H	H	H	L	High	High
Jiang 2025	OPG	H	L	L	H	H	H	L	High	High
Yari 2025	OPG	H	H	L	H	H	H	L	High	High
Kumbasar 2025	OPG	H	H	L	H	H	H	L	High	High
Ulusoy 2025	OPG	H	H	L	H	H	H	L	High	High
Fukuda 2026	OPG	H	H	H	H	H	H	H	High	High
Zhu 2021	OPG	H	L	L	H	H	H	L	High	High
Choi 2022	OPG	L	L	L	H	H	H	L	High	High
Sukegawa 2022	OPG	H	L	L	H	H	H	L	High	High
Soltani 2024	OPG	H	L	L	H	H	H	L	High	High
Fernandes 2025	OPG	H	H	L	H	H	H	L	High	High

Per-study rationale

The judgements above are set out study by study below. Each note states only what the source article reports, and the wording follows the domain that drove the overall rating. Where the analysis domain is high, the reason is given, since it is high for all fifteen studies and the drivers differ. Applicability is high overall in every study for the same underlying reason: none matches the ideal study against which it is judged, in which a consecutive series is enrolled prospectively at the contact prevalence of clinical practice and the model is applied to the panoramic image in its stated triage role at a threshold fixed in advance. The participants domain is therefore of high concern throughout; the index test is of high concern wherever the operating point was chosen on the test data or the image was pre-processed in a way that has no counterpart at the chairside; and the target condition only in the two Fukuda et al. reports.

Jeon et al., 2023. Sampling was retrospective and the 178-unit test set was balanced at 89 contact and 89 non-contact, a composition chosen for training convenience rather than one a clinic would meet. The reference standard is the strongest in the pool: CBCT read independently, with intraoperative confirmation. Three detectors were built and EfficientDet-D4 is the one reported here. Analysis remained at the level of the individual tooth, with no allowance for the two third molars a patient may contribute, which is what carries the analysis domain to high.

Takebe et al., 2023. 579 panoramic images from 384 patients were divided 483 to 96 for training and internal testing, and a further 58 images came from a second institution and were used for testing only. The model localised 94 of the 96 internal images and 55 of the 58 external ones, and accuracy was computed on those. The 2×2 pooled here was reconstructed from recall and precision on the internal set. This is the one study in this review with a genuine external test set, so the analysis domain is high for the per-tooth counting and the reconstruction, not for the validation design.

Fukuda et al., 2024. The test set was balanced by construction and the operating point was fixed after the fact using the Youden index, which is what raises the index-test domain. Training and testing were then repeated over the same 100 patches, so the 200 observations in the published table are 100 images counted twice, and the weight this study carries in the pool is correspondingly overstated. The reference standard is CT or CBCT and is adequate.

Jing et al., 2024. 1570 subjects with paired panoramic and CBCT imaging contributed 2543 sides, split at random into 80% training, 10% validation and 10% testing. Because the split was not stratified by contact status, the test set carries the cohort's own prevalence of roughly 31% rather than an engineered balance, but the participants domain is nonetheless high: the cohort was assembled from patients who had undergone panoramic and CBCT imaging for reasons unrelated to the third molar, among them dentofacial deformity, root canal treatment and periodontal surgery, and images of insufficient quality were excluded. The pipeline segments first and classifies second. The composition of the test split is not reported, which is why the contingency table had to be reconstructed.

Güller et al., 2025. A retrospective series of 290 patients and 546 third molars, with 164 units held out for testing and CBCT as the reference. Six arms were trained, three architectures across two region-of-interest definitions, so the figure that enters the meta-analysis is one selection among several the study reports. The index-test domain is high on two counts. The recall of 86.11% and precision of 87.37% in its Table 5 are macro-averages across the two classes, not the sensitivity and positive predictive value for contact — they are the means of 72.22 and 100, and of 100 and 74.75 — so the contact sensitivity of 65/90, that is 72.2%, had to be recovered from the confusion matrix in Fig. 4c; and mirrored images were added by augmentation before a random 70:30 split, with no stated guarantee that a patient, or its mirrored duplicate, stayed on one side of it. Validation stayed inside the institution. Counts are per tooth.

Jiang et al., 2025. 279 patients contributed 441 third molars, each with one panoramic radiograph and one CBCT. Six architectures were compared using a held-out 20% test set excluded from training, with five-fold cross-validation on the training partition used only for model selection; the reported figures average this held-out test-set performance over five random initialisations (seeds 40–44) rather than resting on a single run. The participants domain is high because 128 of the 407 eligible patients, 31.4%, were excluded for blurred panoramics or for artefacts obscuring the third molar, that is the hardest cases, and eligibility was restricted to ages 18 to 32. The 2×2 was reconstructed from the reported sensitivity, specificity and prevalence, and counting is per tooth.

Yari et al., 2025. 548 images were split 70:15:15, leaving 82 for testing. Panoramic and CBCT imaging were acquired on the same day and the reference standard was set by a maxillofacial radiologist, so the target-condition domain is low. Contact prevalence is 174 of 548 in the dataset and 26 of 82 in the test set, close to 32% in both, so the split was random and the case mix was not balanced by construction. The index-test domain is high because a 256×256 region of interest was cropped by hand around the closest point between root and canal before classification, a step with no counterpart at the chairside. Analysis is per tooth with no correction for clustering. The 2×2 is the one published in the source and reproduces all five reported metrics exactly.

Kumbasar et al., 2025. This report draws on the same 290-patient cohort as Güller et al. and was therefore reserved for sensitivity analysis. Its confusion matrix aggregates ten cross-validation folds across all 546 units, so every observation served as a training example in nine of the ten rounds and none was withheld. The classifier is a majority vote over three networks. Analysis is per tooth.

Ulusoy et al., 2025. Radiographic features drawn from a public dataset of 1010 patients and 1478 third molars were fed to a multilayer perceptron. Six configurations are reported, differing only in the decision threshold, and three of them were added after the default threshold of 0.5 returned a specificity of 44.9%. That post-hoc thresholding is what carries the index-test domain to high. Performance rests on thirty random 80:20 subsamples rather than one fixed test set, and counting is per tooth.

Fukuda et al., 2026. From the same institution and an overlapping cohort as the 2024 report, which is why it was reserved for sensitivity analysis. A 50:50 hold-out was used, the cutoff was not fixed in advance, and feature-based AutoML models were reported alongside convolutional networks. The authors name the absence of external validation among their own limitations.

Zhu et al., 2021. A detection model built on YOLOv4, reported with precision, recall and average precision alone. Without a specificity or the underlying counts there is nothing to reconstruct, and the analysis domain is high on the reporting item rather than on the modelling. Retrospective, per tooth, narrative synthesis.

Choi et al., 2022. The target condition was split as true contact against false or intimate contact, a partition that does not line up with the contact and no-contact dichotomy used elsewhere in this review. The sensitivity and specificity pair comes from a single best split and sits at odds with the averaged contact accuracy of 0.63 reported for the same test set. This is the one study whose participants domain is low, because access to CBCT followed an objective criterion taken from guidance, radiographic overlap on the panoramic, rather than a case-by-case suspicion. Counting is per image. The study was synthesised narratively.

Sukegawa et al., 2022. 1279 units, unevenly divided at 972 against 307, were evaluated by tenfold cross-validation for internal validity. Metrics are macro-averaged across folds and no specificity is given, so the contact and non-contact counts cannot be recovered and no coherent contingency table exists. Counting is per tooth. The study was synthesised narratively.

Soltani et al., 2024. A two-stage pipeline, Faster R-CNN followed by ResNeXt, trained on 387 patients and 619 third molars under an 80:10:10 split that leaves 58 images for testing. The accuracy, recall and specificity that the paper reports cannot be reconciled to any single contingency table, so no 2×2 was derived and the study was synthesised narratively. Counting is per tooth.

Fernandes et al., 2025. A multi-task benchmark in which the contact task appears only as an area under the curve for three architectures. No operating point is given and no contingency table can be derived, which places the analysis domain at high risk on the same item as Zhu et al. and the index-test domain likewise, since without a stated threshold the test as applied cannot be tied to the binary contact decision this review concerns. Retrospective; synthesised narratively.

Supplementary Table 4

GRADE evidence profile for diagnostic test accuracy (FQ1)

Outcome	Studies (teeth/images)	Summary estimate (95% CrI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Publication bias	Certainty
Sensitivity	7 (1,050)	0.81 (0.75–0.87)	Very serious (-2)	Not serious	Serious (-1)	Not serious	Serious (-1)	Very low
Specificity	7 (1,050)	0.87 (0.70–0.95)	Very serious (-2)	Serious (-1)	Serious (-1)	Not serious	Serious (-1)	Very low

Footnotes. Summary estimates are posterior medians from the Bayesian bivariate model (MetaBayesDTA v1.5.3) and the intervals are 95% credible intervals. (RoB) all studies at high risk of bias on QUADAS-3; bias tends to inflate accuracy. (Inconsistency, specificity) marked unexplained heterogeneity (per-study 0.62–1.00; wide 95% prediction region). (Indirectness) heterogeneous architectures, predominantly single-centre retrospective datasets, no deployed triage. (Publication bias) small, predominantly single-centre, positive-result literature; too few studies for a reliable funnel test. Intervals are the exact posterior intervals of the model, 0.746–0.871 for sensitivity and 0.695–0.945 for specificity, rounded to two decimals. Fig. 5 is the GRADEpro GDT export and displays the same intervals truncated rather than rounded, as 0.74 to 0.87 and 0.69 to 0.94.

Summary of findings, per 1000 patients tested at three plausible pretest probabilities of contact (10%, 30% and 50%):

Pretest probability	True positives	False negatives	True negatives	False positives	Certainty
10% (unselected screening)	81 (75–87)	19 (13–25)	780 (625–850)	120 (50–275)	Very low
30% (referred, base case)	243 (224–261)	57 (39–76)	607 (486–662)	93 (38–214)	Very low
50% (selected referral)	406 (373–436)	94 (64–127)	434 (347–473)	66 (27–153)	Very low

Overall certainty very low; AI is not yet supported as a clinical triage tool for the M3M–IAN relationship.

Supplementary Table 5

PROBAST+AI risk-of-bias and applicability judgements (FQ2)

Per-model judgements for all six FQ2 prediction-model studies appraised with PROBAST+AI.

RoB domains: D1 participants, D2 predictors, D3 outcome, D4 analysis. Applicability concern: A1 participants, A2 predictors, A3 outcome.

L = low, H = high.

Study	D1	D2	D3	D4	A1	A2	A3	Overall RoB	Overall applicability
Feher 2021	L	L	L	H	L	L	L	High	Low
Kim BS 2021	H	L	L	H	L	L	L	High	Low
Picoli 2023	H	L	H	H	L	L	H	High	High
Yoon 2026	H	L	L	H	L	L	L	High	Low
Jain 2026	L	H	H	H	L	H	H	High	High
Karadiguddi 2026	H	L	H	H	L	L	H	High	High

Rationale per model

The judgements above are set out model by model below. Each note states only what the source article reports, and names the domain that carried the overall rating. All six models were rated high overall, so what distinguishes them is which domain did the work.

Feher 2021. A dataset of 648 patients and 812 impacted third molars, larger and closer to a routine surgical population than anything else in this group, which is why the participants domain gives least concern. LASSO was tuned by tenfold cross-validation, and the authors respected the dependency created by patients contributing one or two molars, a precaution none of the other five took. What sinks the analysis domain is the event count: 14 neuropathies across 812 molars leaves roughly one event per candidate predictor, far below the ten usually asked for, so the model is fitted on very little signal. Discrimination is reported and calibration is not.

Kim BS 2021. 300 preoperative panoramic images, split evenly at 150 for training and validation and 150 for testing, with an SSD300 detector feeding a ResNet-18 classifier. One hundred of the 300 images are paraesthesia cases, a proportion far above the rate seen in practice, so the sample is enriched by design and the participants domain is high. Augmentation was applied before splitting, which leaves room for the same patient to appear on both sides of the partition. Accuracy, sensitivity, specificity and AUC are given; calibration is not.

Picoli 2023. Twenty-five patients were selected from a database of 6,010 because they had bilateral removal with injury on one side only. Each patient therefore contributes an injured and an uninjured molar, and the 50% event rate is a property of the selection rather than of the disease. The design isolates imaging differences within a patient efficiently and tells you little about absolute risk, which places both the outcome domain and its applicability at high concern. Discrimination is the only measure reported, and at 0.63 it is the weakest in the set.

Yoon 2026. Five architectures were trained on 902 panoramic radiographs and 1,500 CBCT images, with contrastive learning and Bayesian hyperparameter tuning, then compared against oral and maxillofacial surgeons. Classes are unbalanced and the panoramic and CBCT collections overlap, so images from one patient can sit on both sides of the split. Performance is given as accuracy and F1-score without confidence intervals, and neither calibration nor net benefit appears. That is what carries the analysis domain.

Jain 2026. A retrospective chart review of 472 consecutive extractions, the largest sample among the six, so the participants domain gives no concern. Five algorithms were compared and interpreted with SHAP. The difficulty is what goes into them: the authors set out to capture intraoperative and operator-related influences, so the model uses information that does not exist at the moment a clinician would want the prediction. Paraesthesia is one of eight complications, recorded by the operating surgeon over a follow-up of a few days. Validation is internal, and calibration and decision-curve analysis are absent.

Karadiguddi 2026. Forty patients, thirty for training and ten for testing, with a random forest predicting surgical difficulty and a convolutional network reading CBCT for postoperative morbidity. The article describes itself both as cross-sectional and as prospective observational, and the sample is purposive rather than consecutive, so the participants domain is high. The outcome is a composite of pain, swelling and paraesthesia scored at 7 and 14 days, neither specific to nerve injury nor long enough to separate transient from permanent deficit, and that carries the outcome domain and its applicability together. Ten test patients cannot support the discrimination reported.

PRISMA-DTA checklists

Status: R reported; NA not applicable. Line numbers refer to the line-numbered manuscript file and are given as the range spanning the relevant text. They were verified item by item against the line-numbered PDF of the submitted manuscript, which carries 707 numbered lines, and must be re-verified if the text changes. Section numbers are those of the manuscript headings.

PRISMA-DTA main checklist

#	Section / topic	Item	Status	Location
1	Title	Identify as a systematic review of diagnostic test accuracy	R	Lines 1–3 (title)
2	Abstract	Abstract covering the PRISMA-DTA for Abstracts items (unstructured, per npj Digital Medicine Article format, 150-word limit)	R	Lines 13–24 (Abstract)
3	Introduction: rationale	Rationale in the context of what is already known	R	Lines 27–75 (section 1)
4	Introduction: objectives	Question(s) with reference to participants, index test, target condition (PIT)	R	Lines 67–75 (section 1); eligibility elements, lines 83–112
5	Methods: protocol/registration	Registration number and registry; where protocol can be accessed	R	Lines 79–81 (section 2.1; PROSPERO CRD420251000827)
6	Methods: eligibility criteria	Study characteristics and report characteristics used as eligibility	R	Lines 83–112 (section 2.2)
7	Methods: information sources	Databases, registers, other sources; date last searched	R	Lines 114–118 (section 2.3; last search 26 June 2026); Supplementary Note 1
8	Methods: search	Full search strategy for at least one database	R	Supplementary Note 1 (all databases); cited at lines 117–118
9	Methods: study selection	Process for selecting studies (screening, eligibility)	R	Lines 120–125 (section 2.4)
10	Methods: data collection	Data extraction method; processes for obtaining/confirming data	R	Lines 127–131 (section 2.5; piloted form; duplicate independent extraction by F.M. and G.M., adjudicated by P.M.)
11	Methods: definitions	Definitions used in extraction (index tests, target condition, reference standard, units)	R	Lines 83–112 and 127–131; Supplementary Table 1 and Supplementary Table 1b
12	Methods: risk of bias	Methods to assess risk of bias and applicability	R	Lines 133–136 (section 2.6; QUADAS-3, PROBAST+AI, RoBVIS)
13	Methods: diagnostic accuracy measures	Accuracy measures and units (sensitivity, specificity, LR)	R	Lines 138–146; results at lines 238–243
14	Methods: synthesis of results	Methods of handling data and combining results (model, heterogeneity)	R	Lines 138–169 (section 2.7; Bayesian bivariate model, HSROC parameterisation, posterior predictive region)
15	Methods: additional analyses	Sensitivity/subgroup analyses, meta-regression, publication-bias method	R	Lines 152–159 (two covariates fitted; a third specified but not fitted) and 171–176 (section 2.8; certainty, publication bias); Supplementary Table 4 and Supplementary Note 2
16	Methods: meta-analysis	If meta-analysis done, statistical method, software	R	Lines 160–169 (the primary analysis, the sensitivity analyses and the meta-regressions fitted in MetaBayesDTA, with priors, sampler settings and convergence diagnostics reported; the maximum-likelihood fit attempted in MetaDTA did not

#	Section / topic	Item	Status	Location
				converge). Software versions are given in the supplementary material, not in the main text: MetaDTA v2.1.5 in Supplementary Note 2 and MetaBayesDTA v1.5.3 in Supplementary Table 4
17	Results: study selection	Numbers screened, assessed, included/excluded with reasons; flow diagram	R	Lines 178–199 (section 3); Figure 1 (line 205); Supplementary Table 2
18	Results: study characteristics	Characteristics of included studies and citations	R	Lines 200–204 and 359–373; Table 1 (line 207) and Table 3 (line 374)
19	Results: risk of bias within studies	Risk-of-bias and applicability results	R	Lines 226–237 (section 3.1); Figure 2 (line 271); Supplementary Table 3
20	Results: individual studies	Per-study 2×2 (TP, FP, FN, TN) or accuracy with CIs; forest plots	R	Lines 238–243; Figure 3 (line 279); Supplementary Table 1
21	Results: synthesis	Summary estimates with confidence or credible intervals; consistency	R	Lines 238–243 (summary sensitivity and specificity with 95% credible intervals, likelihood ratios, prediction region); Figure 4 (line 282)
22	Results: risk of bias across studies	Publication/selective-reporting risk across studies	R	Lines 266–270; method at lines 175–176; Supplementary Table 4
23	Results: additional analyses	Results of additional analyses	R	Lines 244–261; Supplementary Note 2
24	Discussion: summary of evidence	Summary of main findings; strength of evidence	R	Lines 395–400 (section 4)
25	Discussion: limitations	Limitations at study and review level	R	Lines 474–487 (study level) and 509–538 (review level)
26	Discussion: conclusions	Interpretation; implications for practice and research	R	Lines 540–545 (section 5)
27	Funding	Funding sources and role of funders	R	Lines 550–551 (Funding)

PRISMA-DTA for Abstracts checklist (12 items)

#	Item	Status	Location in abstract
1	Title: identify as a DTA systematic review	R	Lines 1–3 (title)
2	Background: review question (PIT)	R	Lines 14–16; target condition at lines 19–20
3	Objectives: explicit objective(s)	R	Lines 14–17
4	Eligibility criteria	R	Lines 16–17
5	Information sources: databases and last search	R	Lines 15–16 (seven databases and registers; last search June 2026)
6	Risk of bias: methods	R	Line 17 (QUADAS-3, PROBAST+AI)
7	Included studies: number and characteristics	R	Line 18 (21 studies; 7 pooled)
8	Synthesis of results: summary sensitivity and specificity with credible intervals	R	Lines 18–20
9	Strengths/limitations of evidence (certainty)	R	Lines 20–21 (certainty very low)
10	Interpretation/conclusions	R	Lines 21–23
11	Funding (primary source)	NA	Not stated in the abstract; declared at lines 550–551
12	Registration: number and registry	R	Line 24 (PROSPERO CRD420251000827)