Journal of Dentistry 160 (2025) 105868

Contents lists available at ScienceDirect

Journal of Dentistry

journal homepage: www.elsevier.com/locate/jdent

Impact of artificial intelligence assistance on diagnosing periapical radiolucencies: A randomized controlled trial

Utku Pul [a], Antonin Tichy [b,c], Vinay Pitchika [b], Falk Schwendicke [b,*]

[a] DTMD University for Digital Technologies in Medicine and Dentistry, Château de Wiltz, 9516 Wiltz, Luxembourg
[b] Department of Conservative Dentistry and Periodontology, LMU University Hospital, Goethestraße 70, 80336 München, Germany
[c] Institute of Dental Medicine, First Faculty of Medicine of the Charles University and General University Hospital in Prague, Karlovo náměstí 32, 121 11 Prague, Czech Republic

ARTICLE INFO

Keywords:
Artificial intelligence
Periapical radiolucency
Panoramic radiography
Diagnostic accuracy
Randomized controlled trial
Dentistry

ABSTRACT

Objectives: This randomized controlled trial aimed to evaluate the impact of artificial intelligence (AI) assistance on dentists' diagnostic accuracy, confidence, and treatment decisions when detecting periapical radiolucencies (PRs) on panoramic radiographs. We specifically investigated whether AI support influenced diagnostic performance across different levels of clinical experience.

Methods: Thirty dentists with varying levels of experience evaluated 50 panoramic radiographs for the presence or absence of PRs, with and without the aid of AI, using a cross-over design. Diagnostic performance metrics, confidence scores, and clinical decision choices were analyzed. CBCT scans served as the reference standard. Outcomes included sensitivity, specificity, positive and negative predictive values, overall diagnostic accuracy, and area under the ROC and AFROC curves. Statistical analyses were conducted using mixed-effects regression models.

Results: AI assistance significantly improved overall diagnostic accuracy (91.6 % unaided vs. 93.3 % AI-aided; $p <$ 0.001), mainly by reducing false positive diagnoses (false positive rate: 4.3 % unaided vs. 2.0 % AI-aided). Sensitivity remained stable (46.0 % unaided vs. 45.8 % AI-aided). Junior dentists showed the greatest improvements in performance and confidence. AI support shifted treatment decisions toward more conservative approaches.

Conclusions: AI assistance modestly enhanced dentists' diagnostic accuracy for detecting periapical radiolucencies, primarily by decreasing false positive diagnoses. Junior dentists benefited most from AI support. Integration of AI in diagnostic workflows may reduce overtreatment and enhance diagnostic consistency, especially among less experienced clinicians.

Clinical Significance: The integration of AI support in dental diagnostics reduced false positive diagnoses and supported more conservative treatment decisions, particularly benefiting less experienced clinicians. These findings suggest that AI assistance can enhance diagnostic consistency and reduce overtreatment in clinical dental practice.

1. Introduction

Periapical radiolucencies (PRs) are key indicators in dental diagnostics that may signify apical periodontitis and, potentially, endodontic treatment needs. While accurate diagnosis of PRs is essential for proper treatment planning, it remains challenging even for experienced dentists [1,2]. The prevalence of PRs has been estimated to be 5 % at the tooth level [3], which makes it a relatively common finding in oral radiology. While 2D radiographs, such as panoramic or periapical radiographs, are the most commonly used modality in routine dental screening and attentive apical diagnostics (e.g., in symptomatic cases or prior to further interventions), cone beam computed tomography (CBCT) provides markedly improved visualization of PRs, resulting in significantly higher sensitivity compared with 2D radiography. Notably, the routine use of CBCT is constrained by factors including higher exposure to ionizing radiation, higher cost, and limited accessibility [4].

Despite their thorough training, dentists often face challenges in accurately detecting PRs on 2D radiographs due to overlapping

* Corresponding author.
E-mail address: falk.schwendicke@med.uni-muenchen.de (F. Schwendicke).

https://doi.org/10.1016/j.jdent.2025.105868
Received 29 April 2025; Received in revised form 28 May 2025; Accepted 30 May 2025
Available online 3 June 2025

anatomical structures, image quality limitations, and the subtle nature of incipient lesions [5,6]. These diagnostic challenges may lead to both false positive and false negative assessments, potentially resulting in unnecessary treatments and missed pathology, respectively [7]. Closing this gap, artificial intelligence (AI) has shown promise in enhancing diagnostic capabilities in medical imaging, including dental radiography [8]. A recent systematic review and meta-analysis [9] demonstrated AI to have high accuracy for detecting PRs, with a pooled sensitivity of 0.94 (95 % confidence interval (CI): 0.90–0.96) and specificity of 0.96 (95 % CI: 0.91–0.98). Another recent systematic review and meta-analysis also reported high accuracy in PR detection [10]. However, nearly all included studies evaluated the AI performance against a reference standard, usually set by a group of experts who annotated potentially present PRs in each image, rather than assessing its impact on dentists' diagnostic abilities in real-world clinical contexts [11]. The limited number of studies in dentistry and related fields that truly assessed the diagnostic benefit or harm of AI indicates an "AI chasm." This refers to the difference in diagnostic performance between AI-aided and unaided examiners, which tends to be much smaller than anticipated when compared to the standalone diagnostic performance of the AI system against a reference standard [12,13]. Moreover, there is scarce data on how AI impacts examiners' diagnostic confidence and decision-making, as well as whether any of the observed effects differ across examiners with variable levels of experience [14].

In the present randomized controlled trial, we aimed to explore the diagnostic benefit or harm of using a commercial AI application to detect PRs on panoramic radiographs, with corresponding CBCT scans serving as the reference standard. Despite their known limitations in sensitivity and specificity for detecting PRs, we chose to evaluate AI assistance on panoramic radiographs because they are more commonly used in routine dental practice [1,6]. Furthermore, PR detection on panoramic radiographs represents a more challenging diagnostic task where AI assistance might provide greater benefit. We further evaluated potential differences in diagnostic confidence and decision-making in aided versus unaided examiners with varying clinical experience. We hypothesized that utilizing AI would significantly increase examiners' diagnostic accuracy across various experience levels. A confirmation of the hypothesis could potentially reduce the need for additional imaging in some cases.

2. Methods

2.1. Study design

We conducted a randomized, controlled, non-blinded, cross-over superiority trial with a 1:1 allocation ratio. The cross-over approach enabled each participating dentist to act as their own control by randomly assessing half of the radiographs AI-aided and the other unaided.

2.2. Trial registration

The study was prospectively registered in the German Clinical Trials Register (DRKS) under the registration number DRKS00034916 on September 5th, 2024.

2.3. Ethical considerations

This study was conducted in accordance with the Declaration of Helsinki and received ethical approval from the ethics committee of LMU München (protocol number 24-0580). All participating dentists provided written informed consent before enrollment, and patient data in the radiographic materials were anonymized. This was an unblinded study with respect to the intervention, as dentists were necessarily aware of whether they were using AI for each evaluation block. However, partial blinding was maintained as dentists were blinded to the reference standard results, the radiograph selection, and the allocation process. Confidentiality was maintained throughout the study.

2.4. Participant selection

Thirty dentists with varying levels of clinical experience were recruited, ranging from 3 to 23 years (mean, 12.0 years). Experience was categorized as junior (≤10 years, $n = 17$), intermediate (11–15 years, $n = 6$), and senior (>15 years, $n = 7$) to enable subgroup analysis. Participants included 19 males and 11 females, practicing in Germany ($n = 15$) and Turkey ($n = 15$). More than half of the participants were generalists ($n = 19$), while the remainder were specialists ($n = 11$). The specialists comprised endodontists ($n = 5$), periodontists ($n = 3$), oral radiologists ($n = 1$), and conservative dentistry specialists ($n = 2$).

2.5. Data and reference test

Fifty panoramic radiographs were selected from a database of patients who had been examined with both panoramic radiography (exposure: 73 kV, 8 mA, 11.9 s; dose: 91 mGy cm^2) and CBCT (exposure: 90 kV, 2.5 mA, 15 s; dose: 1178.74 mGy cm^2; acquisition mode: High contrast 8 × 9). Both modalities were acquired using CS 8200 3D (Carestream Dental, Atlanta, GA, USA). For each patient, both imaging modalities were acquired within a maximum interval of 7 days.

The CBCT scans were used as the reference standard; they were not submitted to AI analysis or evaluated for AI assistance. Two experts in CBCT diagnostics independently evaluated each CBCT to determine the presence or absence of PRs. In cases of disagreement, consensus was reached through joint re-evaluation. Raw interrater agreement was 90.2 %, with a Cohen's Kappa of 0.74 (95 % CI: 0.68–0.80), indicating substantial agreement. The reference test was established before and independently of the index test (see below).

According to the reference standard, 106 teeth (8.5 %) exhibited PRs, whereas 1145 (91.5 %) did not. PRs were distributed across the 50 radiographs with an average of 2.1 PRs (0–9) per radiograph. The majority of radiographs (72 %) contained 1–3 PRs, while 14 % had no PR and 14 % had four or more PRs. Approximately 60 % of the PRs were located in the maxilla and 40 % in the mandible. Molars, premolars, and anterior teeth accounted for 45 %, 30 %, and 25 % of the PRs, respectively, with first molars and first premolars being more commonly affected than other teeth.

2.6. AI

An AI application designed for 2D radiographic analysis used in previous evaluations [15–17], dentalXrai Pro (version 3.0.9, dentalXrai, Berlin, Germany), was used in this study. The software utilizes a convolutional neural network architecture based on a modified ResNet-50 backbone with a feature pyramid network for multi-scale feature extraction. The system was trained on over 100,000 annotated dental radiographs and was locked (not fine-tuned) for this study. The interface presents detected findings through color-coded overlays (Fig. 1) and allows toggling between standard and high-sensitivity modes, which dentists could adjust according to their preferences. The system operates as a supportive diagnostic tool that highlights suspected areas. Previous data generated on an independent dataset demonstrated a sensitivity of 0.65 and a specificity of 0.87, respectively, for detecting PRs using this application [18].

2.7. Randomization and study procedure

The study employed a block-randomized cross-over design. The 50 panoramic radiographs were divided into 10 blocks, each containing five radiographs. During the study, dentists stepwise randomly selected one block of radiographs and one decision slip concealed in an opaque envelope to determine whether to diagnose AI-aided or unaided. This

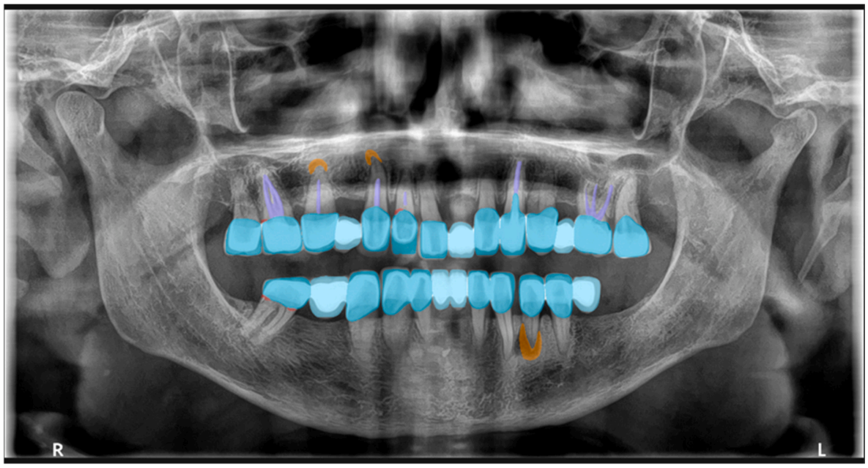

Fig. 1. Screenshot of dentalXrai Pro software interface showing a panoramic radiograph with AI-assisted findings. Color-coded highlights indicate crowns (blue), obturated root canals (purple), caries (red), and suspected periapical radiolucencies (orange).

process was repeated until each dentist had evaluated all blocks, as shown in Fig. 2.

The randomization process ensured that the radiographic blocks and the conditions (AI-aided vs. unaided) were assigned randomly. Each radiograph contained an average of 22.5 [13-28] assessable teeth, with the remainder being missing. To avoid selection bias, the radiographs were randomly assigned to blocks without any stratification by PR prevalence or diagnostic challenge, maintaining the integrity of the randomization process. No washout period was implemented between evaluations of different blocks since this design allows for immediate alternation between conditions, reflecting realistic clinical usage patterns.

For each evaluation, dentists recorded:

1. Diagnosis: determine if a PR was present or absent for each tooth.
2. Confidence: rate confidence in diagnosis on a scale from 1 to 5 (1 = very low, 5 = very high)
3. Decision: 0 = no further intervention, 1 = further diagnostics (e.g. CBCT), 2 = root-canal treatment or other endodontic (invasive) therapy

2.8. Outcomes

The primary outcome was diagnostic accuracy and additional performance metrics, including:

- Sensitivity and specificity
- Positive and negative predictive values and false positive rates
- Area under the receiver operating characteristic (ROC) curve (AUC)

Secondary outcomes included the following:

- Confidence scores (mean and distribution)
- Alternative free-response ROC (AFROC) analysis
- Decisions (distribution across categories)
- Impact of dentist experience on the above measures
- Impact of specialization (generalist vs. specialists) on diagnostic performance
- Impact of country of practice (Germany vs. Turkey) on diagnostic performance

2.9. Sample size calculation

The sample size was calculated using Hillis et al.'s Multiple Readers and Multiple Cases (MRMC) approaches [19] with correlation values [20] as described in Obuchowski [21] when using large estimates of intra-observer and inter-observer variability. Estimated effect sizes were drawn from Nardi et al. [22]. Sample size needs to consider clustering effects, determined via intra-class correlations found in Meinhold et al. [23]. Power was set at 95 % with a two-sided alpha of 0.05. This calculation demonstrated that 30 dentists, each evaluating 50 radiographs (25 AI-aided and 25 unaided), would provide 95 % power to detect the expected effect size. Our sample size calculation was intended to detect the overall effect of AI assistance on diagnostic accuracy. While we conducted subgroup analyses by experience level, we acknowledge that the smaller sample sizes in the intermediate ($n = 6$) and senior ($n = 7$) groups limit statistical power for detecting interaction effects.

2.10. Statistical analysis

Diagnostic accuracy metrics, including sensitivity, specificity, positive predictive value, negative predictive value, and overall accuracy, were calculated with 95 % CIs. AUC values were derived from the ROC curve analysis. Additionally, we conducted the AFROC analysis to comprehensively assess diagnostic performance based on confidence scores.

For comparing AI-aided vs. unaided conditions, we used:

- Independent samples *t*-tests for performance metrics at the dentist level (AUC, AFROC parameters),
- McNemar's test for paired binary outcomes (correct/incorrect diagnosis),
- Chi-square tests for categorical variables (treatment decisions),
- Mixed-effects linear regression for confidence scores.

To account for the hierarchical nature of the data, we employed mixed-effects regression models with a three-level nested structure: teeth (level 1) within radiographs (level 2) within dentists (level 3). For diagnostic accuracy, we used mixed-effects logistic regression; for confidence scores, we utilized mixed-effects linear regression; and for treatment choices, we applied mixed-effects ordinal logistic regression [24,25]. For the mixed-effects regression models, we used generalized linear mixed models for binary outcomes (diagnostic accuracy) using the *glmer* function from the *lme4* package in R with a binomial distribution and logit link function. For confidence scores, we used linear mixed models with the lmer function, and for treatment decisions, we employed cumulative link mixed models with the *clmm* function from the ordinal package. In all models, we specified AI assistance, experience level, and country as fixed effects. Random effects included dentist, radiograph, and tooth to account for the three-level nested structure of

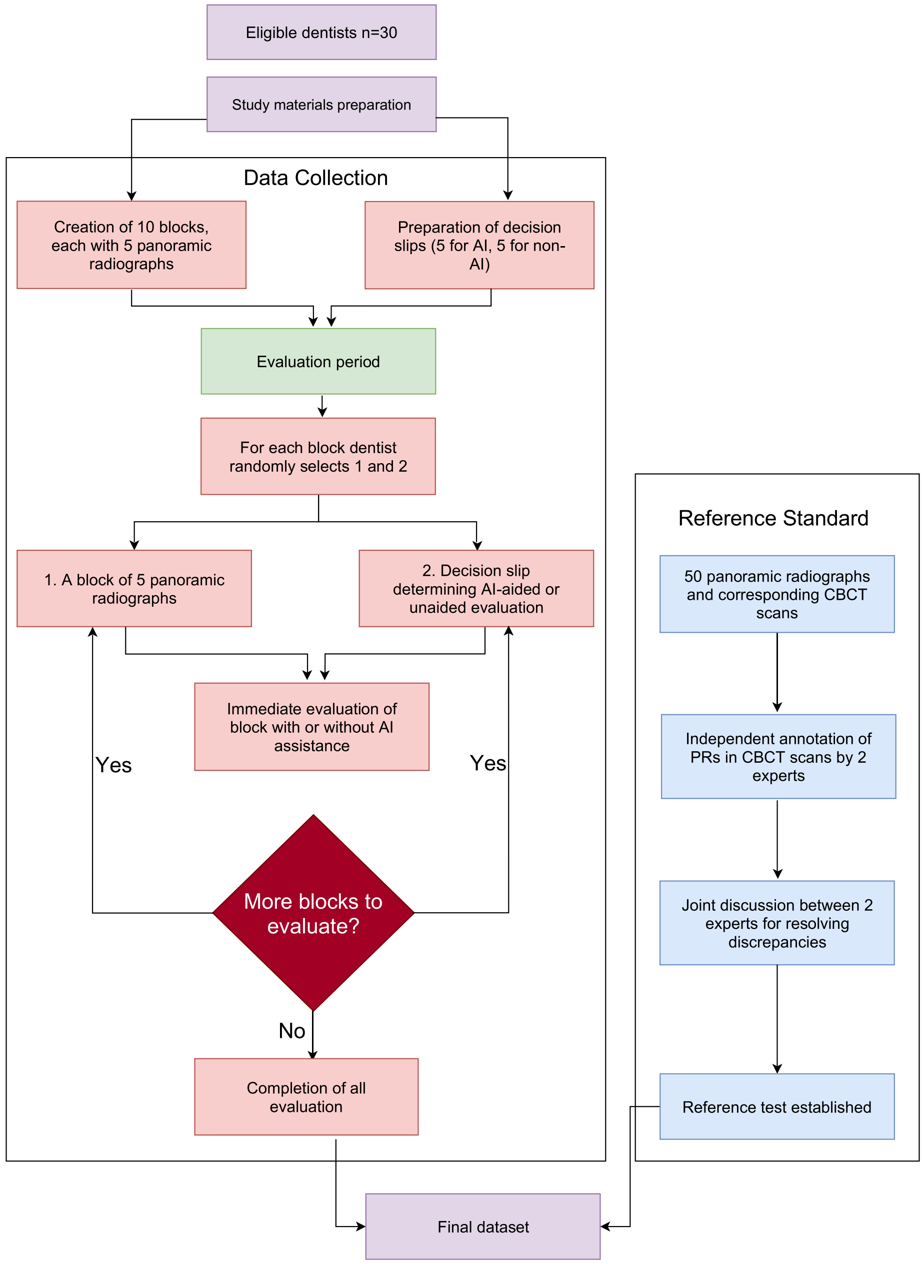

Fig. 2. Study design flow diagram. The flowchart illustrates the methodology for evaluating AI-assisted detection of periapical radiolucency using panoramic radiographs. Thirty eligible dentists evaluated 10 blocks of 5 panoramic radiographs each, with randomized AI assistance. The reference standard was established through expert annotation of corresponding CBCT scans.

our data: teeth (level 1) within radiographs (level 2) within dentists (level 3). This structure appropriately models the hierarchical nature of our dataset and controls for clustering effects that could otherwise inflate Type I error rates [26]. We also ran models without interaction terms, which showed similar or slightly better fit according to AIC values and R-squared metrics. Subgroup analyses by dentist experience, specialization, and country were incorporated using interaction terms in the respective models.

Shapiro-Wilk tests were performed to assess the normality of continuous variables, including confidence scores and AUC values [27]. When normality assumptions were violated ($p < 0.05$), non-parametric alternatives (Wilcoxon rank-sum tests) were employed [28]. For the mixed-effects models, we examined residual diagnostics to confirm appropriate model fit, including QQ plots and residual versus fitted value plots [25]. All analyses were performed using R (version 4.4.3). Statistical significance was set at a two-sided alpha level of 0.05.

3. Results

3.1. Descriptive statistics

The study involved 30 dentists who evaluated 50 panoramic radiographs (in total, 1251 teeth) each, resulting in 37,530 tooth assessments.

Table 1
Absolute numbers of true positives, false positives, true negatives, and false negatives for both AI-aided and unaided conditions across all experience levels. AI = Artificial intelligence.

Assessment	True Positive	False Positive	True Negative	False Negative	Total
Unaided	704	746	16,561	785	18,877
AI-aided	750	345	16,608	851	18,623

Of these assessments, 18,636 (49.66 %) were AI-aided, while 18,894 (50.34 %) were unaided. Table 1 presents the absolute numbers of true positives, false positives, true negatives, and false negatives for both AI-aided and unaided conditions across all experience levels.

3.2. Overall diagnostic performance

AI assistance significantly improved overall diagnostic accuracy from 91.6 % (95 % CI: 91.1–92.1 %) to 93.3 % (95 % CI: 92.9–93.7 %) ($p < 0.001$). The most significant change was in specificity, which increased from 95.7 % (95 % CI: 95.4–96.0 %) unaided to 98.0 % (95 % CI: 97.8–98.2 %) AI-aided, while sensitivity remained relatively stable at 46.0 % (95 % CI: 42.7–49.3 %) unaided compared to 45.8 % (95 % CI: 42.6–49.0 %) AI-aided.

AI-aided ROC AUC showed a modest yet statistically significant improvement (AUC: 0.719 ± 0.018 vs. 0.708 ± 0.017; $p = 0.042$) (Fig. 3). AFROC AUC increased across nearly all experience levels with the aid of AI. Overall, it rose from 0.701 ± 0.027 unaided to 0.710 ± 0.026 AI-aided, although this difference was not statistically significant ($p > 0.05$). The most notable effect was the reduction of false positive diagnoses, with the false positive rate decreasing from 4.3 % unaided to 2.0 % AI-aided. This translated to 401 fewer false positive diagnoses across all evaluations.

The positive predictive value (PPV) showed significant improvement with AI-aided, rising to 68.6 % (95 % CI: 65.1–72.1 %) compared to 48.7 % (95 % CI: 45.1–52.3 %). Meanwhile, the negative predictive value (NPV) remained stable at 94.9 % (95 % CI: 94.6–95.2 %) compared to 95.2 % (95 % CI: 94.9–95.5 %) unaided.

3.3. Impact of experience

The benefits of AI assistance varied significantly depending on the level of experience of dentists. Junior dentists (≤10 years of experience) exhibited the most remarkable improvement with the aid of AI, with the AUC increasing from 0.70 to 0.74. Intermediate dentists (11–15 years of experience) demonstrated a modest improvement, while senior dentists (>15 years of experience) did not benefit from AI (Table 2). A similar pattern was evident in diagnostic accuracy and false positive rates (Fig. 4).

3.4. Impact of country of practice and specialty

In additional subgroup analyses, diagnostic accuracy did not differ significantly between German and Turkish dentists, or between general dentists and specialists, regardless of AI use. Among dentists using AI, the average accuracy was 93.7 % (93.0–94.4 %) for German dentists and 93.6 % (93.2–94.0 %) for Turkish dentists. Without AI, German dentists achieved 91.8 % (91.2–92.4 %), while Turkish dentists reached 92.0 % (91.6–92.5 %). Similarly, specialists achieved an average accuracy of 93.9 % (93.3–94.6 %) with AI and 92.2 % (91.6–92.9 %) without AI. General dentists performed comparably, with 93.5 % (93.0–94.0 %) AI-aided and 91.7 % (91.2–92.2 %) unaided. No statistically significant

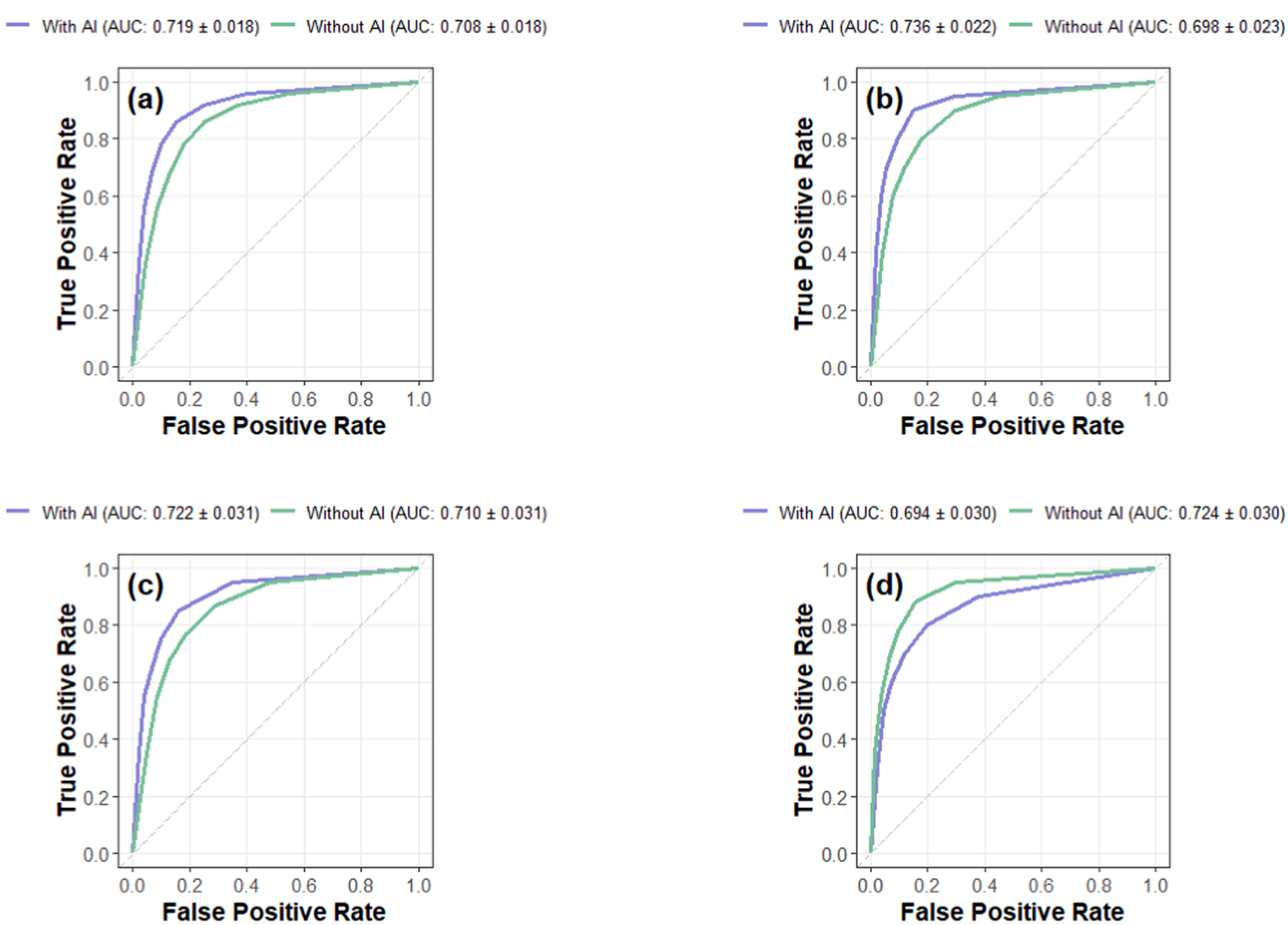

Fig. 3. Receiver Operating Characteristic (ROC) curves plot true positive rate against false positive rate to evaluate diagnostic performance, with the diagonal line representing random chance (AUC = 0.5). AI assistance provided the greatest improvement for junior dentists, while senior and intermediate dentists showed more modest benefits. (a) Overall performance across all dentists, (b) Junior dentists (≤10 years of experience), (c) Intermediate dentists (11–15 years of experience), (d) Senior dentists (>15 years of experience). Table 2 shows the corresponding Area Under the Curve (AUC) values and differences.

Table 2
Diagnostic performance with and without AI assistance, presented overall and by dentist experience level. 95 % confidence intervals are shown in parentheses for AI-aided and Unaided columns; ranges are shown for the Difference column. AI = Artificial Intelligence; AUC = area under the receiver operating characteristic curve; FPR = false positive rate; NPV = negative predictive value; PPV = positive predictive value. * Statistically significant difference ($p < 0.05$).

Experience	Metric	AI-aided	Unaided	Difference
Overall (n = 30)	Sensitivity (%)	45.8 (41.7–49.9)	46.0 (41.9–50.1)	−0.2 (−5.5; 5.1)
	Specificity (%)	98.0 (97.4–98.6)	95.7 (94.9–96.5)	2.3 (1.4; 3.2)*
	PPV (%)	68.6 (62.8–74.4)	48.7 (42.1–55.3)	19.9 (11.7; 28.1)*
	NPV (%)	94.9 (94.3–95.5)	95.2 (94.6–95.8)	−0.3 (−1.1; 0.5)
	Accuracy (%)	93.3 (92.6–94.0)	91.6 (90.8–92.4)	1.7 (1.0; 2.4)*
	FPR (%)	2.0 (1.8–2.2)	4.3 (4.0–4.6)	−2.3 (−2.6; −2.0)*
	AUC	0.7 (0.7–0.7)	0.7 (0.7–0.7)	0.011 (0.001; 0.02)*
Junior (n = 17)	Sensitivity (%)	49.1 (44.2–54.0)	44.0 (39.1–48.9)	5.1 (0.9; 9.3)*
	Specificity (%)	98.1 (97.5–98.7)	95.6 (94.6–96.6)	2.5 (1.5; 3.5)*
	PPV (%)	70.0 (63.5–76.5)	48.8 (41.2–56.4)	21.2 (12.0; 30.4)*
	NPV (%)	95.5 (94.8–96.2)	94.7 (93.9–95.5)	0.8 (−0.1; 1.7)
	Accuracy (%)	94.0 (93.2–94.8)	91.2 (90.2–92.2)	2.8 (1.7; 3.9)*
	FPR (%)	1.9 (1.6–2.2)	4.4 (4.0–4.8)	−2.5 % (−2.9; −2.1)*
	AUC	0.7 (0.7–0.8)	0.7 (0.7–0.7)	0.04 (0.02; 0.06)
Intermediate (n = 6)	Sensitivity (%)	42.6 (36.5–48.7)	46.9 (40.8–53.0)	−4.3 (−11.0; 2.4)
	Specificity (%)	98.1 (97.3–98.9)	95.1 (93.8–96.4)	3.0 (1.7; 4.3)*
	PPV (%)	68.6 (60.2–77.0)	46.1 (37.1–55.1)	22.5 (10.9; 34.1)*
	NPV (%)	94.6 (93.6–95.6)	95.3 (94.4–96.2)	−0.7 (−1.9; 0.5)
	Accuracy (%)	93.2 (92.0–94.4)	91.0 (89.6–92.4)	2.0 (−0.5; 3.5)
	FPR (%)	1.9 (1.5–2.3)	4.9 (4.3–5.5)	−3.0 (−3.6; −2.4)*
	AUC	0.7 (0.7–0.8)	0.7 (0.7–0.7)	0.01 (−0.01; 0.03)
Senior (n = 7)	Sensitivity (%)	41.4 (35.5–47.3)	50.8 (44.9–56.7)	−9.4 (−16.0; −2.8)*
	Specificity (%)	97.5 (96.5–98.5)	95.2 (95.0–97.4)	2.3 (1.0; 3.6)*
	PPV (%)	65.1 (56.3–73.9)	50.8 (42.0–59.6)	14.3 (3.0; 25.6)*
	NPV (%)	93.7 (92.6–94.8)	96.2 (95.2–97.1)	−2.5 (−3.8; −1.2)*
	Accuracy (%)	91.9 (90.5–93.3)	93.0 (91.8–94.2)	−1.1 (−2.7; 0.5)
	FPR (%)	2.5 (2.0–3.0)	3.8 (3.2–4.4)	−1.3 (−1.9; −0.7)*
	AUC	0.7 (0.7–0.7)	0.7 (0.7–0.8)	−0.03 (−0.05; −0.01)

differences were found across these groups ($p > 0.05$ in all cases), indicating that the benefit of AI was consistent regardless of country or specialization.

3.5. Diagnostic confidence

Dentists reported slightly higher confidence when being assisted by AI (mean score 3.59 ± 1.06) compared to when not being assisted (3.53 ± 1.10), though this finding was not statistically significant ($p > 0.05$). When AI assistance was utilized, confidence scores were more consistent, with a lower standard deviation.

Mixed-effects linear regression indicated a positive effect of AI on confidence scores (+0.055 points, 95 % CI: −0.023 to 0.134) after accounting for dentist-specific variation; however, this result neither reached statistical significance ($p > 0.05$). Among all dentists, 19 (63.3 %) showed higher mean confidence scores when using AI, while 11 (36.7 %) reported lower scores.

3.6. Decision-making

AI assistance was linked to a shift toward less interventionist approaches. With AI, dentists recommended a "wait and see" approach in 23.0 % of cases compared to 20.7 % without AI. Further invasive treatments were recommended in 26.8 % of cases with AI, compared to 29.0 % without AI. The percentage of recommendations for further diagnostics remained virtually unchanged at 50.2 % with AI versus 50.3 % without AI. While consistent, AI's effect on treatment decisions did not achieve statistical significance ($p = 0.265$); this was confirmed by mixed-effects ordinal logistic regression (OR: 0.91, 95 % CI: 0.82–1.01, $p = 0.062$) (Table 3).

4. Discussion

AI is entering healthcare fast, with dentistry being no exception. In clinical practice, AI-assisted detection of PRs on panoramic radiographs may serve as an effective initial screening tool [29]. When the AI system flags a potential PR, clinicians review the findings and may consider follow-up imaging, such as periapical radiographs or CBCT, to confirm the diagnosis before proceeding with treatment planning [4]. In this role, AI is expected to enhance sensitivity, particularly supporting junior practitioners [30]. While there is a large and growing body of evidence demonstrating the theoretical diagnostic performance of AI in dentistry, mainly in image analysis, there is a significant gap in assessing the true benefits of AI assistance for practitioners' diagnostic accuracy, confidence, and decision-making.

The present study, for the first time, evaluated these aspects in a randomized controlled design, with dentists detecting PRs on panoramic radiographs with and without AI assistance. We found AI to modestly but consistently improve dentists' diagnostic accuracy, primarily by reducing false positive diagnoses, with effects varying significantly based on the dentist's experience. Similarly, AI increased the diagnostic confidence of less experienced dentists, but not necessarily that of experienced dentists. AI also tended to drive dentists towards less invasive decisions made based on the assessment of the radiograph, potentially reducing both overtreatment and unnecessary radiation exposure from additional imaging [31].

The overall diagnostic accuracy for detecting PR (AI-aided AUC: 0.719 [95 % CI: 0.701–0.737] vs. unaided AUC: 0.708 [95 % CI: 0.691–0.725]) was lower than previously reported [32], likely routed in different radiograph modalities being analyzed. In periapical radiographs, for example, the resolution will be higher, and tooth overlap will be lower than in panoramic radiographs, affecting dentists' accuracy in detecting PRs. Moreover, previous studies – summarized in a recent meta-analysis [10] – reported higher sensitivity (0.93) and specificity (0.85) for deep learning models detecting PRs than our evaluation. Notably, these studies primarily evaluated AI performance in isolation rather than in real-world usage by dentists. Furthermore, most previous studies [33] relied on the same type of radiograph for generating the reference test. In contrast, we used CBCT scans to establish the reference standard, which is noteworthy because dentists detect approximately 25–30 % more PRs on CBCT scans than on 2D radiographs [1]. On the one hand, this increases the validity of our reference standard; on the other hand, it challenges examiners, as both humans and AI may have faced PRs that were not detectable on 2D radiographs, lowering their sensitivity. We presume that this also contributed to the observation that sensitivity remained virtually unchanged even with AI assistance.

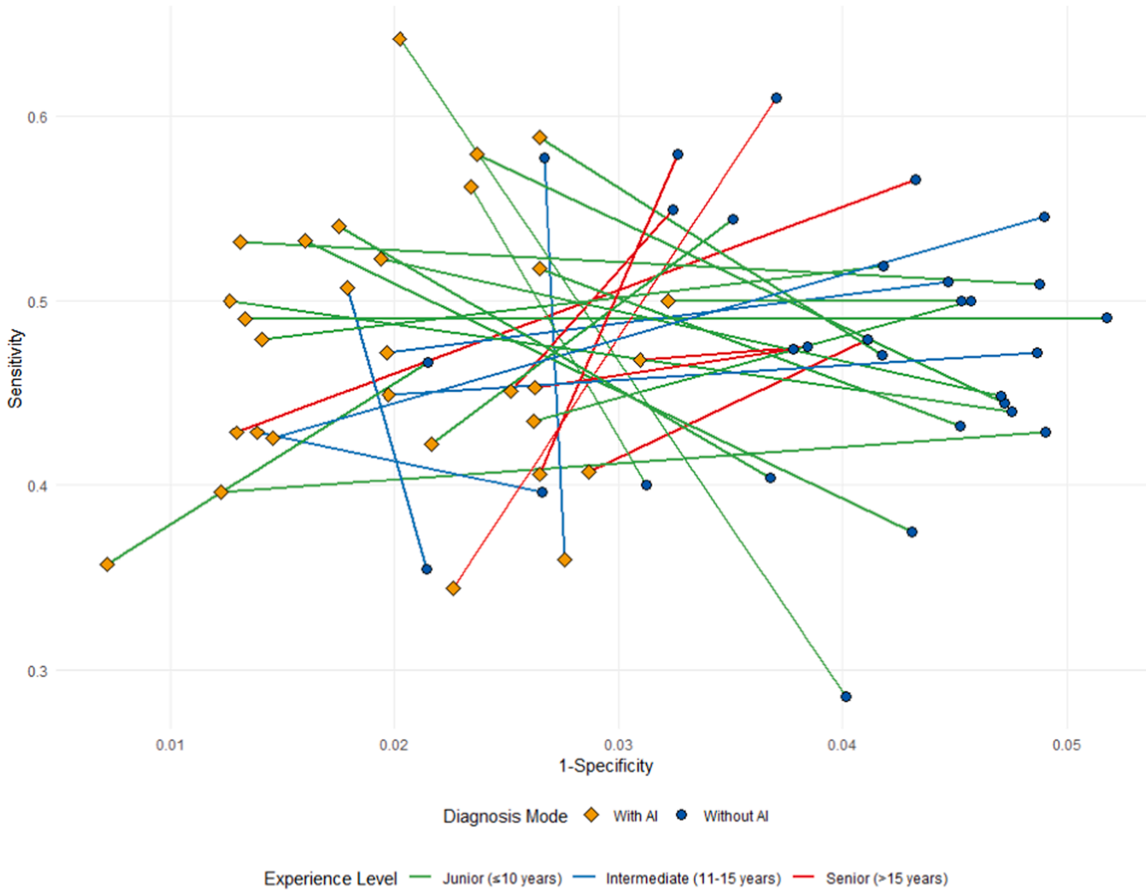

Fig. 4. Arrows indicate a change in performance metrics for individual dentists when using AI, with each arrow connecting the same dentist's performance under both conditions. Diamond markers represent the average sensitivity and 1-specificity values across all dentists within each diagnostic mode (with/without AI). Line colors reflect dentists' experience levels: green for junior (≤10 years), blue for intermediate (11–15 years), and red for senior (>15 years) practitioners. The figure demonstrates that AI assistance generally improved specificity with minimal impact on sensitivity, though individual responses varied considerably across experience levels. Most arrows point upward and slightly rightward, indicating that AI typically decreased false positive rates while maintaining acceptable true positive detection.

Table 3
Summary of mixed-effects regression models assessing the impact of AI assistance on diagnostic accuracy, confidence, and treatment decisions.

Variable	Category	Model 1		Model 2		Model 3*	
		aOR (95 % C.I.)	p-value	β (95 % C.I.)	p-value	aOR (95 % C.I.)	p-value
Dependent variable:	No	Ref.	–	Ref.	–	Ref.	–
AI Assistance	Yes	1.36 (1.25; 1.48)	<0.001	0.04 (−0.03; 0.11)	0.249	0.86 (0.74; 1.00)	<0.05
Dependent variable:	0	n/a	n/a	n/a	n/a	Ref.	–
Treatment decision	1	n/a	n/a	n/a	n/a	0.22 (0.16; 0.30)	<0.001
	2	n/a	n/a	n/a	n/a	2.52 (1.83; 3.46)	<0.001
Experience level	Junior	Ref.	–	Ref.	–	Ref.	–
	Intermediate	0.94 (0.84; 1.04)	0.208	−0.05 (−0.14; 0.04)	0.235	0.84 (0.69; 1.02)	0.081
	Senior	0.96 (0.87; 1.06)	0.468	0.04 (−0.05; 0.12)	0.412	0.92 (0.77; 1.11)	0.404
Country	Germany	Ref.	–	Ref.	–	Ref.	–
	Turkey	1.00 (0.92; 1.09)	0.951	−0.03 (−0.10; 0.04)	0.363	0.89 (0.77; 1.03)	0.128

Model 1: Mixed-effects logistic regression assessing the effect of AI assistance on diagnostic accuracy.
Model 2: Mixed-effects linear regression assessing the effect of AI assistance on confidence.
Model 3: Mixed-effects ordinal regression assessing the effect of AI assistance on treatment decisions.
* AI Assistance was considered as a covariate in Model 3.
aOR: Adjusted odds ratios; β: Coefficient estimate.

In terms of accuracy, the absolute improvement yielded by AI assistance was modest. This was not surprising, as the discrepancy between AI performance in diagnostic accuracy studies (where AI is compared to a reference standard not established in clinical settings) and clinical application, is well documented in dentistry [15,34], radiology [35], and other medical fields [36,37]. While AI tools frequently demonstrate high performance on curated datasets and retrospective analyses, their benefit tends to diminish in real-world settings due to factors such as workflow misalignment, trust issues, and variability in clinical decision-making [37,38].

Although AI aid did not markedly improve accuracy, its clinical significance should be considered in context. Table 1 shows that AI helped dentists avoid misdiagnosing normal findings as pathological; of the 37,500 tooth assessments, the number of false positives decreased from 746 to 345 with AI assistance. The 53.5 % decrease in false positive rate presents one of the main clinical implications of this study, as overdiagnosis is a well-known problem in dental practice, often resulting in unnecessary treatments that incur costs and possible complications. While overtreatment in our context may involve relatively minor interventions such as additional radiographic imaging, reducing false positives still represents meaningful clinical value by minimizing patient exposure to unnecessary procedures and associated healthcare costs [39, 40].

Our discovery that AI cut false positive diagnoses by over half directly tackles this issue. This reduction in false positives may be explained by mechanisms such as attention anchoring [41], where AI highlights suspect regions and prevents overinterpretation of ambiguous areas. Additionally, cognitive off-loading [42] allows dentists to delegate part of the visual pattern recognition to AI, reducing mental burden and enhancing evaluation of critical areas. Moreover, AI's consistent application of diagnostic thresholds may reduce interobserver variability and promote more standardized diagnostic decisions.

The impact of AI assistance varied notably depending on the dentists' level of experience. Junior dentists showed significant improvements in diagnostic performance when supported by AI, while those with intermediate experience exhibited more modest gains, and senior dentists did not at all perform better with AI support. These findings may reflect how clinical experience shapes interaction with AI. Less experienced dentists, who may lack diagnostic confidence, tend to rely more on AI suggestions, enhancing their accuracy and realizing the theoretical benefits of AI (measured in the discussed diagnostic accuracy studies). In contrast, experienced practitioners might be less influenced by AI or even experience conflict when AI contradicts their internalized patterns. In some cases, AI cues may disrupt intuitive reasoning or reinforce errors through confirmation bias, especially when initial impressions are unconsciously validated by matching AI outputs [43]. This inverse relationship between clinical experience and the benefit of AI suggests that while AI can be a valuable aid, it may, in some cases, interfere with the well-developed pattern recognition skills of more experienced clinicians. Supporting this, in everyday practice, junior clinicians seem to embrace AI as a safety net, routinely consulting AI recommendations and frequently following them [44] while more experienced practitioners tend to develop more pragmatic approaches, using AI selectively for complex or uncertain cases [45] or as a confirmatory tool after forming their own conclusions [46]. Generally, experience has been found to modulate the benefit-harm ratio of AI in medical diagnostics. Majkowska et al. [30]. demonstrated in their study that AI assistance provided greater benefits to less experienced radiologists when detecting lung nodules on chest radiographs. Similarly, Bejnordi et al. [47]. found that AI algorithms could improve diagnostic consistency among pathologists with varying levels of experience.

Previous studies in medicine have also shown that decision support systems can influence clinical decision-making, often leading to more evidence-based and standardized care [48,49]. In this study, the slight increase in diagnostic confidence observed with AI support and a shift toward more conservative treatment recommendations suggests that AI may also contribute to greater consistency in clinical decision-making. The finding contrasts with the study by Mertens et al. [15]. on caries detection, where the use of AI led to more invasive treatment planning. This discrepancy likely reflects differences in the use case. First, while models for caries detection tend to increase sensitivity [50] (and hence the number of positive cases), in our study, the number of PRs detected with AI decreased from 1450 to 1095 (Table 1), indicating an increase in specificity and precision. As a consequence of the perceived absence of PRs, dentists seemed more inclined to choose a wait-and-see approach rather than immediately recommending any interventions. This tendency was particularly evident in cases where dentists initially reported low confidence, indicating that the influence of AI is most significant in situations characterized by diagnostic uncertainty or hesitation.

The shift toward more conservative treatment decisions with AI support warrants further investigation. Unlike caries, which prompts intervention due to its progressive nature, the decision-making process related to PRs is more complicated, as they may or may not gradually diminish after treatment. Moreover, the observed association of AI-assisted diagnoses with more conservative treatment approaches may reflect the inherent limitations of detecting PRs on panoramic radiographs. It is possible that increased specificity helped dentists better distinguish between clear pathology requiring intervention and equivocal findings better suited for monitoring [51]. However, we cannot exclude that this conservative trend also reflects dentists' awareness of the diagnostic limitations of panoramic radiographs for confirming PRs [22]. Future studies comparing treatment decisions across different imaging modalities, with and without AI assistance [52,53], are needed to disentangle these contributing factors. Furthermore, in the future, the human-AI interaction should be better personalized, and interface designs should be prioritized to align with dentists' natural visual workflows, as dentists have been shown to use AI selectively [16] and frequently toggle between views, creating workflow interruptions. Optimizing interfaces to minimize attention shifts between images and AI findings could reduce this issue and enhance human-AI interaction. Additionally, adopting educational approaches where AI explanations supplement detections [11] can help dentists develop improved pattern recognition rather than creating dependence on AI assistance.

This study has several strengths. It was conducted in real-world clinical settings, achieving high ecological validity by evaluating dentists using their everyday diagnostic equipment and workflows. While dentists did not use the tool in their daily workflow on real patients, this caveat had to be accepted as part of our study design. Besides that, the randomized cross-over design enabled us to control individual dentist factors and directly compare their performance with and without AI support. We demonstrated that AI assistance may influence clinical decision-making, encouraging more conservative approaches, an aspect rarely examined in dental AI studies [15,54]. Additionally, our multilevel modeling approach rigorously accounts for the hierarchical structure of dental radiographic data, addressing a crucial methodological gap in earlier studies that typically use simpler analytical approaches [55,56]. By utilizing CBCT as the reference standard, this study also established a strong benchmark for evaluating diagnostic accuracy. Finally, with 30 dentists and more than 37,000 assessment points, our study had adequate statistical power to identify significant differences. However, there were some limitations as well, including a risk of selection bias which could limit the generalizability of our results. The radiographs used in this study were sourced from patients who had undergone both panoramic radiography and CBCT imaging, which may represent a specific subgroup with a higher likelihood of underlying pathology. Moreover, dentists were aware that they participated in a clinical trial, which may have increased their diagnostic diligence, but also how they framed their decisions. As outlined, assessing AI in routine use would address this, but a randomized design is likely unfeasible in this case. Lastly, the generalizability of our results may be further limited by the radiographs that were sourced from a single clinical center using one imaging system and using a single AI application. The presented outcomes may not extend to other AI technologies with different algorithms, training datasets, or performance characteristics.

Based on our findings and the identified limitations, we propose several further directions for future research: Studies examining the impact of AI-assisted diagnosis on treatment success rates, complications, and patient satisfaction seem relevant to gauge the true, long-term impact of diagnostic AI in dentistry. Similarly, evaluating the economic implications of AI implementation, including potential savings from reduced overtreatment, should be considered. Future studies should also include a more detailed characterization of lesion size and an analysis of how anatomical location affects diagnostic performance. Additionally larger and more balanced experience subgroups would be suitable be confirm our findings regarding the impact of AI across experience levels. Such efforts could provide valuable insights for optimizing AI systems and refining clinical protocols tailored to different tooth regions. Lastly, studies investigating how AI systems can be adapted to provide optimal support for dentists with different levels of experience and in different environments or contexts seem warranted as well, as research examining whether ongoing use of AI applications leads to improved diagnostic capabilities even when AI is not available.

5. Conclusion

AI assistance modestly enhanced dentists' diagnostic accuracy for detecting PRs on panoramic radiographs, mainly by reducing false positive diagnoses by over 50 %. Benefits varied by experience level, with junior dentists showing the most significant improvements. While our findings suggest AI can improve dental care by minimizing over-diagnosis and standardizing diagnostic quality, several limitations must be considered in clinical translation, including the inherent limitations of panoramic radiography for PR diagnosis, the single-center design, and the use of one specific AI system. It is essential to recognize that AI assistance will not always be beneficial, and clinicians must maintain critical judgement in determining when and how to incorporate AI recommendations into their diagnostic workflow. Future multi-center studies with diverse imaging modalities and AI systems are needed to validate these findings across broader clinical contexts.

Ethics

The study was conducted with the approval of the Ethics Committee of LMU München (Project Number: 24–0580) on August 14, 2024.

CRediT authorship contribution statement

Utku Pul: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Antonin Tichy:** Writing – review & editing, Validation, Supervision, Methodology. **Vinay Pitchika:** Writing – review & editing, Validation, Formal analysis, Data curation. **Falk Schwendicke:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors have no conflicts of interest to declare. Prof. Dr. Falk Schwendicke is a co-founder of dentalXrai, a startup focusing on dental radiograph analytics using artificial intelligence. However, the present work was conducted independently of this affiliation.

Acknowledgements and Funding

This study was supported by the German Research Foundation DFG KR 5457/1-1.

References

[1] K. Leonardi Dutra, et al., Diagnostic accuracy of cone-beam computed tomography and conventional radiography on apical periodontitis: a systematic review and meta-analysis, J. Endod. 42 (3) (2016) 356–364.
[2] C. Kruse, et al., Diagnostic accuracy of cone beam computed tomography used for assessment of apical periodontitis: an ex vivo histopathological study on human cadavers, Int. Endod. J. 52 (4) (2019) 439–450.
[3] J.G. Pak, et al., Prevalence of periapical radiolucency and root canal treatment: a systematic review of cross-sectional studies, J. Endod. 38 (9) (2012) 1170–1176.
[4] S. Patel, et al., European Society of Endodontology position statement: the use of CBCT in endodontics, Int. Endod. J. 47 (6) (2014) 502–504.
[5] I.B. Bender, S. Seltzer, Roentgenographic and direct observation of experimental lesions in bone: II. 1961, J. Endod. 29 (11) (2003) 707–712, discussion 1.
[6] C. Estrela, et al., Accuracy of cone beam computed tomography and panoramic and periapical radiography for detection of apical periodontitis, J. Endod. 34 (3) (2008) 273–279.
[7] A. Balasundaram, et al., Comparison of cone-beam computed tomography and periapical radiography in predicting treatment decision for periapical lesions: a clinical study, Int. J. Dent. 2012 (2012) 920815.
[8] F. Schwendicke, et al., Convolutional neural networks for dental image diagnostics: a scoping review, J. Dent. 91 (2019) 103226.
[9] U. Pul, F. Schwendicke, Artificial intelligence for detecting periapical radiolucencies: a systematic review and meta-analysis, J. Dent. 147 (2024) 105104.
[10] S. Sadr, et al., Deep Learning for detection of periapical radiolucent lesions: a systematic review and meta-analysis of diagnostic test accuracy, J. Endod. 49 (3) (2023) 248–261, e3.
[11] F. Schwendicke, et al., Artificial intelligence in dentistry: chances and challenges, J. Dent. Res. 99 (7) (2020) 769–774.
[12] M.A. Gianfrancesco, et al., Potential biases in machine learning algorithms using electronic health record data, JAMA Intern. Med. 178 (11) (2018) 1544–1547.
[13] H. Alami, et al., Artificial intelligence and health technology assessment: anticipating a new level of complexity, J. Med. Internet. Res. 22 (7) (2020) e17707.
[14] P. Tschandl, et al., Human-computer collaboration for skin cancer recognition, Nat. Med. 26 (8) (2020) 1229–1234.
[15] S. Mertens, et al., Artificial intelligence for caries detection: randomized trial, J. Dent. 115 (2021) 103849.
[16] N. Castner, et al., Expert gaze as a usability indicator of medical AI decision support systems: a preliminary study, NPJ Digital Med. 7 (1) (2024) 199.
[17] A.G. Cantu, et al., Detecting caries lesions of different radiographic extension on bitewings using deep learning, J. Dent. 100 (2020) 103425.
[18] T. Ekert, et al., Deep learning for the radiographic detection of apical lesions, J. Endod. 45 (7) (2019) 917–922, e5.
[19] S.L. Hillis, et al., Power estimation for multireader ROC methods an updated and unified approach, Acad. Radiol. 18 (2) (2011) 129–142.
[20] H.E. Rockette, et al., Empiric assessment of parameters that affect the design of multireader receiver operating characteristic studies, Acad. Radiol. 6 (12) (1999) 723–729.
[21] N.A. Obuchowski, Sample size tables for receiver operating characteristic studies, AJR Am. J. Roentgenol. 175 (3) (2000) 603–608.
[22] C. Nardi, et al., Is panoramic radiography an accurate imaging technique for the detection of endodontically treated asymptomatic apical periodontitis? J. Endod. 44 (10) (2018) 1500–1508.
[23] L. Meinhold, et al., Clustering effects of oral conditions based on clinical and radiographic examinations, Clin. Oral Investig. (2020) 24.
[24] S.L. Hillis, et al., Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis, Acad. Radiol. 15 (5) (2008) 647–661.
[25] D. Bates, et al., Fitting linear mixed-effects models using lme4, J. Stat. Softw. 67 (1) (2015) 1–48.
[26] D.J. Barr, et al., Random effects structure for confirmatory hypothesis testing: keep it maximal, J. Mem. Lang. 68 (3) (2013) 255–278.
[27] S.S. Shapiro, M.B. Wilk, An analysis of variance test for normality (complete samples), Biometrika 52 (3/4) (1965) 591–611.
[28] F. Wilcoxon, Individual comparisons by ranking methods, Biom. Bull. 1 (6) (1945) 80–83.
[29] K. Hung, et al., The use and performance of artificial intelligence applications in dental and maxillofacial radiology: a systematic review, Dentomaxillofac. Radiol. 49 (1) (2020) 20190107.
[30] A. Majkowska, et al., Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation, Radiology. 294 (2) (2020) 421–431.
[31] E. Berkhout, The ALARA-principle. Backgrounds and enforcement in dental practices], Ned. Tijdschr. Tandheelkd. 122 (2015) 263–270.
[32] M.H. Hamdan, et al., The effect of a deep-learning tool on dentists' performances in detecting apical radiolucencies on periapical radiographs, Dentomaxillofac. Radiol. 51 (7) (2022) 20220122.
[33] T. Ekert, et al., Deep learning for the radiographic detection of apical lesions, J. Endod. 45 (7) (2019) 917–922, e5.
[34] S. Sitaras, et al., Applications of artificial intelligence in dental medicine: a critical review, Int. Dent. J. 75 (2) (2025) 474–486.
[35] A. Aristidou, et al., Bridging the chasm between AI and clinical implementation, Lancet 399 (10325) (2022) 620.
[36] J.C.C. Kwong, et al., Integrating artificial intelligence into healthcare systems: more than just the algorithm, NPJ Digital Med. 7 (1) (2024) 52.
[37] C.J. Kelly, et al., Key challenges for delivering clinical impact with artificial intelligence, BMC. Med. 17 (1) (2019) 195.
[38] J. Shaw, et al., Artificial intelligence and the implementation challenge, J. Med. Internet. Res. 21 (7) (2019) e13659.
[39] R.J. Elderton, Overtreatment with restorative dentistry: when to intervene? Int. Dent. J. 43 (1) (1993) 17–24.
[40] D. Ericson, et al., Minimally invasive dentistry–concepts and techniques in cariology, Oral Health Prev. Dent. 1 (1) (2003) 59–72.
[41] L. Zhang, et al., Diagnostic error and bias in the department of radiology: a pictorial essay, Insights. ImAging 14 (1) (2023) 163.
[42] S.-Y. Yoon, et al., Spectrum of cognitive biases in diagnostic radiology, Radiographics. 44 (7) (2024) e230059.
[43] V. Sivaraman, et al., Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care, in: Proceedings of the 2023

CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, Association for Computing Machinery, 2023. Article 754.
[44] S. Weber, et al., How do medical professionals make sense (or not) of AI? A social-media-based computational grounded theory study and an online survey, Comput. Struct. Biotechnol. J. 24 (2024) 146–159.
[45] C. Cai, et al., Hello AI": uncovering the onboarding needs of medical practitioners for Human-AI collaborative decision-making, Proc. ACM Hum.-Comput. Interact. 3 (2019) 1–24.
[46] J. Scheetz, et al., A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology, Sci. Rep. 11 (1) (2021) 5193.
[47] B. Ehteshami Bejnordi, et al., Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer, JAMA 318 (22) (2017) 2199–2210.
[48] K. Kawamoto, et al., Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success, BMJ 330 (7494) (2005) 765.
[49] E.H. Shortliffe, M.J. Sepúlveda, Clinical decision support in the era of artificial intelligence, JAMA 320 (21) (2018) 2199–2200.
[50] A.M. Luke, N.N.F. Rezallah, Accuracy of artificial intelligence in caries detection: a systematic review and meta-analysis, Head. Face Med. 21 (1) (2025) 24.
[51] Quality guidelines for endodontic treatment: consensus report of the European society of endodontology, Int. Endod. J. 39 (12) (2006) 921–930.
[52] J. Ee, et al., Comparison of endodontic diagnosis and treatment planning decisions using cone-beam volumetric tomography versus periapical radiography, J. Endod. 40 (7) (2014) 910–916.
[53] C. Kruse, et al., Impact of cone beam computed tomography on periapical assessment and treatment planning five to eleven years after surgical endodontic retreatment, Int. Endod. J. 51 (7) (2018) 729–737.
[54] X. Liu, et al., A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis, Lancet Digit. Health 1 (6) (2019) e271–ee97.
[55] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nat. Med. 25 (1) (2019) 44–56.
[56] H.F. Duncan, et al., European Society of Endodontology position statement: management of deep caries and the exposed pulp, Int. Endod. J. 52 (7) (2019) 923–934.