Enhancing Diagnostic Precision: A Calibration Study Integrating Artificial Intelligence for Dental Caries Detection in Dentistry Training

Rina Fadillah[1,2], Rasmi Rikmasari[3], Saiful Akbar[4], Arlette Suzy Setiawan[5*]

1. Faculty of Dentistry, Universitas Padjadjaran, Bandung Indonesia.
2. Department of Dental Public Health, Faculty of Dentistry, Universitas Jenderal Achmad Yani, Cimahi Indonesia.
3. Department of Prosthodontics, Faculty of Dentistry, Universitas Padjadjaran, Bandung Indonesia.
4. School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Bandung Indonesia.
5. Department of Pediatric Dentistry, Faculty of Dentistry, Universitas Padjadjaran, Bandung Indonesia.

Abstract

Dental caries is a significant public health issue, especially in children, with high prevalence rates that highlight the necessity for effective detection and management strategies. Current diagnostic methods vary in consistency, prompting the integration of Artificial Intelligence (AI) to enhance reliability and standardization.

This study aims to assess the inter-rater reliability (IRR) in dental caries detection among dental students using traditional and AI-enhanced calibration methods, preparing them for accurate assessments in clinical and educational settings. Eight fifth-year dental students participated in a structured calibration program. Calibration effectiveness was measured using Cohen's Kappa for IRR, and assessments were conducted across two sessions to evaluate the consistency and reliability of the examiners. The calibration showed an average agreement level of 69.87% for slide-based methods and 74% for dental phantoms. Kappa values ranged from 0.61 to 0.90, suggesting substantial to almost perfect agreement. AI applications demonstrated the potential to streamline and enhance the diagnostic process.

The study successfully bridged theoretical knowledge and practical skills, showing that dental phantoms combined with AI applications can significantly improve the accuracy and consistency of dental caries assessments. This calibration approach enhances diagnostic skills and facilitates the integration of advanced technologies in dental education and practice.

Experimental article (J Int Dent Med Res 2024; 17(3): 1100-1105)

Keywords: agreement; dental caries, kappa score, digital dentistry, artificial intelligence.

Received date: 14 July 2024 **Accept date:** 07 August 2024

Introduction

Dental caries represent a significant public health concern across various age groups, particularly alarming in children, with prevalences reaching 92.5% in ages 5-9 and 73.4% in ages 10-14 in Indonesia.[1,2] Despite various efforts to combat this issue, the effectiveness of prevention strategies remains a challenge[3], emphasizing the need for early detection and standardized methods for dental caries examination, such as the World Health Organization's Oral Health Survey Basic Method and the International Caries Detection and Assessment System (ICDAS).[1,4–6] Current methods predominantly involve manual examination by dentists, which, while generally reliable, may face challenges in consistency and accuracy over time, suggesting the potential for Artificial Intelligence (AI) to mitigate these issues by standardizing and enhancing the diagnostic process.[7–10]

In 2021, we launched the digital application Halo Indonesia Bersama Dentist (HI BOGI), primarily aimed at dental health education for school children. Since its introduction, HI BOGI has been actively used by the public [11]. The digitalization of dentistry, mainly through applications like HI BOGI, highlights the growing role of Artificial Intelligence (AI) in enhancing dental services. One significant area of focus has been the early detection of dental caries—a process that typically demands considerable time from dentists. Our development team has been

Corresponding author:
Arlette Suzy Setiawan
Department of Pediatric Dentistry, Faculty of Dentistry, Universitas Padjadjaran, Bandung Indonesia.
E-mail: arlette.puspa@unpad.ac.id

enhancing the HI BOGI application with AI capabilities to address this. These improvements simplify and accelerate the caries detection process, making it more user-friendly and efficient.

Additionally, in relevance to improving HI BOGI, our study emphasizes the importance of calibration and training among dental examiners to ensure the quality and consistency of dental caries assessments. We underline using statistical tools such as Cohen's Kappa to evaluate reproducibility and inter-examiner reliability.[12–17]

The overarching goal of this calibration was to prepare examiners for subsequent tasks in an Artificial Intelligence (AI) system training series, explicitly labelling photo data based on ICDAS criteria. This step is crucial for enhancing the accuracy and reliability of AI-assisted dental caries detection, representing a significant advancement in integrating AI technologies within dental diagnostics. The study aims to enhance the calibration process among dentists assessing dental caries based on the ICDAS classification through AI-based application training sessions. This involves theoretical training on dental caries and practical calibration exercises, including patient simulations and analysis of intra-oral photographs with AI-based application, aiming to improve the inter-rater reliability (IRR) in dental caries assessment.

Materials and methods

Study Design

This study involved eight fifth-year dental students from the Faculty of Dentistry at Universitas Jenderal Achmad Yani (UNJANI) as participants who had previous experience in dental examination. Conducted in September 2023 at UNJANI's community laboratory, the training program aimed to align the examiners' skills with the International Caries Detection and Assessment System (ICDAS) standards through several structured phases:

Theoretical and Practical Training: Conducted by the leading investigator (RPNF) as gold-standard examiners, this phase focused on the fundamentals of dental caries, the ICDAS classification for assessment, and hands-on clinical demonstrations.

Knowledge Assessment: Participants were evaluated on their understanding of the ICDAS criteria, requiring a minimum inter-rater reliability (IRR) kappa score of 0.8 to pass.

Clinical Practice: Trainees performed supervised examinations to ensure practical application skills and consistency in diagnosing.

AI-based Application Training: Participants were introduced to an AI-based application to assist in dental caries detection, including how to interpret AI-generated assessments.

Examiner Calibration: This critical phase ensured uniformity and reliability in examination techniques across all participants. During this phase, examiners used an AI-based application to assess dental caries from slide-based images and dental phantoms. This integration of AI technology facilitated a standardised approach to detecting and evaluating caries, enhancing the consistency and accuracy of the assessments.

Examiner Callibration Procedure

The calibration aimed at achieving a high standard of consistency and reliability in dental caries assessments, guided by the leading investigator (RPNF) recognised for their expertise and standardised training as the lead trainer in Basic Health Research 2018. This procedure included:

Assessment Comparison: Evaluations were systematically compared against those conducted by the leading investigator to measure consistency and accuracy.

Kappa Statistics: Used as the primary measure of agreement, with scores categorised from poor (<0.20) to almost perfect (0.81-1.00), targeting a minimum kappa score of 0.8.

Statistical Analysis

Univariate and bivariate analyses were employed to evaluate examiner agreement and reliability in dental caries assessments using the ICDAS classification. Techniques included calculating percentage agreement and kappa statistics for inter-examiner agreement, supplemented by paired t-tests to compare mean kappa values across different time points. Data analysis was performed using SPSS Version 25.

Results

The calibration process engaged eight dental students from the Faculty of Dentistry at Universitas Jenderal Achmad Yani, consisting of two males and six females. The exercise was conducted over two sessions, each designed to enhance the student's diagnostic skills through

different modalities.

The calibration involved a slide-based presentation of sixty cases in the initial session. This session provided a baseline for the participant's ability to identify healthy teeth and various stages of dental caries using a visual medium. The results from this session showed an average calibration value of 69.87%.

The subsequent session incorporated 30 dental phantoms embedded with natural teeth, adhering to the International Caries Detection and Assessment System (ICDAS) criteria. The outcomes from this session demonstrated a higher percentage agreement of 74%, indicating a more effective training tool than slide-based images (Table 1).

The comparative analysis between the two methods suggests that dental phantoms offer a more tactile and visually accurate training experience, which is crucial in helping students understand the ICDAS criteria deeply. This approach improves the participants' diagnostic accuracy and provides a standardized and reproducible training environment. Additionally, it circumvents the ethical and logistical issues of using actual patients.

Overall, the study successfully bridged the theoretical knowledge with practical application, enhancing the examiners' diagnostic capabilities in a controlled setting. This is particularly advantageous in preparing them for future roles involving dental health evaluations and integrating AI technology in dental diagnostics.

Calibration method	n	Mean (%)	SD	Minimum (%)	Maximum (%)
Slide	8	69,87	57,24-82,50	45	89
Dental Phantom	8	74	65,08-82,91	60	90

Table 1. The Percentage agreement between raters.

Variable	Mean (%)	SD	Z	P
Slide	69,87	15,10	-0,631	0,538
Dental Phantom	74,00	10,66		

Table 2. Differences in per cent agreement between photographic slide and phantom.

Degree of Confidence Analysis

Table 2 illustrates the comparative analysis of confidence levels between the slide-based and dental phantom methods. The analysis aimed to determine if the two calibration techniques significantly differed in examiner confidence. The results indicate no statistically significant difference in the confidence levels, with a p-value greater than 0.05. This suggests that both methods provided similar confidence levels in assessing dental caries despite their different modalities.

Inter-examiner Agreement

Table 3 details the Kappa values and the percentage agreement among examiners. The results demonstrate that all but one examiner achieved a Kappa value within the range indicating substantial agreement (0.61-0.80), highlighting effective calibration across most participants. However, Examiner 3 showed a lower percentage agreement of 68.3%, with a Kappa value below 0.61, suggesting that further training or calibration might be required for this individual to reach the consistency level of their peers.

These findings help underline the importance of using both visual and tactile modalities in training and highlight the need for individual assessment and possible re-calibration to ensure all examiners meet the required standard of diagnostic accuracy. This dual approach in calibration supports a comprehensive understanding of dental caries assessment, which is crucial for accurate diagnostics and effective integration of AI technology in future dental evaluations.

	Kappa (95& CI)	Overall percentage agreement
Examiner 1	0.900 (0.86-0.98)	95.0%
Examiner 2	0.867 (0.84-0.97)	93.3%
Examiner 3	0.737 (0.75-0.93)	86.7%
Examiner 4	0.867 (0.84-0.97)	93.3%
Examiner 5	0.602 (0.68-0.88)	80.0%
Examiner 6	0.368 (0.55-0.78)	68.3%
Examiner 7	0.737 (0.65-0.85)	76.5%
Examiner 8	0.867 (0.84-0.97)	93.3%

Table 3. Kappa values and overall percentage agreement for interexaminer reproducibility.

Consistency of Inter-rater Reliability (IRR)

Table 4. comprehensively evaluates the Cronbach's alpha values associated with the different calibration techniques used in the study. The results show that all calibration techniques yielded Cronbach's alpha values greater than 0.9. This high alpha level indicates inter-rater solid reliability (IRR) consistency among all the raters involved.

The robust Cronbach's alpha values

reinforce the reliability of using photographic slides and dental phantoms as practical tools for identifying dental decay. This level of consistency across different raters underscores the trustworthiness of the calibration methods employed, ensuring that all examiners are well-aligned in their assessments and capable of delivering accurate and consistent diagnoses. This alignment is crucial for applying these methods in educational settings and clinical practice, particularly as part of training programs that prepare dental students for real-world diagnostics and the integration of artificial intelligence in dentistry.

Calibration methods	n	Cronbach's alpha
Slide	8	0,927
Phantom	8	0,942

Table 4. Inter-rater reliability based on Cronbach's alpha.

Discussion

Kappa statistics are paramount in biomedical sciences for evaluating the consistency of examiners in interpreting categorical data. These statistics range from -1, indicating complete disagreement, to +1, denoting total agreement among examiners, with 0 reflecting agreement that could merely be due to chance. Notably, Kappa values are generally lower than the Overall Percentage Agreement (OPA) since they account for agreement beyond chance occurrences. This distinction is significant as studies have illustrated instances where high OPA corresponds with low, or even negative, Kappa values, highlighting the nuanced interpretation required when evaluating examiner repeatability.[12]

Our study's calibration exercises for dental caries detection were meticulously designed, selecting representative samples with the highest disease prevalence to ensure reliable repeatability values [1,12]. Calibration aims to harmonize examiner perceptions, culminating in a Kappa score that signifies acceptable agreement for research purposes at a minimum of 0.61. Remarkably, the average interexaminer Kappa in this study reached 0.91, resonating with similar high-performing studies, while Cronbach's alpha for both slide and phantom methods was 0.9, indicating near-perfect agreement.[16]

Contrary to expectations and previous literature, our study found no significant difference in percentage agreement between slide photo and dental phantom methods, supporting the hypothesis that examiners would achieve consistent results regardless of the method employed. This finding diverges from prior research, where notable discrepancies were observed between these two calibration techniques. Such consistency underscores the robustness of our calibration approach, even as one examiner required retraining due to a low initial score, highlighting the importance of individual competence in achieving high interexaminer reliability.[18–20]

Our findings, especially the high Kappa values and OPA above 90%, signify exemplary interexaminer reproducibility, essential for the success of AI in dentistry. Following calibration, the focus shifts to training the AI model using Google Colab, marking a pivotal step towards automating caries detection, potentially revolutionizing diagnostic approaches in dental health care.

This study, while contributing valuable insights into the calibration of dental examiners and the integration of Artificial Intelligence (AI) in dental caries detection, encounters several limitations:

Sample Size and Diversity: The calibration process was confined to a relatively small group of dental students from a single institution. This limited sample size and lack of diversity among participants may not fully represent the variability in examiner proficiency that could be encountered in broader clinical settings.

Use of Dental Phantoms: Although employing dental phantoms embedded with natural teeth offers a controlled and ethical training environment, it cannot entirely replicate the complexity and variability of real patient cases. Phantoms lack dynamic biological factors and patient-specific challenges, such as saliva flow and patient movement, which can impact caries detection in clinical practice.

AI Training Data Set: The study's focus on a specific set of 3,000 images from the HI BOGI application for AI model training may not encompass the full range of dental caries presentations. This limitation might affect the AI model's ability to generalize across unseen cases, particularly those with rare or atypical manifestations of dental caries.

Kappa Statistics Interpretation: While high Kappa values indicate strong inter-examiner agreement, they do not necessarily reflect the absolute accuracy of caries detection. The reliance on Kappa statistics and percentage agreement as primary measures may only capture some nuances of diagnostic accuracy and decision-making processes.

Future research could address these limitations by expanding the pool of examiners to include a broader range of professionals, incorporating actual patient examinations to compare with phantom-based training, broadening the AI training dataset, and exploring additional measures of diagnostic accuracy. Further studies might also investigate integrating newer AI algorithms and software platforms to enhance the model's diagnostic capabilities and applicability across diverse dental healthcare settings.

Conclusions

In conclusion, this calibration study has laid a solid foundation for future research and development in integrating AI technologies within dental diagnostics. It has shown that it is possible through meticulous training and calibration to achieve high levels of inter-rater reliability, which are essential for the practical application of AI in enhancing the quality and consistency of dental caries assessments.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of Komite Etik Penelitian Kesehatan Universitas Padjadjaran with document number 1291/UN6.KEP/EC/2023.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Acknowledgments

The authors wish to thank DRPM Universitas Padjadjaran, who supported this manuscript.

Declaration of Interest

The authors declare no conflicts of interest.

AI Usage Statement: In preparing this manuscript, Artificial Intelligence (AI)-based technology, specifically Grammarly and Smodin, was employed to enhance the readability and coherence of the text. These tools assist in ensuring grammatical accuracy, clarity of expression, and the overall quality of the manuscript's language. The use of Grammarly aimed to support the authors in conveying their research findings more effectively, making the scientific discourse accessible to a broader audience while maintaining the integrity and precision of the technical content.

Declaration of Interest

The authors report no conflict of interest.

References

1. World Health Organization. Oral Health Surveys Basic Methods. 5th Edition. Geneva: World Health Organization; 2013: 1–123.
2. Kementerian Kesehatan Republik Indonesia. Laporan Nasional Riskesdas 2018. Jakarta: Badan Penerbit Penelitian dan Pengembangan Kesehatan; 2019: 1–627.
3. Hirsch G, Edelstein B, Frosh M, Anselmo T. A Simulation Model for Designing Effective Interventions in Early Childhood Caries. Prev Chronic Dis. 2012;9(1):E66–74.
4. Saini D, Jain R, Thakur A. Dental Caries early detection using Convolutional Neural Network for Tele dentistry. In: 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE; 2021: 958–63.
5. Young DA, Nový BB, Zeller GG, Hale R, Hart TC, Truelove EL, et al. The American Dental Association Caries Classification System for Clinical Practice. The Journal of the American Dental Association. 2015 Feb;146(2):79–86.
6. Gugnani N, Pandit I. International Caries Detection and Assessment System (ICDAS): A New Concept. Int J Clin Pediatr Dent. 2011 Aug;4(2):93–100.
7. Schwendicke F, Samek W, Krois J. Artificial Intelligence in Dentistry: Chances and Challenges. J Dent Res. 2020 Jul 21;99(7):769–74.
8. Schwendicke F, Elhennawy K, Paris S, Friebertshäuser P, Krois J. Deep learning for caries lesion detection in near-infrared light transillumination images: A pilot study. J Dent. 2020 Jan;92:103260.
9. Mertens S, Krois J, Cantu AG, Arsiwala LT, Schwendicke F. Artificial intelligence for caries detection: Randomized trial. J Dent. 2021 Dec;115:103849.
10. Handayani N, Budiarto A, Rachman A, Setiawan AS. Enhancing Denture Care Efficiency: Mobile Prosto Open-Source Software for Indonesian National Army Soldiers. Journal of International Dental and Medical Research. 2024;17(1):209–14.
11. Fadilah RPN, Pribadi AP, Aji RW, Kusaeri R. Effectiveness of the novel teledentistry “HI BOGI” an android-based oral health application in increasing oral health knowledge of elementary school children. Padjadjaran Journal of Dentistry. 2021 Dec 1;33(3):243.
12. Tonello AS, Silva RP da, Assaf AV, Ambrosano GMB, Peres SH de CS, Pereira AC, et al. Interexaminer agreement dental caries epidemiological surveys: the importance of disease prevalence in the sample. Revista Brasileira de Epidemiologia. 2016 Jun;19(2):272–9.
13. Markovic D, Vukovic A, Soldatovic I, Peric T, Kilibarda B, Rosianu RS, et al. Multilevel calibration procedure for the oral

health national multicenter survey in primary teeth. Int J Paediatr Dent. 2023 Nov 9;33(6):585–94.
14. Koç Vural U, Kütük ZB, Ergin E, Yalçın Çakır F, Gürgan S. Comparison of two different methods of detecting residual caries. Restor Dent Endod. 2017;42(1):48.
15. Rechmann P, Jue B, Santo W, Rechmann BMT, Featherstone JDB. Calibration of dentists for Caries Management by Risk Assessment Research in a Practice Based Research Network - CAMBRA PBRN. BMC Oral Health. 2018 Dec 4;18(1):2.
16. Nabarrette M, Santos PR dos, Assaf AV, Ambrosano GMB, Meneghim M de C, Vedovello SAS, et al. Longitudinal study for dental caries calibration of dentists unexperienced in epidemiological surveys. Braz Oral Res. 2023;37.
17. Agbaje JO, Mutsvari T, Lesaffre E, Declerck D. Examiner performance in calibration exercises compared with field conditions when scoring caries experience. Clin Oral Investig. 2012 Apr 23;16(2):481–8.
18. Susilawati S, Monica G, Fadilah RPN, Bramantoro T, Setijanto D, Wening GRS, et al. Building team agreement on large population surveys through inter-rater reliability among oral health survey examiners. Dent J. 2018 Mar 31;51(1):42–6.
19. Hartman H, Nurdin D, Akbar S, Cahyanto A, Setiawan AS. Exploring the potential of artificial intelligence in paediatric dentistry: A systematic review on deep learning algorithms for dental anomaly detection. Int J Paediatr Dent. 2024 Jan 31;1–8.
20. Yilmaz AE, Demirhan H. Weighted kappa measures for ordinal multi-class classification performance. Appl Soft Comput. 2023 Feb;134:110020.