International Journal of Community Medicine and Public Health
Jathar AA et al. Int J Community Med Public Health. 2026 Sep;13(9):5130-5135
https://www.ijcmph.com
pISSN 2394-6032 | eISSN 2394-6040

DOI: https://dx.doi.org/10.18203/2394-6040.ijcmph20263196

Original Research Article

Can AI see what the surgeons see? Analyzing the accuracy of ChatGPT: publicly available multimodal AI for radiographic interpretation and treatment plan in oral surgery

Astha A. Jathar[1]*, Varsha S. Manekar[1], Akash M. Bakale[1], Komal Rewatkar[1], Ayush A. Jathar[2]

[1]Department of Oral and Maxillofacial Surgery, Government Dental College and Hospital, Nagpur, Maharashtra, India
[2]Department of Computer Science, Government College of Engineering (GCOEN), Nagpur, Maharashtra, India

Received: 24 June 2026
Revised: 18 August 2026
Accepted: 20 August 2026

***Correspondence:**
Dr. Astha A. Jathar,
E-mail: jatharastha@gmail.com

ABSTRACT

Background: With the rise of AI, patients have been actively doing their own research. With emergence of easily accessible multimodal AI, such as ChatGPT-5 that can interpret images, there is a growing need to evaluate their accuracy, reliability, and potential clinical application. Therefore, this study aimed to evaluate the accuracy of radiographic diagnostics and surgical treatment planning of ChatGPT-5 in the context of oral and maxillofacial surgery.
Methods: We conducted a cross-sectional study using ChatGPT-5's free version and departmental OPG. Oral surgeons opinions regarding diagnosis and treatment plan were recorded, and a Cohen's Kappa was calculated for inter-rater agreeability. Furthermore, Cohen's Kappa was calculated amongst the Surgeons and AI's response for inter-rater reliability analysis.
Results: The mean number of positive findings per radiograph was significantly higher for ChatGPT (3.69±1.28) compared to oral surgeons (2.8±1.18), with this difference reaching statistical significance (p=0.001). Indicating that AI consistently reported more diagnostic and treatment related observations per OPG.
Conclusions: The findings demonstrated that ChatGPT-5 identified a significantly higher number of diagnostic and treatment related findings per radiograph compared to clinicians. While this suggests high sensitivity, qualitative and agreement analysis (Cohen's Kappa) revealed that many of these additional findings represented false positive interpretations.

Keywords: Artificial intelligence, ChatGPT, Maxillofacial surgery

INTRODUCTION

Artificial intelligence (AI) has rapidly advanced in recent years, demonstrating significant potential in a every field.[1] ChatGPT (chat generative pre-trained transformer), developed by OpenAI, is a large language model capable of generating human-like responses and analyzing images.[2] Recent developments have led to the emergence of multimodal AI systems, which can process and integrate diverse forms of data such as texts and images, ChatGPT (latest model being ChatGPT-5) being one.[3]

In clinical context these models have the potential to combine textual input with computer vision, enabling the interpretation of radiographic images.[4,5] Its image interpretation capabilities have attracted significant attention in the medical community, prompting both excitement and caution regarding its clinical applications. AI, especially ChatGPT, is impacting healthcare through applications in patient communication, patients doing their own research, and information.[6-9]

In dentistry and oral and maxillofacial surgery, OPGs

(orthopantomograms) have been one of the most frequently prescribed screening examination in diagnosis and treatment plans.[7] In surgical context, oral surgeons have a key role, as they are often responsible for interpreting OPGs and taking responsibility for analyzing and explaining the results to patients.[8]

At the same time, there has been a significant shift in patient behavior, patients are increasingly using AI tools to conduct preliminary research, understand diagnoses, and even question clinical recommendations.[9-12] This phenomenon has been driven by the widespread availability of mobile devices and digital platforms that facilitate access to medical information.[9,11,12]

Given the increasing accessibility of multimodal AI tools, there is a growing need to evaluate their accuracy, reliability, and potential clinical application. Despite its growing popularity and impressive general capabilities, ChatGPT's diagnostic utility particularly in the interpretation of OPGs and formulation of oral surgical treatment plans has not yet been systematically studied.[13]

Therefore, this study aimed to evaluate the accuracy of radiographic diagnostics and surgical treatment planning of ChatGPT-5 in the context of oral and maxillofacial surgery.

Aims and objectives

Evaluation of accuracy of ChatGPT-5 in interpreting radiographic features relevant to oral surgery. And to assess the accuracy and appropriateness of the Surgical treatment plan suggested by the ChatGPT-5.

METHODS

Study design and setting

This cross-sectional study was conducted in the department of oral and maxillofacial surgery at our hospital, between November 2025 and March 2026. The study aimed to evaluate the accuracy of radiographic diagnostics and surgical treatment planning of ChatGPT-5 (publicly available multimodal AI) in the context of oral surgery.

Ethical approval

Ethical approval for the study was obtained from the institutional ethics committee of our institute (reference number- IEC/GDCHN/93/2025). As the study required no patient identifiers, the requirement for informed consent was waivered.

Study sample and study tool

A publicly available multimodal version of ChatGPT-5 was used in this study. A freely accessible web-based interface available at the time of the study was utilized. No additional training or fine tuning of the model was performed.

Sample size was calculated using McNemar's test for paired proportions. The expected diagnostic accuracies were assumed to be 81% and 50%, respectively, based on findings reported by Jeong et al, with α=0.005 and 80% power, the estimated proportion of discordant pairs was 0.50.[14]

The minimum required sample size was 39 paired OPG images.

Inclusion and exclusion criteria

High-quality digital panoramic images (OPG) with high contrast and minimal distortion from the department's radiographic archives were chosen. The OPGs that were included showcased either one or some of the conditions, such as impacted third molars, retained roots, periapical radiolucency, odontogenic cysts, fractures, missing teeth with implant fixtures, and bone loss relevant to oral surgery.

Blurred or distorted images, orthodontic cases and pediatric OPGs were excluded from the present study.

Methodology

High resolution OPGs reporting to the department of oral and maxillofacial surgery were taken for this study according to the inclusion criteria.

Surgeon's assessments

Three oral and maxillofacial surgeons with more than 10 years of experience who were willing to participate, were informed about the study, their role in the study and what was expected of them, and their clinical opinions were recorded using a tabular checklist (blinded to the AI's response).

The surgeons were provided with two standardized checklists, one containing the predefined radiographic features for diagnosis, and other containing relevant treatment options. They were then asked to select the one or more appropriate options for each OPG in both the checklists.

All three surgeons systematically assessed the OPGs for presence or absence of relevant radiographic findings for diagnosis, including impacted teeth, retained roots, periapical radiolucency, cystic or other pathological lesions, implants, localized or generalized bone loss, fractures, and proximity to vital anatomical structures (such as the inferior alveolar canal and maxillary sinus).

Based on the radiographic findings, the surgeons selected the appropriate treatment options from the second predefined checklist, that included surgical extraction,

enucleation, apicoectomy, implant evaluation or placement, open or closed reduction, and other surgical interventions.

The checklist responses were collected anonymously for inter-rater reliability analysis and Cohen's Kappa was calculated.

ChatGPT-5's assessments

The same OPGs were uploaded on ChatGPT-5's free version with vision with a standard prompt.

"*Please identify the radiographic findings relevant to oral surgery. And based on the findings complete the following diagnostic and surgical treatment plan checklist that I will provide*".

The same two checklist tables that were given to the Surgeons were then uploaded. The prompt structure and the checklist format remained identical for all the OPGs to ensure consistency.

No additional feedback, clarification or re-prompting was performed. The AI was blinded to clinical history and expert interpretations. And the responses were recorded without modification.

After collecting the data, Cohen's Kappa was calculated amongst the surgeons and AI's response. For both ChatGPT-5 and surgeons, each selected finding or treatment option was coded as 1 (present) and each non-selected option as 0 (absent) for subsequent statistical analysis.

RESULTS

Cohen kappa between the surgeons

The inter-rater reliability among oral surgeons for the diagnostic parameters showed excellent agreement across all features assessed. The Cohen's Kappa values ranged from 0.91 to 1.00, indicating almost perfect agreement.

Table 1: Agreement between surgeons for diagnostic parameters.

Parameters	Kappa
Impacted teeth	1
Retained roots	0.92
Periapical radiolucency	1
Cystic pathology	1
Implants	1
Bone Loss	0.91
Fractures	1
Proximity to vital structures	1

Similarly, the inter-rater reliability for the treatment parameters also indicated excellent agreement between surgeons, with Kappa values ranging from 0.90 to 1.00.

Table 2: Agreement between surgeons for treatment parameters.

Parameters	Kappa
Surgical extraction	0.90
Enucleation	1
Apicoectomy	1
Implant placement	1
Open /close reduction	1
Removal of retained roots	1

Overall, these results demonstrate that the oral surgeons were highly consistent in both diagnostic interpretation and surgical decision-making, establishing a robust reference standard for comparison with the AI's findings.

Agreement between AI and surgeons for radiographic interpretation

The agreement between the multimodal AI (ChatGPT-5) and oral surgeons varied across radiographic parameters. The highest level of agreement was observed for fractures (κ=0.41), indicating a moderate level of agreement.

Table 3: Agreement between AI and surgeons for radiographic interpretation.

Radiographic features	Kappa (k)	Level of agreement
Impacted teeth	0.20	Slight
Retained roots	0.35	Fair
Periapical radiolucency	0.21	Fair
Cystic pathology	0.08	Slight
Implants	0.37	Fair
Bone loss	0.21	Fair
Fractures	0.41	Moderate
Proximity to vital structures	0.30	Fair

Fair agreement was observed for parameters like retained roots (κ=0.35), implants (κ=0.37), and proximity to vital structures (κ=0.30). On the other hand, impacted teeth (κ=0.20) and bone loss (κ=0.21) showed only slight to fair agreement. Cystic pathology (κ=0.08) demonstrated the lowest level of agreement.

Agreement between AI and surgeons for treatment planning parameters

The level of agreement between the ChatGPT-5 and oral surgeons in determining treatment plans was found to be low overall, with Cohen's Kappa values ranging from 0.02 to 0.41. This corresponds to a slight to moderate agreement.

Table 4: Agreement between AI and surgeons for treatment planning parameters.

Treatment parameters	Kappa (k)	Level of agreement
Surgical extraction	0.05	Slight
Enucleation	0.16	Slight
Apicoectomy	0.11	Slight
Implant placement	0.19	Slight-Fair
Open/close reduction	0.41	Moderate
Removal of retained roots	0.22	Fair
Other surgical treatment	0.02	Poor

The highest level of agreement was observed for open/closed reduction procedures (κ=0.41). Fair agreement was observed for removal of retained roots (κ=0.22) and implant placement (κ=0.19). These findings suggest that while the AI occasionally identified these surgical indications, it lacked consistency in differentiating between retained roots and residual bone fragments or prosthetic materials.

The agreement for enucleation (κ=0.16) and apicoectomy (κ=0.11) was minimal, reflecting ChatGPT's difficulty in recognizing lesions or pathologies that require precise radiographic interpretation. Similarly, surgical extraction (κ=0.05) and other surgical treatments (κ=0.02) demonstrated negligible agreement, indicating that the AI frequently over or under recommended procedures in these categories.

Table 5: Mean, standard deviation and p value.

	Mean	N	Std deviation	P value
Total ChatGPT-5	3.6923	39	1.28050	0.001
Total oral surgeon	2.8462	39	1.18185	

Overall, the results indicate that ChatGPT-5 demonstrated limited accuracy and poor concordance with clinician-derived treatment plans, especially for procedures that require context-specific decision-making.

Comparison of the total number of diagnostic and treatment related findings per radiograph revealed statistically significant difference between AI and surgeon groups (p=0.001). The mean number of positive findings per radiograph was significantly higher for ChatGPT (3.69±1.28) compared to oral surgeons (2.8±1.18), with this difference reaching statistical significance (p=0.001). Indicating that AI consistently reported more diagnostic and treatment related observations per OPG.

While this finding suggests higher sensitivity, it more likely reflects a tendency towards over diagnosis. The AI frequently reported additional findings that were not pointed out by the clinicians and it misinterpreted normal structures and radiographic artifacts.

Thus, the increased mean score for ChatGPT-5 indicates towards reduced specificity rather than superior diagnostic performance.

DISCUSSION

This cross-sectional study was designed to assess the ability of a publicly accessible multimodal AI (ChatGPT-5) to interpret panoramic radiographs (OPGs) and propose surgical treatment plans relevant to oral surgery, comparing its performance to that of experienced oral surgeon.

The study revealed that number of positive findings per radiograph was significantly higher for ChatGPT (3.69±1.28) compared to oral Surgeons (2.8±1.18), with this difference reaching statistical significance (p=0.001). Indicating that AI consistently reported more diagnostic and treatment related observations per OPG.

The model demonstrated the highest reliability for the parameter fractures (κ=0.41). However, among the 39 panoramic radiographs analysed, only one case exhibited a mandibular fracture, which the AI failed to identify. The apparently high Kappa coefficient thus reflects the agreement in negative diagnosis (i.e., both AI and clinicians correctly identifying the absence of fracture in the remaining cases) rather than true diagnostic sensitivity. This phenomenon represents a known limitation of the Kappa statistic in datasets with low prevalence of positive findings, where agreement on absence can disproportionately inflate reliability values.

For implants, retained roots, cystic pathologies and bone loss, AI's performance was notably weaker. These results suggest that ChatGPT's diagnostic accuracy is feature-dependent and it performs relatively well when the radiographic features are distinct, but struggles with subtle radiolucent variations.

The agreement between ChatGPT-5 and surgeons in treatment planning parameters was even lower, the only statistically significant result was found for open/closed reduction procedures (κ=0.41). Again, the observed moderate Kappa value does not indicate effective fracture detection by the AI but rather a statistical artifact related to class imbalance. In contrast, treatment decisions requiring more nuanced clinical judgment such as surgical extraction, enucleation, and apicoectomy, demonstrated poor agreement.

The model frequently over-recommended surgical interventions, sometimes suggesting apicoectomy or extraction for clinically sound teeth. This pattern aligns with its observed over-diagnosis behaviour in

radiographic interpretation, where textual generalizations override nuanced clinical reasoning.

The findings of the present study are consistent with those reported by Suárez et al who compared radiologists' interpretations of 900 panoramic radiographs with AI-generated analyses.[13] Their study demonstrated that while ChatGPT could identify general radiographic features in some cases, it frequently produced incomplete or fabricated interpretations, particularly in complex images involving overlapping anatomical structures or developing roots. These limitations closely parallel the diagnostic inaccuracies observed in the present study.

Another study by Hui Jeong et al evaluated the performance of ChatGPT and other large language models in oral and maxillofacial radiology compared with dental students, and suggested similar findings to the present study.[14] In their study, dental students achieved an overall diagnostic accuracy of 81.2%, whereas ChatGPT demonstrated a substantially lower accuracy of 50.0%. The authors concluded that chatbot's performance in oral and maxillofacial radiology was unsatisfactory and emphasized the need for verification of AI-generated responses.

The study was conducted with a relatively small sample size of 39 OPGs. The limited number of positive cases for certain conditions like mandibular fractures lead to artificially elevated values due to agreement on negative findings.

Future scope

Future studies should include larger data set with balanced representation of various oral surgical pathologies.

CONCLUSION

The findings demonstrated that ChatGPT-5 identified a significantly higher number of diagnostic and treatment related findings per radiograph compared to clinicians. While this suggests high sensitivity, qualitative and agreement analysis (Cohen's Kappa) revealed that many of these additional findings represented false positive interpretations, including misidentification of normal anatomical structures and radiographic artifacts, and unwarranted surgical recommendations.

Importantly, the inter-rater agreement among surgeons was nearly perfect for both diagnostic and treatment parameters, reinforcing the high reliability of the human reference standard used for comparison.

While its ability to recognize distinct radiographic features is promising, its overall diagnostic and treatment-planning accuracy remain insufficient for independent clinical use. Within the scope of this study, it may therefore be concluded that ChatGPT-5 possesses emerging but limited potential as an adjunctive tool in oral surgery.

Declaration

The authors declare the use of ChatGPT as research tools in the methodology of this study. The questions were entered into these models to generate data for comparative analysis. The human authors were responsible for the study design, data collection, analysis, and validation of all results. The authors did not use any generative AI or AI-assisted technology for the writing or editing of this manuscript.

Funding: No funding sources
Conflict of interest: None declared
Ethical approval: The study was approved by the Institutional Ethics Committee (reference number-IEC/GDCHN/93/2025)

REFERENCES

1. Xu Y, Liu X, Cao X, Huang C, Liu E, Qian S, et al. Artificial intelligence: A powerful paradigm for scientific research. Innovation. 2021;2(4):100179.
2. Iqbal U, Tanweer A, Rahmanti AR, Greenfield D, Lee LT, Li YJ. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. J Biomed Sci. 2025;32(1):45.
3. Wang Q, Amugo I, Rajakaruna H, Irudayam MJ, Xie H, Shanker A, et al. Evaluating GPT-5 for melanoma detection using dermoscopic images. Diagnostics. 2025;15(23):3052.
4. Choi WC. Chang CI. ChatGPT-5 in Education: New Capabilities and Opportunities for Teaching and Learning. Preprints.org10. 2025.
5. Dursun D, Bilici Geçer R. Dental age estimation from panoramic radiographs: a comparison of orthodontist and ChatGPT-4 evaluations using the London Atlas, Nolla, and Haavikko Methods. Diagnostics. 2025;15(18):2389.
6. World Health Organization, 2023; Nature Medicine, 2023; JAMA Network Open. 2023.
7. Izzetti R, Nisi M, Aringhieri G, Crocetti L, Graziani F, Nardi C. Basic knowledge and new advances in panoramic radiography imaging techniques: a narrative review on what dentists and radiologists should know. Applied Sciences. 2021;11(17):7858.
8. Chen J, Mullins CD, Novak P, Thomas SB. Personalized strategies to activate and empower patients in health care and reduce health disparities. Health Educ Behav. 2016;43:25-34.
9. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Fact. 2023;10:e47564.
10. Mentzou A, Rogers A, Carvalho E, Daly A, Malone M, Kerasidou X. Artificial intelligence in digital self-diagnosis tools: a narrative overview of

reviews. Mayo Clin Proc Digit Health. 2025;3(3):100242.
11. Kamerman J. AI in healthcare has moved past the hype- now the real work begins. Available from: https://www.usa.philips.com/healthcare/article/ai-in-healthcare-has-moved-past-the-hype-now-the-real-work-begins. Accessed on 13 August 2026.
12. Foresman G, Biro J, Tran A, MacRae K, Kazi S, Schubel L, et al. Patient perspectives on artificial intelligence in health care: focus group study for diagnostic communication and tool implementation. J Particip Med. 2025;17:e69564.
13. Suárez A, Arena S, Herranz Calzada A, Castillo Varón AI, Diaz-Flores García V, Freire Y. Decoding wisdom: evaluating ChatGPT's accuracy and reproducibility in analyzing orthopantomographic images for third molar assessment. Comput Struct Biotechnol J. 2025;28:141-7.
14. Jeong H, Han SS, Yu Y, Kim S, Jeon KJ. How well do large language model-based chatbots perform in oral and maxillofacial radiology? Dentomaxillofac Radiol. 2024;53(6):390-5.

Cite this article as: Jathar AA, Manekar VS, Bakale AM, Rewatkar K, Jathar AA. Can AI see what the surgeons see? Analyzing the accuracy of ChatGPT: publicly available multimodal AI for radiographic interpretation and treatment plan in oral surgery. Int J Community Med Public Health 2026;13:5130-5.