Preprints.org

Article

Not peer-reviewed version

Generative AI vs. Dentists: Reliability and Reproducibility in Mandibular Cortical Index Classification

Keisuke Seki *, Minori Kashima, Taiki Akiyama, Atsushi Kobayashi, Ko Dezawa, Yoshimasa Takeuchi, Mika Furuchi, Atsushi Kamimoto

Posted Date: 1 May 2026

doi: 10.20944/preprints202605.0011.v1

Keywords: mandibular cortical index; generative AI; inter-rater reliability; dental panoramic radiography; osteoporosis screening; NotebookLM; kappa coefficient; image recognition

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2026 doi:10.20944/preprints202605.0011.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Generative AI vs. Dentists: Reliability and Reproducibility in Mandibular Cortical Index Classification

Keisuke Seki [1,*], Minori Kashima [1], Taiki Akiyama [1], Atsushi Kobayashi [1], Ko Dezawa [2], Yoshimasa Takeuchi [1], Mika Furuchi [1] and Atsushi Kamimoto [1]

1 Department of Comprehensive Dentistry and Clinical Education, Nihon University School of Dentistry, Tokyo, 101-8310, Japan

2 Department of Oral and Maxillofacial Radiology, Nihon University School of Dentistry, Tokyo, 101-8310, Japan

* Correspondence: seki.keisuke@nihon-u.ac.jp

Abstract

The mandibular cortical index (MCI) is a valuable screening tool for osteoporosis on dental panoramic radiographs; however, inter-examiner variability remains a significant challenge. This study aimed to evaluate the diagnostic performance and reproducibility of a closed-type generative AI (NotebookLM, Google) compared with eight dentists of varying experience levels. One hundred radiographs were evaluated in two sessions with an interval of at least two weeks. The intra-examiner reliability for the AI was exceptionally high ($\kappa = 0.987$), and its processing speed was approximately six times faster than that of the dentists. However, the agreement between the AI and the dentists remained at "slight agreement" or lower ($\kappa < 0.2$), statistically rejecting the null hypothesis of diagnostic equivalence. Notably, a "two-level discrepancy" was observed, where the AI interchanged Class 1 (normal) and Class 3 (severe) in over 10% of cases. In contrast, dentists demonstrated a significant learning effect, with inter-examiner agreement improving between sessions. These results suggest that while generative AI offers superior speed and reproducibility, its current decision-making logic deviates fundamentally from human expert criteria. Future integration should focus on hybrid models where AI serves as a standardized feedback tool while dentists provide final confirmatory diagnoses.

Keywords: mandibular cortical index; generative AI; inter-rater reliability; dental panoramic radiography; osteoporosis screening; NotebookLM; kappa coefficient; image recognition

1. Introduction

Morphological changes in the mandibular cortex are widely recognized as indicators of systemic bone metabolism abnormalities, including age-related changes and osteoporosis. In particular, the mandibular cortical index (MCI) has played an important role as a screening tool for bone quality assessment, as it can be readily evaluated on dental panoramic radiographs (DPR) [1,2]. The MCI is a classification system proposed by Klemetti et al., used to assess osteoporosis risk and estimate bone mineral density by visually evaluating the continuity and degree of resorption of the cortical bone [3]. Because osteoporosis is a silent chronic disease, diagnosis is often delayed, and the resulting low patient consultation rates have become a major public health concern [4,5]; there is growing interest in the potential of dental care, with MCI assessment playing a central role, to help address this problem. Nevertheless, MCI assessment is highly dependent on the observer's experience, and inter-rater variability has been identified as a challenge [2,6].

In recent years, the application of artificial intelligence (AI) in medical imaging diagnostics has advanced rapidly, and its utility has been reported across various areas of dentistry, including caries

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2026 doi:10.20944/preprints202605.0011.v1

diagnosis, detection of periodontal disease, implant treatment planning, and detection of jawbone lesions [7,8]. In particular, large language models (LLMs) and image understanding models based on them have been reported to demonstrate high accuracy in visual interpretation and classification tasks, and are expected to serve as auxiliary tools for dental image interpretation [9]. However, studies directly comparing the diagnostic accuracy of AI with that of experienced dentists in evaluating subtle differences in anatomical findings, such as those required for MCI classification, remain limited [10,11]. Furthermore, although MCI classification is of considerable clinical significance in osteoporosis risk assessment and geriatric medicine, its accurate diagnosis requires a certain level of experience and image interpretation skill. Given the substantial inter-examiner and inter-facility variability in radiographic assessment, AI-based automated evaluation could enable consistent bone quality assessment and is expected to improve screening accuracy. It also has the potential to reduce the burden of complex image interpretation workflows and support the training of junior dentists.

The aim of this study was to use NotebookLM, a closed-type generative AI developed by Google, to perform MCI classification on DPR images and compare the results with those of eight dentists, in order to identify any tendencies in AI diagnostic behavior. We focused particularly on inter-rater and intra-rater reliability to examine whether AI possesses the reproducibility required for clinical application in MCI assessment. The null hypothesis tested in this study was that "the MCI diagnostic results of dentists and generative AI are equivalent." The findings of this study are expected to provide foundational data for the future implementation of AI as an auxiliary tool for dental image interpretation.

2. Materials and methods

2.1. Study design

This cross-sectional study evaluated the mandibular cortical index (MCI) classification of patients who visited the Nihon University School of Dentistry Mishima Dental Center. The study protocol was approved by the Ethics Committee of the Nihon University School of Dentistry (Approval No. EP25D026). The study was conducted in strict accordance with the Declaration of Helsinki (1975) [12], as revised in 2013, and followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines for observational studies [13].

2.2. Patient selection and data sources

The subjects of this study were patients who visited the Nihon University School of Dentistry Mishima Dental Center between December 2015 and March 2022. Digital panoramic radiograph (DPR) images obtained from these patients were utilized for analysis. To ensure patient anonymity, the images were exported and saved as JPEG files, excluding DICOM metadata and personal annotations. The inclusion criteria were female patients aged 20 years or older who underwent DPR imaging as part of their initial diagnostic workup at the center. The exclusion criteria were as follows: (1) images taken at external facilities; (2) patients who were pregnant or breastfeeding; (3) images where morphological diagnosis of the mandibular cortex was not possible due to artifacts or suboptimal positioning; (4) a history of mandibular resection or reconstruction; (5) bone destruction due to neoplastic lesions; and (6) a history of radiotherapy to the head and neck region.

The primary outcome was to evaluate the inter-rater reliability of MCI classification between dentists and AI. The secondary outcome was to assess the inter-rater reliability among dentists with varying levels of clinical experience.

2.3. DPR imaging data

All DPR images utilized in this study were obtained during preoperative implant examinations and were acquired using the same radiographic unit (Veraviewepocs X700, Morita, Kyoto, Japan). A

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2026 doi:10.20944/preprints202605.0011.v1

random sample of 100 images was selected and exported as JPEG files. These images were individually embedded into a presentation software (Microsoft PowerPoint, Microsoft Corp., Redmond, WA, USA) to create a 100-page document.

To ensure patient privacy and standardized evaluation, each DPR image was cropped to show only the inferior border of the left mandible. Specifically, the images were split at the midline, and the dental region was masked using a black elliptical overlay within the software (Figure 1). To minimize potential bias, as the number of remaining teeth might suggest a patient's age, the same masking procedure was applied to all cases, including edentulous patients. Finally, reference illustrations for the MCI classification criteria were included on the first page, and the entire document was converted to a PDF format to serve as the standardized evaluation sample.

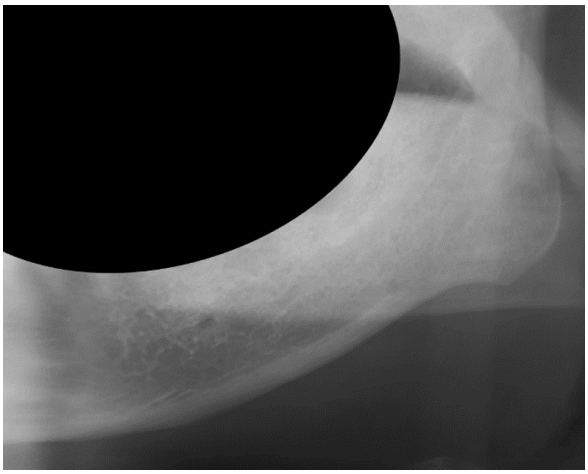

Figure 1. Cropped and masked panoramic radiograph prepared for MCI evaluation. To protect patient privacy, the image was trimmed at the midline to display only the left mandibular inferior cortex, and the dentition region was obscured with a black oval overlay to prevent age-related diagnostic bias. Patient identification data were removed, and images were stored as JPEG files without DICOM annotation.

2.4. Evaluation of the Mandibular Cortical Index (MCI).

Figure 2(A–C) presents the reference illustrations provided on the initial page of the evaluation sample. Based on the classification reported by Klemetti et al. [3], the morphology of the mandibular cortical bone at the inferior border was categorized into three classes:

・Class 1 (C1): The endosteal margin of the cortex is even and sharp on both sides.

・Class 2 (C2): The endosteal margin shows semilunar defects (lacunar resorption) or forms endosteal cortical residues.

・Class 3 (C3): The cortical layer is heavy with endosteal residues and is clearly porous.

To facilitate the evaluation process for the participants, these definitions were simplified as follows. Class 1 (C1) was described as having a smooth inner surface, Class 2 (C2) as having an irregular inner surface with linear resorption, and Class 3 (C3) as characterized by extensive resorption and porous changes throughout the cortical bone.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2026 doi:10.20944/preprints202605.0011.v1

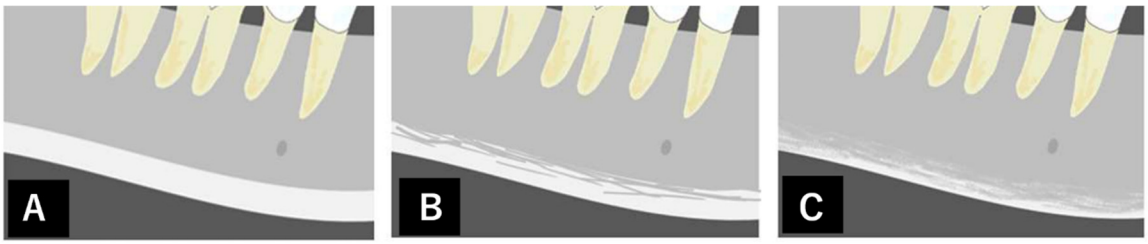

Figure 2. A–C. Reference illustrations for MCI classification. (A) Class 1 (C1): smooth and intact endosteal margin; (B) Class 2 (C2): irregular endosteal margin with endosteal cortical resorption; (C) Class 3 (C3): heavy endosteal resorption with cortical layer discontinuity.

2.5. Examiners: Dentists.

The panel of examiners consisted of eight dentists with varying levels of clinical experience:

- A male periodontist (PER-M28; 28 years of clinical experience)
- A male endodontist (END-M33; 33 years)
- A female prosthodontist (PRO-F29; 29 years)
- A male prosthodontist (PRO-M19; 19 years)
- A male dental radiologist (RAD-REF; 14 years), who served as the reference examiner
- A male general practitioner (GP-M02; 2 years)
- A female postgraduate resident (PGR-F01; 1 year)
- A male postgraduate resident (PGR-M01; 1 year)

The clinical experience of each examiner is indicated within the respective identifiers (e.g., "28" in PER-M28 refers to 28 years of experience). In the present study, as no absolute gold standard exists for the diagnostic assessment of MCI, the results provided by the dental radiologist (RAD-REF) were utilized as the reference data. The MCI classification was evaluated based on the reference illustrations provided on the first page of the sample file, and the results were recorded in separate Excel spreadsheets. Evaluations were conducted in two sessions: an initial session and a second session held at least two weeks later to minimize recall bias. Separate response sheets were used for each session, and the total time required for each evaluation was recorded.

2.6. Examiners: Generative AI

To evaluate the sample images, NotebookLM (Google, Mountain View, CA, USA), a closed-source generative AI tool, was utilized to ensure the protection of personal information. As the AI examiner, NotebookLM processed the sample PDF file following a specific protocol. After the file was uploaded, the following prompt was entered in Japanese to request the MCI assessment: “Please examine the images of the inferior border of the left mandible in this file and classify them according to the MCI criteria, referring to the reference illustrations on the first page. Please provide the results for each case as Class 1, Class 2, or Class 3 in text format.” The evaluation time for the AI was defined as the processing duration displayed within the browser environment. As with the dental examiners, the AI evaluation was conducted in two separate sessions with an interval of at least two weeks to evaluate intra-examiner reliability. The prompt and the provided sample file were identical in both sessions.

2.7. Statistical analysis

To evaluate the reliability of the MCI classification (Classes 1–3), the Kappa coefficient was calculated. Since the MCI is an ordinal scale in which the degree of bone resorption increases as the class category increases, the analysis was performed using linear weights that reflect the distance of disagreement, rather than a simple Kappa coefficient. To assess inter-rater reliability among all nine raters—including eight human raters and one generative AI (NotebookLM)—the weighted Fleiss’ kappa coefficient was used. In addition, the weighted Cohen’s kappa coefficient was used to assess

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2026 doi:10.20944/preprints202605.0011.v1

inter-rater reliability between the two raters and the agreement between two assessments by the same rater (intra-rater reliability). All analyses were performed using EZR (Saitama Medical Center, Jichi Medical University), a graphical user interface for R (The R Foundation for Statistical Computing, Vienna, Austria, version 4.0.0) [14], and 95% confidence intervals (95% CI) were calculated for each coefficient. The Kappa coefficient is interpreted according to the Landis and Koch criteria: 0.00 or less is "Poor," 0.01–0.20 is "Slight," 0.21–0.40 as "Fair," 0.41–0.60 as "Moderate," 0.61–0.80 as "Substantial," and 0.81–1.00 as "Almost perfect" [15].

3. Results

3.1. Inter-rater reliability

Table 1 summarizes the results of the inter-examiner reliability analysis for MCI classification among the eight dentists and NotebookLM (NBLM). The weighted Fleiss' kappa coefficient for all examiners (eight dentists and NotebookLM) was 0.498 for the first session and 0.542 for the second session. According to the Landis and Koch criteria, these values indicated "moderate agreement." In contrast, the inter-examiner agreement among the dentists alone was 0.606 for the first session and 0.627 for the second session, representing "substantial agreement." When the AI was included in the analysis, the overall agreement rate tended to be slightly lower compared to the agreement among human examiners alone. Figure 3 illustrates a representative case where the assessments were consistent across all examiners.

Table 1. Overall inter-examiner reliability (weighted Fleiss' κ) for MCI classification. Agreement levels across all nine examiners (eight dentists and NotebookLM) and for the dentists only are shown for both sessions, including 95% confidence intervals.

	1st Session	2nd Session
All Examiners (AI + Dentists)	0.498 [0.441, 0.555]	0.542 [0.485, 0.598]
Dentists Only	0.606 [0.548, 0.664]	0.627 [0.571, 0.683]

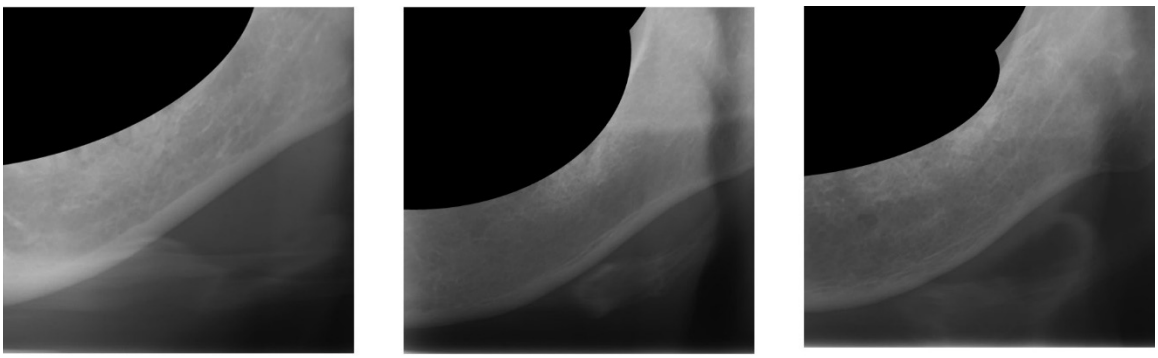

Figure 3. Representative panoramic radiograph cases showing MCI Class 1, 2, and 3 on which all examiners reached consensus.

3.2. Intra-examiner reliability and evaluation time.

Table 2 summarizes the agreement between the two assessments by the same examiner and the time required for each session. The intra-examiner reliability for the NBLM was 0.987 [95% CI: 0.962–1.000], which was the highest among all examiners and represented "almost perfect" agreement. In contrast, the intra-examiner reliability among the dentists ranged from 0.257 to 0.761, showing substantial individual variation. The two dental residents were categorized into the "fair agreement"

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2026 doi:10.20944/preprints202605.0011.v1

group, which was the lowest reliability level among all examiners. Even the dentist with the highest score achieved 0.761 ("substantial agreement"), highlighting that the AI's reproducibility is significantly higher than that of human examiners.

The average time required to evaluate the 100 cases for the dentists ranged from a minimum of 5.5 minutes (RAD-REF) to a maximum of 20.5 minutes (PRO-F29). In comparison, the average evaluation time for NotebookLM was 1.5 minutes, which was approximately six times faster than the average processing speed of the dentists.

Table 2. Intra-examiner reliability (weighted Cohen's κ) and mean evaluation time per examiner. Results are based on 100 panoramic radiographs across two evaluation sessions.

Examiner ID	Weighted Cohen's κ [95% CI]	Interpretation	Time 1 (min)	Time 2 (min)	Mean Time (min)
NBLM	0.987 [0.962, 1.000]	Almost perfect	2	1	1.5
END-M33	0.685 [0.578, 0.792]	Substantial	5	7	6
PER-M28	0.64 [0.521, 0.758]	Substantial	7	6	6.5
PRO-F29	0.645 [0.512, 0.778]	Substantial	30	11	20.5
RAD-REF	0.703 [0.589, 0.817]	Substantial	6	5	5.5
PRO-M19	0.761 [0.651, 0.871]	Substantial	10	9	9.5
GP-M02	0.619 [0.466, 0.772]	Substantial	8	6	7
PGR-M01	0.416 [0.252, 0.581]	Fair	15	15	15
PGR-F01	0.257 [0.108, 0.406]	Fair	6	8	7

3.3. Discrepancies in diagnostic criteria between dentists and generative AI

Table 3 presents the degree of agreement across the 36 possible pairings among examiners during the first session. The linearly weighted kappa coefficients among the dentists ranged from 0.055 to 0.631. Among these combinations, "substantial agreement" was observed between GP-M02 and PRO-M19 (0.631) and between GP-M02 and PRO-F29 (0.623), representing the highest reliability among the human pairings. Moreover, "moderate agreement" was the most frequent category, occurring in 13 pairs. Conversely, some pairings, such as the two dental residents (PGR-F01 and PGR-M01), showed only "slight agreement" (0.055), highlighting significant individual variation in diagnostic judgment.

In contrast, the agreement between NBLM and the dentists ranged from -0.067 to 0.128, indicating a significantly low level of consensus. The highest value for NBLM-dentist combinations was 0.128 ("slight agreement") with GP-M02. However, the agreement values were negative for both dental residents (PGR-F01: -0.067; PGR-M01: -0.012), resulting in a "poor agreement" classification. These findings suggest that while certain combinations of dentists maintain a reliable level of consistency through shared expertise, the decision-making logic of NBLM deviates substantially from the evaluation criteria utilized by human experts for MCI classification. Notably, significant discrepancies were observed where the AI and human assessments were completely reversed, specifically between Class 1 (normal/mild) and Class 3 (severe).

Table 3. Pairwise inter-examiner agreement matrix (linearly weighted Cohen's κ) for the first session.

	AI-NBLM	END-M33	PER-M28	PRO-F29	PRO-M19	RAD-REF	GP-M02	PGR-M01	PGR-F01
AI-NBLM	-								
END-M33	0.053 [−0.101, 0.207]	-							
PER-M28	0.049 [−0.106, 0.204]	0.56 [0.412, 0.708]	-						
PRO-F29	0.022	0.437	0.346	-					

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2026 doi:10.20944/preprints202605.0011.v1

	[−0.119, 0.163]	[0.285, 0.589]	[0.198, 0.495]						
PRO-M19	0.089 [−0.063, 0.241]	0.491 [0.334, 0.649]	0.498 [0.344, 0.652]	0.539 [0.389, 0.689]	-				
RAD-REF	0.005 [− 0.146, 0.156]	0.439 [0.287, 0.591]	0.308 [0.158, 0.457]	0.352 [0.201, 0.503	0.51 [0.353, 0.667]	-			
GP-M02	0.128 [−0.038, 0.294]	0.397 [0.245, 0.549]	0.344 [0.198, 0.489]	0.623 [0.473, 0.773]	0.631 [0.485, 0.777]	0.438 [0.287, 0.589]	-		
PGR-M01	-0.012 [− 0.165, 0.141]	0.264 [0.118, 0.410]	0.19 [0.054, 0.325]	0.455 [0.301, 0.609]	0.385 [0.155, 0.615]	0.401 [0.239, 0.563]	0.454 [0.302, 0.606]	-	
PGR-F01	-0.067 [− 0.205, 0.071]	0.48 [0.324, 0.636]	0.444 [0.284, 0.604]	0.144 [− 0.012, 0.300]	0.26 [0.114, 0.406]	0.227 [0.076, 0.378]	0.139 [− 0.005, 0.283]	0.055 [− 0.093, 0.203]	-

3.4. Consistency and persisting discrepancies in the second evaluation session

Table 4 presents the agreement levels for the 36 possible examiner pairings during the second session. Among the pairings of dentists, the agreement levels ranged from "moderate" (14 pairs) to "substantial" (7 pairs). Particularly high agreement was observed between GP-M02 and PER-M28 (0.688) and between PGR-F01 and RAD-REF (0.682), suggesting that the dentists maintained highly consistent diagnostic criteria across sessions. Conversely, pairings involving PGR-M01 showed relatively low agreement (range: 0.242–0.523). This may be attributed to a specific bias in PGR-M01's assessments; for instance, this examiner classified 69 out of 100 cases as "Class 1." In contrast, the agreement between NBLM and human examiners remained low across all combinations (range: 0.003 [END-M33] to 0.161 [PGR-F01]). Unlike the first session, no pairings were categorized as "poor"; however, all AI-human combinations remained within the "slight agreement" range. For all AI-related pairings, the lower limit of the 95% confidence interval was either zero or a negative value, indicating that NBLM's judgments did not achieve statistically significant agreement with those of the human examiners. Detailed analysis of these discrepancies revealed a recurring "significant two-tier discrepancy"—where Class 1 (normal/mild) and Class 3 (severe) were interchanged—across all AI-dentist pairings. These results further demonstrate that the decision-making logic of current generative AI differs fundamentally from the diagnostic criteria utilized by human experts.

Table 4. Pairwise inter-examiner agreement matrix (linearly weighted Cohen's κ) for the second session.

	AI-NBLM	**END-M33**	**PER-M28**	**PRO-F29**	**PRO-M19**	**RAD-REF**	**GP-M02**	**PGR-M01**	**PGR-F01**
AI-NBLM	-								
END-M33	0.003 [−0.158, 0.164]	-							
PER-M28	0.097 [−0.068, 0.261]	0.534 [0.385, 0.683]	-						
PRO-F29	0.129 [−0.036, 0.294]	0.205 [0.046, 0.364]	0.501 [0.344, 0.657]	-					
PRO-M19	0.091 [−0.073, 0.256]	0.436 [0.277, 0.595]	0.649 [0.508, 0.790]	0.444 [0.281, 0.607]	-				
RAD-REF	0.129 [−0.032, 0.290]	0.345 [0.198, 0.493]	0.52 [0.367, 0.673]	0.504 [0.351, 0.656]	0.55 [0.405, 0.695]	-			
GP-M02	0.097 [−0.068, 0.262]	0.314 [0.158, 0.470]	0.688 [0.557, 0.820]	0.671 [0.532, 0.810]	0.566 [0.417, 0.715]	0.532 [0.381, 0.682]	-		
	0.117	0.242	0.341	0.424	0.315	0.523	0.4	-	

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2026 doi:10.20944/preprints202605.0011.v1

PGR-M01	[−0.047, 0.282]	[0.101, 0.383]	[0.180, 0.501]	[0.261, 0.587]	[0.154, 0.475]	[0.368, 0.677]	[0.239, 0.561]		
	0.161	0.382	0.672	0.539	0.667	0.682	0.678	0.422	-
PGR-F01	[0.000, 0.321]	[0.237, 0.527]	[0.533, 0.811]	[0.386, 0.692]	[0.528, 0.807]	[0.549, 0.815]	[0.545, 0.811]	[0.263, 0.581]	

3.5. Agreement between individual examiners and the reference standard

An analysis of the agreement between individual examiners and the dental radiologist (RAD-REF), who served as the reference standard, revealed several key trends. Among the group of dentists, the degree of agreement with the reference ranged from "fair" (weighted $\kappa = 0.227$; first session of PGR-F01) to "substantial" (weighted $\kappa = 0.682$; second session of PGR-F01) for most participants. In contrast, the agreement between NBLM and the reference was consistently low, with weighted kappa values of 0.005 in the first session and 0.129 in the second session, both of which were categorized as "poor." These results suggest that while the AI demonstrates exceptionally high intra-examiner reproducibility, it performs classifications based on an internal logic that deviates significantly from the clinical diagnostic criteria utilized by dental professionals.

3.6. Qualitative analysis of diagnostic disagreements

A detailed analysis of the cases with conflicting results revealed a distinct divergence between human and AI assessments. A "two-level discrepancy"—defined as a case where the AI classified an image as Class 3 (severe) while a dentist classified it as Class 1 (normal/mild), or vice versa—was observed in 11 to 15 cases (over 10%) across all dentist-AI pairings.

3.7. Specific classification bias and misinterpretation of indicators

Among some postgraduate residents (PGRs), a systematic tendency, or classification bias, toward assigning cases to a specific category (e.g., Class 1) was observed. In contrast, the discrepancies noted in NBLM appeared more stochastic, resulting from the misinterpretation of anatomical indicators, specifically the morphology of the endosteal margin of the cortical bone.

3.8. Visualization of Inter-examiner Agreement Using Heatmaps

The inter-examiner agreement (linearly weighted κ coefficients) for the first session was visualized using a heatmap (Figure 4). In this heatmap, the rows and columns corresponding to NBLM appeared in predominantly light tones, indicating consistently low agreement and forming a visual pattern distinct from the human examiners. Among the dentists, areas representing high agreement (indicated in the red spectrum) were concentrated among mid-career and senior practitioners, particularly involving PRO-M19, GP-M02, and PRO-F29. In contrast, the rows and columns associated with the postgraduate residents (PGR-M01 and PGR-F01) generally exhibited lighter colors. These visual gradients effectively reflect the variations in agreement levels relative to clinical proficiency and experience.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2026 doi:10.20944/preprints202605.0011.v1

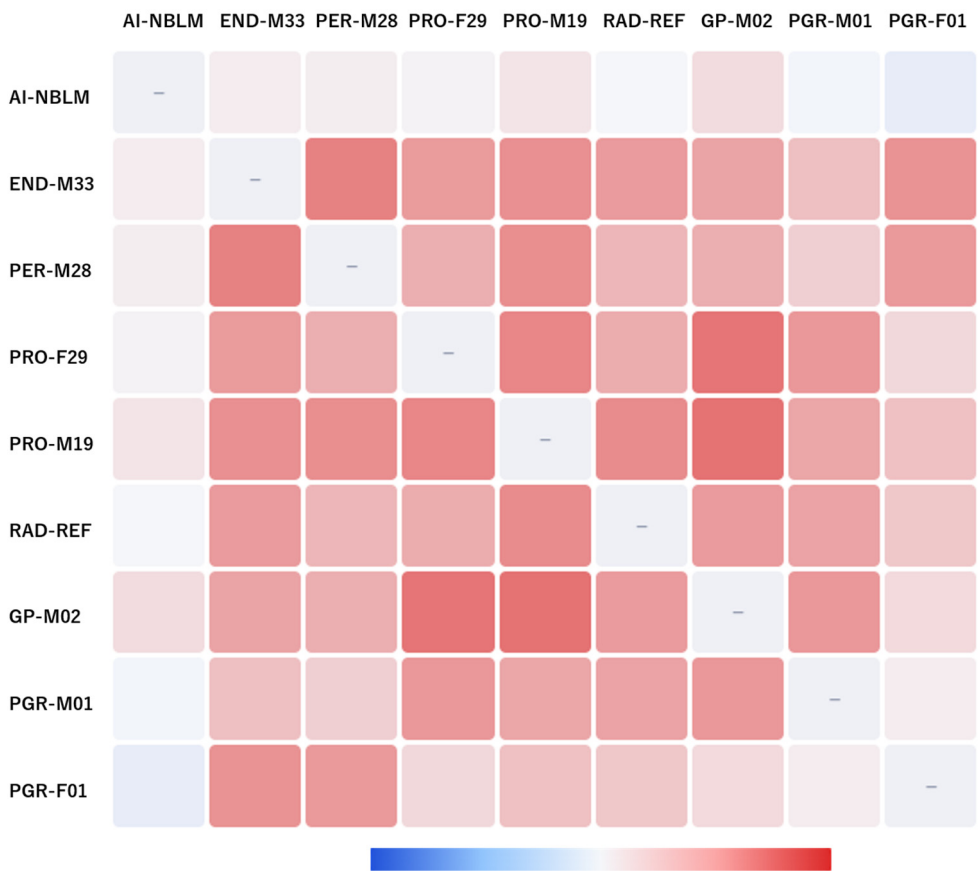

Figure 4. Heatmap of pairwise inter-examiner agreement (linearly weighted Cohen's κ) from the first session. Darker red colors indicate higher agreement. The rows and columns corresponding to the generative AI (NotebookLM) show consistently low values, forming a pattern distinct from the dentist-only regions.

Next, the inter-examiner agreement for the second session is presented as a heatmap in Figure 5. In this second heatmap, the red areas representing high agreement among the dentists were generally more extensive and intense compared to those in the first session. This visually confirms a convergence of diagnostic criteria among the human examiners as they gained more experience with the evaluation process. In contrast, the rows and columns corresponding to NBLM remained isolated as light-colored regions. This indicates that the AI's divergence from the group of dentists was persistent and consistent across both evaluation sessions.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2026 doi:10.20944/preprints202605.0011.v1

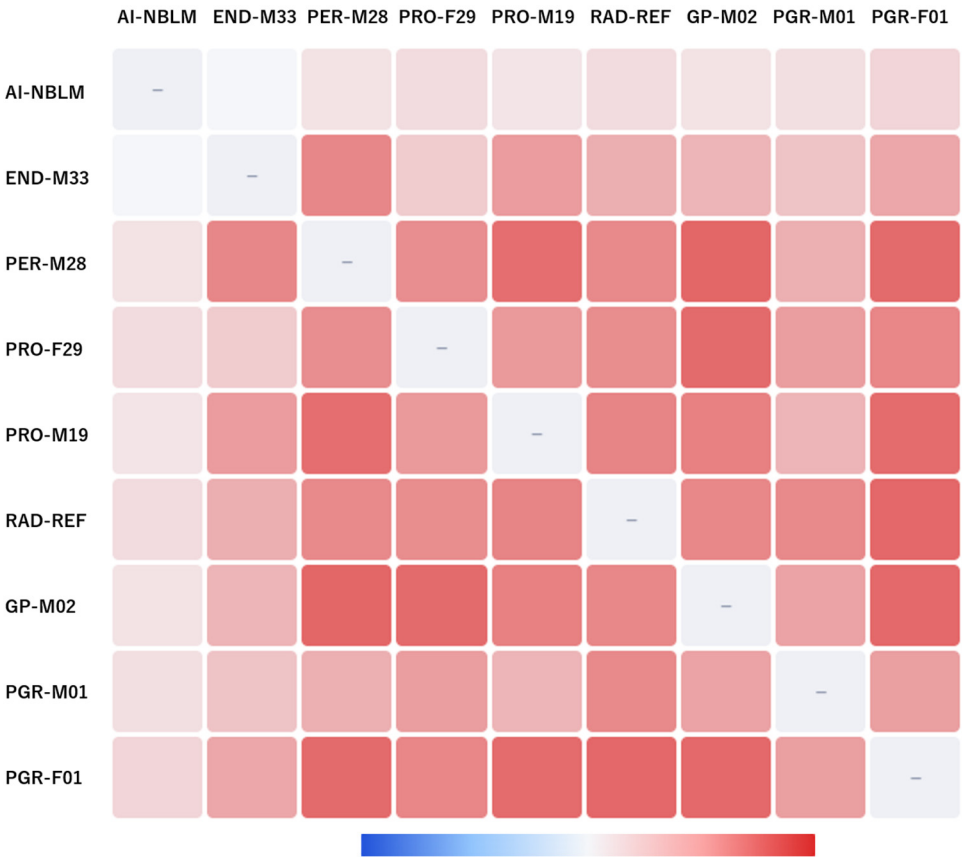

Figure 5. Heatmap of pairwise inter-examiner agreement (linearly weighted Cohen's κ) from the second session. Compared with Figure 4, the high-agreement regions among dentists expanded, visually reflecting the convergence of diagnostic criteria through repeated evaluation. NotebookLM remained isolated in the low-agreement zone.

3.9. Cluster Analysis

Hierarchical cluster analysis was performed using the results of the second session (Table 4), for which the convergence of diagnostic criteria had been confirmed via heatmap analysis. The agreement between each pair of examiners (linearly weighted κ coefficients) was converted into a distance metric ($d = 1 - \kappa$), and clustering was executed using Ward's method (Figure 6). To determine the optimal number of clusters, the changes in clustering distances on the dendrogram were examined. A cutoff height of 0.701 was selected, as it represented the most significant jump in distance, resulting in the classification of the examiners into three distinct clusters. Cluster 1 consisted solely of NBLM, joining the hierarchy at the furthest distance from all other examiners. Cluster 2 consisted solely of END-M33, reflecting a unique position within the dentist group. Cluster 3, the largest group, comprised seven dentists: PGR-M01, PRO-M19, RAD-REF, PGR-F01, PRO-F29, PER-M28, and GP-M02, representing the primary diagnostic framework shared among the human examiners. These results visually and structurally demonstrate that NBLM's decision-making logic differs fundamentally from the diagnostic framework shared by the dentists.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2026 doi:10.20944/preprints202605.0011.v1

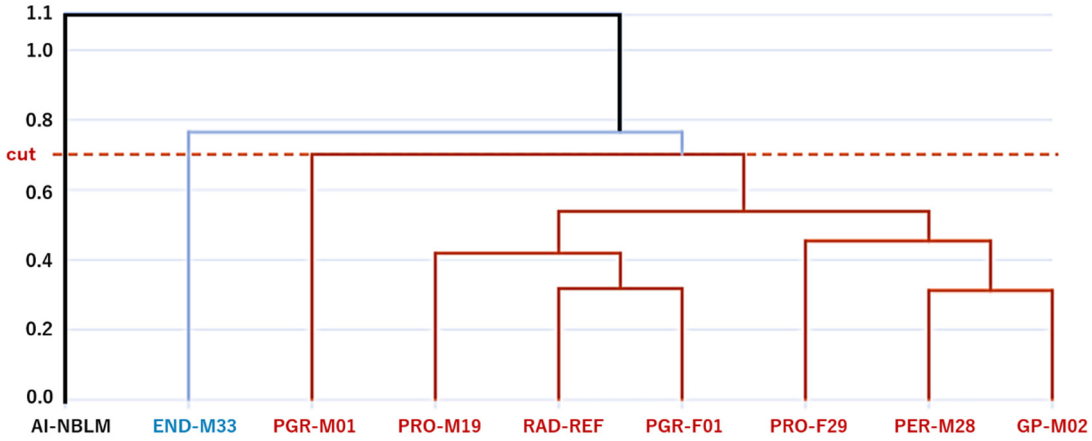

Figure 6. Dendrogram from hierarchical cluster analysis (Ward's method) based on second-session pairwise κ values (distance d = 1 - κ; cutoff = 0.701). Three clusters were identified: Cluster 1 (NBLM, isolated at the furthest hierarchical distance), Cluster 2 (END-M33, occupying a unique intermediate position), and Cluster 3 (seven dentists comprising the primary diagnostic framework), illustrating the fundamental divergence of AI decision-making logic from human diagnostic criteria.

4. Discussion

In the present study, we evaluated the degree of agreement between a generative AI tool (NotebookLM) and eight dentists in the classification of the mandibular cortical index (MCI). Regarding the null hypothesis that "the MCI diagnostic performance of generative AI is equivalent to that of dentists," the agreement between the AI and all human examiners remained at "slight agreement" or lower in both sessions. Furthermore, as the lower limit of the 95% confidence interval was zero or negative, the null hypothesis was statistically rejected. Consequently, at present, MCI assessments performed by NotebookLM cannot be considered equivalent to those made by dental professionals.

4.1. Comparison of Diagnostic Performance

The primary factor underlying the discrepancy between AI and human judgment is the inherent limitation of NotebookLM's image analysis capabilities. This system is a large language model (LLM) primarily designed to process text and structured document data; therefore, it may lack the specialized computer vision architecture required for the high-precision recognition of subtle morphological changes along the endosteal margin of the cortical bone [16,17]. The observed "two-level discrepancy"—where Class 1 and Class 3 classifications were interchanged—likely reflects this fundamental limitation in visual processing. Conversely, NBLM demonstrated an intra-examiner reliability of 0.987, indicating that its reproducibility in response to an identical prompt is exceptionally high. This finding underscores the necessity of evaluating accuracy and reproducibility as independent characteristics when assessing the clinical utility of generative AI.

The selection of a closed-type system, NotebookLM, for this study was primarily driven by ethical considerations regarding data security. While utilizing multimodal AI systems with advanced computer vision capabilities, such as GPT-4o or Gemini 1.5 Pro, might have been ideal from a purely technical perspective, these open-type systems pose a potential risk of data transmission to external servers. Consequently, obtaining approval from the institutional ethics committee was not feasible due to concerns over the protection of patient privacy. Compliance with Japan's Act on the Protection of Personal Information, national medical information guidelines [18,19], and the EU's General Data Protection Regulation (GDPR) [20] necessitates strict control over the external sharing of medical data. This remains a significant international challenge for the research and development of medical

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2026 doi:10.20944/preprints202605.0011.v1

AI. Recently, there has been progress in on-premises AI systems and dedicated medical-grade closed systems that operate entirely within local infrastructures [21]. As technical solutions and legal frameworks evolve, we anticipate the establishment of environments where high-precision AI can be utilized in an ethically sound manner. Previous studies have reported that dedicated deep-learning-based diagnostic models demonstrate high accuracy in evaluating mandibular morphology, including MCI classification [22,23]. Therefore, comparative validation between such specialized systems and general-purpose generative AI remains a critical task for future research.

4.2. Learning Effects in Human Examiners

Regarding diagnostic proficiency among the dentists, the average inter-examiner agreement improved from 0.386 to 0.491 between the first and second sessions, and the number of pairings demonstrating "substantial agreement" increased from two to seven. This convergence of diagnostic criteria, also visually confirmed by the heatmap and dendrogram, suggests a learning effect resulting from repeated evaluations [2,24]. However, as the assessments were conducted only twice, longitudinal studies are required to identify the learning plateau and evaluate long-term diagnostic stability. Significant individual variation was observed among the postgraduate residents; while PGR-F01 achieved a high degree of agreement with multiple experienced practitioners in the second session, PGR-M01 showed relatively lower agreement with other examiners and was positioned as an intermediate node between END-M33 and the main dentist cluster in the dendrogram, suggesting a residual tendency toward classification bias. While AI could serve as a valuable training tool for residents by providing consistent feedback and standardized evaluation criteria [25,26], the current version of NotebookLM lacks the diagnostic accuracy necessary for such educational applications. Looking forward, a hybrid model—combining AI-based initial screening with a final confirmatory diagnosis by a dentist—represents a realistic and promising approach [27,28]. From the perspective of standardizing diagnostic criteria across different clinical facilities and promoting large-scale osteoporosis screening, the integration of high-precision computer vision AI remains highly significant [29,30].

4.3. Limitations

This study has several limitations. First, the small sample size of 100 cases from a single facility, combined with the evaluation by only eight dentists, may limit the generalizability of the findings. Second, due to ethical constraints, a direct comparison with the latest AI systems specialized in computer vision was not feasible; therefore, these findings are restricted to a specific general-purpose LLM (NotebookLM). Additionally, a detailed analysis of the specific causes of AI diagnostic errors (such as identifying focal points) and a longitudinal assessment of the stability of the observed learning effects among dentists remain insufficient. Future research involving large-scale validation across multiple institutions and a wider variety of AI systems is warranted to address these challenges.

5. Conclusions

This study compared the diagnostic performance of a closed-type generative AI (NotebookLM) with that of dentists in classifying the mandibular cortical index (MCI) using dental panoramic radiographs. Our findings reveal that while NotebookLM demonstrates exceptional reproducibility and diagnostic speed—significantly outperforming human examiners in these areas—its diagnostic accuracy remains insufficient, leading to the rejection of the null hypothesis regarding its equivalence to expert clinicians. Although NotebookLM is primarily a large language model (LLM) tailored for text analysis, it exhibits a nascent capability for image-based classification. However, accurately identifying subtle morphological changes in cortical bone remains a significant challenge for this general-purpose system. Future efforts should leverage the AI's high reproducibility to develop standardized feedback tools for dental education. Furthermore, pursuing comparative validation

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2026 doi:10.20944/preprints202605.0011.v1

with multimodal AI systems specialized in computer vision will be essential to objectify and streamline MCI diagnosis in clinical practice.

Author Contributions: Conceptualization, K.S. and A.Ka.; methodology, K.S. and Y.T.; software, K.S. and M.F.; validation, K.S., T.A. and M.K.; formal analysis, K.S.; investigation, K.S., M.K., T.A., A.Ko., K.D., Y.T., M.F. and A.Ka.; data curation, K.S. and A.Ko.; writing—original draft preparation, K.S.; writing—review and editing, Y.T. and M.F.; visualization, K.S.; supervision, A.Ka.; project administration, K.S.; funding acquisition, K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by JSPS KAKENHI Grant Number JP23K16235 (Grant-in-Aid for Early-Career Scientists).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Nihon University School of Dentistry (Approval No. EP25D026, date of approval: 16 October 2025).

Informed Consent Statement: Patient consent was waived due to the retrospective nature of the study and the use of de-identified archival radiographic data, as approved by the Ethics Committee.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy and ethical restrictions.

Acknowledgments: The authors would like to thank the staff of the Nihon University School of Dentistry Mishima Dental Center for their support in data collection. After writing the original manuscript, the authors utilized Google's Gemini (Gemini 3 Flash) for English proofreading to enhance the linguistic quality and readability of the text.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Taguchi, A.; Tanaka, R.; Kakimoto, N.; Morimoto, Y.; Arai, Y.; Hayashi, T.; Kurabayashi, T.; Katsumata, A.; Asaumi, J.; Japanese Society for Oral and Maxillofacial Radiology. Clinical guidelines for the application of panoramic radiographs in screening for osteoporosis. *Oral Radiol.* **2021**, *37*, 189–208. https://doi.org/10.1007/s11282-021-00518-6
2. Seki, K.; Nagasaki, M.; Yoshino, T.; Yano, M.; Kawamoto, A.; Shimizu, O. Radiographical diagnostic evaluation of mandibular cortical index classification and mandibular cortical width in female patients prescribed antiosteoporosis medication. *Diagnostics* **2024**, *14*, 1009. https://doi.org/10.3390/diagnostics14101009
3. Klemetti, E.; Kolmakov, S.; Kröger, H. Pantomography in assessment of the osteoporosis risk group. *Scand. J. Dent. Res.* **1994**, *102*, 68–72. https://doi.org/10.1111/j.1600-0722.1994.tb01156.x
4. Nguyen, T.V.; Center, J.R.; Eisman, J.A. Osteoporosis: underrated, underdiagnosed and undertreated. *Med. J. Aust.* **2004**, *180*, S18–S22. https://doi.org/10.5694/j.1326-5377.2004.tb05906.x
5. Singer, A.J.; Sharma, A.; Deignan, C.; Borgermans, L. Closing the gap in osteoporosis management: the critical role of primary care in bone health. *Curr. Med. Res. Opin.* **2023**, *39*, 387–398. https://doi.org/10.1080/03007995.2022.2141483
6. Jowitt, N.; MacFarlane, T.; Devlin, H.; Klemetti, E.; Horner, K. The reproducibility of the mandibular cortical index. *Dentomaxillofacial Radiol.* **1999**, *28*, 141–144. https://doi.org/10.1038/sj/dmfr/4600427
7. Revilla-León, M.; Gómez-Polo, M.; Vyas, S.; Barmak, A.B.; Özcan, M.; Att, W.; Krishnamurthy, V.R. Artificial intelligence applications in restorative dentistry: A systematic review. *J. Prosthet. Dent.* **2022**, *128*, 867–875. https://doi.org/10.1016/j.prosdent.2021.02.010
8. Revilla-León, M.; Gómez-Polo, M.; Vyas, S.; Barmak, B.A.; Galluci, G.O.; Att, W.; Krishnamurthy, V.R. Artificial intelligence applications in implant dentistry: A systematic review. *J. Prosthet. Dent.* **2023**, *129*, 293–300. https://doi.org/10.1016/j.prosdent.2021.05.008

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2026 doi:10.20944/preprints202605.0011.v1

9. Liu, Z.; Nalley, A.; Hao, J.; Ai, Q.Y.H.; Yeung, A.W.K.; Tanaka, R.; Hung, K.F. The performance of large language models in dentomaxillofacial radiology: A systematic review. *Dentomaxillofacial Radiol.* **2025**, *54*, 613–631. https://doi.org/10.1093/dmfr/twaf060
10. Ogawa, R.; Ogura, I. Quantitative analysis of mandibular cortical morphology using artificial intelligence-based computer assisted diagnosis for panoramic radiography on underlying diseases and dental status in women over 20 years of age. *J. Dent. Sci.* **2024**, *19*, 937–944. https://doi.org/10.1016/j.jds.2023.07.030
11. Nguyen, V.A.; Vuong, T.Q.T.; Nguyen, V.H. Benchmarking large-language-model vision capabilities in oral and maxillofacial anatomy: A cross-sectional study. *PLoS One* **2025**, *20*, e0335775. https://doi.org/10.1371/journal.pone.0335775
12. World Medical Association. World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. *JAMA* **2013**, *310*, 2191–2194. https://doi.org/10.1001/jama.2013.281053
13. von Elm, E.; Altman, D.G.; Egger, M.; Pocock, S.J.; Gøtzsche, P.C.; Vandenbroucke, J.P.; STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *BMJ* **2007**, *335*, 806–808. https://doi.org/10.1136/bmj.39335.541782.AD
14. Kanda, Y. Investigation of the freely available easy-to-use software 'EZR' for medical statistics. *Bone Marrow Transplant.* **2013**, *48*, 452–458. https://doi.org/10.1038/bmt.2012.244
15. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. https://doi.org/10.2307/2529310
16. Urooj, B.; Ali, S.; Naqvi, S.K.H.; Xiao, F.; Huang, P.C. Large language models in medical image analysis: A systematic survey and future directions. *Biomed. J.* **2025**, *48*, 100932. https://doi.org/10.1016/j.bj.2025.100932
17. Bhayana, R. Chatbots and large language models in radiology: A practical primer for clinical and research applications. *Radiology* **2024**, *310*, e232756. https://doi.org/10.1148/radiol.232756
18. Personal Information Protection Commission; Ministry of Health, Labour and Welfare. Guidance on Appropriate Handling of Personal Information by Medical and Long-Term Care Service Providers. Available online: https://www.ppc.go.jp/personalinfo/legal/iryoukaigo_guidance/ (accessed on 26 April 2026).
19. Conduah, A.K.; Ofoe, S.; Siaw-Marfo, D. Data privacy in healthcare: Global challenges and solutions. *Digit. Health* **2025**, *11*, 20552076251343959. https://doi.org/10.1177/20552076251343959
20. Meszaros, J.; Minari, J.; Huys, I. The future regulation of artificial intelligence systems in healthcare services and medical research in the European Union. *Front. Genet.* **2022**, *13*, 927721. https://doi.org/10.3389/fgene.2022.927721
21. Ng, M.Y.; Helzer, J.; Pfeffer, M.A.; Seto, T.; Hernandez-Boussard, T. Development of secure infrastructure for advancing generative artificial intelligence research in healthcare at an academic medical center. *J. Am. Med. Inform. Assoc.* **2025**, *32*, 586–588. https://doi.org/10.1093/jamia/ocaf005
22. Tassoker, M.; Öziç, M.U.; Yuce, F. Comparison of five convolutional neural networks for predicting osteoporosis based on mandibular cortical index on panoramic radiographs. *Dentomaxillofacial Radiol.* **2022**, *51*, 20220108. https://doi.org/10.1259/dmfr.20220108
23. Nakamoto, T.; Taguchi, A.; Kakimoto, N. Osteoporosis screening support system from panoramic radiographs using deep learning by convolutional neural network. *Dentomaxillofacial Radiol.* **2022**, *51*, 20220135. https://doi.org/10.1259/dmfr.20220135
24. Yasar, F.; Sener, S.; Yesilova, E.; Akgünlü, F. Mandibular cortical index evaluation in masked and unmasked panoramic radiographs. *Dentomaxillofacial Radiol.* **2009**, *38*, 86–91. https://doi.org/10.1259/dmfr/56808511
25. Claman, D.; Sezgin, E. Artificial intelligence in dental education: Opportunities and challenges of large language models and multimodal foundation models. *JMIR Med. Educ.* **2024**, *10*, e52346. https://doi.org/10.2196/52346
26. Uribe, S.E.; Maldupa, I.; Schwendicke, F. Integrating generative AI in dental education: A scoping review of current practices and recommendations. *Eur. J. Dent. Educ.* **2025**, *29*, 341–355. https://doi.org/10.1111/eje.13074

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 May 2026 doi:10.20944/preprints202605.0011.v1

27. Ezhov, M.; Gusarev, M.; Golitsyna, M.; Yates, J.M.; Kushnerev, E.; Tamimi, D.; Aksoy, S.; Shumilov, E.; Sanders, A.; Orhan, K. Clinically applicable artificial intelligence system for dental diagnosis with CBCT. *Sci. Rep.* **2021**, *11*, 15006. https://doi.org/10.1038/s41598-021-94093-9
28. Ding, H.; Wu, J.; Zhao, W.; Matinlinna, J.P.; Burrow, M.F.; Tsoi, J.K.H. Artificial intelligence in dentistry—A review. *Front. Dent. Med.* **2023**, *4*, 1085251. https://doi.org/10.3389/fdmed.2023.1085251
29. Ghasemi, N.; Rokhshad, R.; Zare, Q.; Shobeiri, P.; Schwendicke, F. Artificial intelligence for osteoporosis detection on panoramic radiography: A systematic review and meta-analysis. *J. Dent.* **2025**, *156*, 105650. https://doi.org/10.1016/j.jdent.2025.105650
30. Khadivi, G.; Akhtari, A.; Sharifi, F.; Zargarian, N.; Esmaeili, S.; Ahsaie, M.G.; Shahbazi, S. Diagnostic accuracy of artificial intelligence models in detecting osteoporosis using dental images: A systematic review and meta-analysis. *Osteoporos. Int.* **2025**, *36*, 1–19. https://doi.org/10.1007/s00198-024-07229-8

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.