Comparative analysis of the diagnostic accuracy of an artificial intelligence software and dental practitioners in assessing caries (using the CPO-D index) and periapical lesions on panoramic radiographs

Authors: Diogo de Almeida Costa[1], Eduardo Guerreiro[1,2]
[1] Egas Moniz School of Health and Science
[2] Egas Moniz Center for Interdisciplinary Research

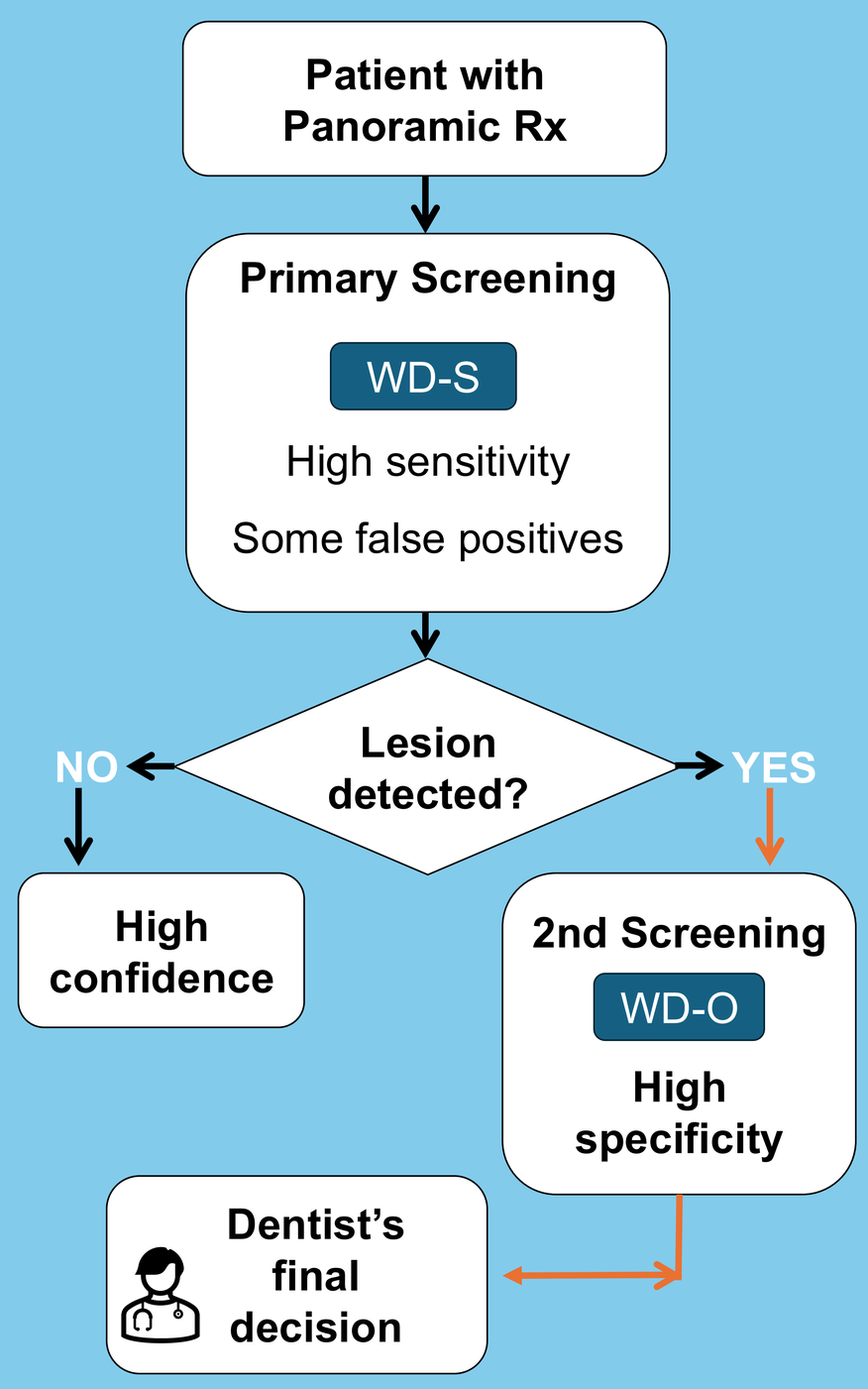
Introduction

Dental caries remains the most prevalent non-communicable disease worldwide, affecting approximately 2.5 billion individuals (Kassebaum et al., 2015 ; Guerreiro et al., 2024).
Human interpretation of orthopantomograms (OPGs) shows considerable inter-observer variability, with κ values rarely exceeding 0.60 (Landis & Koch, 1977).
Convolutional neural networks currently achieve receiver-operating-characteristic (ROC) area-under-the-curve (AUC) values above 0.90 in dentistry (Schwendicke et al., 2019). However, gaps persist regarding sample representativeness and algorithmic bias (Ezhov et al., 2021).
The equivalence—or superiority—of artificial intelligence (AI) compared with experienced observers still needs to be confirmed in a Portuguese sample.

Materials and Methods

Study design: Observational, comparative study evaluating—on panoramic radiographs—the diagnostic accuracy for dental caries (CPO-D index) and apical lesions achieved by an artificial-intelligence (AI) software versus dentists with more than five years' experience.
Sample: 200 adult OPGs (September–December 2024), quality Grade 1, selected from 780 examinations performed at the Egas Moniz University Clinic.
Gold standard: Consensus of the three evaluators (κ = 0.75, "substantial").
AI system: WeDiagnostix v23.10 operated in two modes—WD-S (sensitive) and WD-O (optimal).

From screening to confirmation: AI modes WD-S & WD-O match dentists' diagnostic accuracy



Aims

Determine the sensitivity, specificity, predictive values, accuracy, F1 score and area under the ROC curve (AUC) of **WeDiagnostix** in its Sensitive (WD-S) and Optimal (WD-O) modes.
Compare the performance of each mode with that of three dentists with ≥ 5 years of clinical practice.
Assess inter-observer agreement (Cohen's κ) among the clinicians.

Results and Conclusions

Table 1: Main results of the different modes compared with the established gold standard.

Lesion	Mode	Sensitivity	Specificity	F1	AUC
Caries	WD-S	**85,3 %**	88,7 %	0,49	**0,870**
Caries	WD-O	55,5 %	**99,0 %**	**0,66**	0,773
Apical Lesion	WD-S	**88,2 %**	95,9 %	0,31	**0,920**
Apical Lesion	WD-O	75,0 %	**99,2 %**	**0,60**	0,871

The findings align with the meta-analysis by Pul and Schwendicke (2024), which reports an overall sensitivity of 94 % and a specificity of 96 %. WD-S maximises sensitivity and is therefore recommended for screening, whereas WD-O prioritises specificity, making it suitable for diagnostic confirmation.
Limitations include the monocentric and retrospective nature of the sample. The AI training set is not publicly available, leaving a residual risk of over-fitting. Diagnostic responsibility remains with the clinician; AI constitutes a decision-support tool (Topol, 2019).

References: Ezhov, M., Gusarev, M., Golitsyna, M., Yates, J. M., Kushnerev, E., Tamimi, D., Aksoy, S., Shumilov, E., Sanders, A., & Orhan, K. (2021). Clinically applicable artificial intelligence system for dental diagnosis with CBCT. *Scientific Reports, 11*(1), 15006. https://doi.org/10.1038/s41598-021-94093-9
Guerreiro, E., Botelho, J., Machado, V., Proença, L., Mendes, J. J., & Manso, A. C. (2024). Caries experience before and after COVID-19 restrictions: An observational study. *Journal of Clinical Medicine, 13*(4). https://doi.org/10.3390/jcm13041164
Kassebaum, N. J., Bernabé, E., Dahiya, M., Bhandari, B., Murray, C. J., & Marcenes, W. (2015). Global burden of untreated caries: A systematic review and metaregression. *Journal of Dental Research, 94*(5), 650–658. https://doi.org/10.1177/0022034515573272
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.
Pul, U., & Schwendicke, F. (2024). Artificial intelligence for detecting periapical radiolucencies: A systematic review and meta-analysis. *Journal of Dentistry, 147*, 105104. https://doi.org/10.1016/j.jdent.2024.105104
Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7