



## Research article

## Decoding wisdom: Evaluating ChatGPT's accuracy and reproducibility in analyzing orthopantomographic images for third molar assessment

Ana Suárez<sup>a</sup>, Stefania Arena<sup>a</sup>, Alberto Herranz Calzada<sup>a,c</sup>, Ana Isabel Castillo Varón<sup>b</sup>, Victor Diaz-Flores García<sup>c,\*</sup>, Yolanda Freire<sup>a</sup>

<sup>a</sup> Department of Pre-Clinic Dentistry II, Faculty of Biomedical and Health Sciences, Universidad Europea de Madrid, Calle Tajo s/n, Villaviciosa de Odón, Madrid 28670, Spain

<sup>b</sup> Department of Medicine, Faculty of Medicine, Health and Sports, Universidad Europea de Madrid, Calle Tajo s/n, Villaviciosa de Odón, Madrid 28670, Spain

<sup>c</sup> Department of Pre-Clinic Dentistry I, Faculty of Biomedical and Health Sciences, Universidad Europea de Madrid, Calle Tajo s/n, Villaviciosa de Odón, Madrid 28670, Spain

## ARTICLE INFO

## Keywords:

Artificial Intelligence  
ChatGPT  
Oral surgery  
Orthopantomography  
Dentistry

## ABSTRACT

The integration of Artificial Intelligence (AI) into healthcare has opened new avenues for clinical decision support, particularly in radiology. The aim of this study was to evaluate the accuracy and reproducibility of ChatGPT-4o in the radiographic image interpretation of orthopantomograms (OPGs) for assessment of lower third molars, simulating real patient requests for tooth extraction. Thirty OPGs were analyzed, each paired with a standardized prompt submitted to ChatGPT-4o, generating 900 responses (30 per radiograph). Two oral surgery experts independently evaluated the responses using a three-point Likert scale (correct, partially correct/incomplete, incorrect), with disagreements resolved by a third expert. ChatGPT-4o achieved an accuracy rate of 38.44 % (95 % CI: 35.27 %–41.62 %). The percentage agreement among repeated responses was 82.7 %, indicating high consistency, though Gwet's coefficient of agreement (60.4 %) suggested only moderate repeatability. While the model correctly identified general features in some cases, it frequently provided incomplete or fabricated information, particularly in complex radiographs involving overlapping structures or underdeveloped roots.

These findings highlight ChatGPT-4o's current limitations in dental radiographic interpretation. Although it demonstrated some capability in analyzing OPGs, its accuracy and reliability remain insufficient for unsupervised clinical use. Professional oversight is essential to prevent diagnostic errors. Further refinement and specialized training of AI models are needed to enhance their performance and ensure safe integration into dental practice, especially in patient-facing applications.

## 1. Introduction

The interpretation of radiographic images, particularly Orthopantomograms (OPGs) or panoramic radiographs, is a fundamental aspect of dental diagnostics, as it provides a comprehensive view of the oral cavity and adjacent structures [1]. Its accessibility, low cost and reduced radiation exposure makes it a widely used technique, particularly in the planning of complex surgical procedures such as the extraction of impacted third molars. It also allows the identification of

anatomical key factors that may influence the risk of intraoperative complications [2].

However, OPGs have inherent limitations, such as distortion and overlapping of anatomical structures, which can complicate their interpretation [3]. Although they serve as a basic diagnostic tool for the evaluation of impacted mandibular third molars, in certain cases they need to be complemented by three-dimensional imaging techniques, such as cone beam computed tomography (CBCT) [4]. CBCT provides more detailed information while respecting radiation protection

\* Correspondence to: Department of Pre-Clinic Dentistry I, School of Biomedical Sciences, Universidad Europea de Madrid, Calle Tajo s/n, Villaviciosa de Odón, Madrid 28670, Spain.

E-mail addresses: [ana.suarez@universidadeuropea.es](mailto:ana.suarez@universidadeuropea.es) (A. Suárez), [stefania.arena@universidadeuropea.es](mailto:stefania.arena@universidadeuropea.es) (S. Arena), [alberto.herranz@universidadeuropea.es](mailto:alberto.herranz@universidadeuropea.es) (A. Herranz Calzada), [anaisabel.castillo@universidadeuropea.es](mailto:anaisabel.castillo@universidadeuropea.es) (A.I. Castillo Varón), [victor.diaz-flores@universidadeuropea.es](mailto:victor.diaz-flores@universidadeuropea.es) (V. Diaz-Flores García), [yolanda.freire@universidadeuropea.es](mailto:yolanda.freire@universidadeuropea.es) (Y. Freire).

<https://doi.org/10.1016/j.csbj.2025.04.010>

Received 24 February 2025; Received in revised form 8 April 2025; Accepted 9 April 2025

Available online 11 April 2025

2001-0370/© 2025 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

principles [5].

In this context, general dentists have a key role, as they are often responsible for performing and interpreting OPGs in their own clinics and taking responsibility for analysing and explaining the results to patients [6]. At the same time, there has been a significant shift in patient behaviour, with individuals taking a more active role in managing their health, moving away from the traditional passive approach [7]. This phenomenon has been driven by the widespread availability of mobile devices and digital platforms that facilitate access to medical information, clinical data monitoring and health record management [8]. In fact, the Internet has become the primary resource that patients consult before seeking professional medical advice [9].

This massive access to information has led to a growing interest in independently understanding radiological images without the intervention of a healthcare professional [10]. This shift, driven by digital empowerment and the democratisation of medical knowledge [11], has encouraged the development of artificial intelligence (AI)-based tools that could facilitate the interpretation of health-related information.

In particular, large language models (LLMs) such as ChatGPT have shown significant potential in the medical field. These AI models, trained on large amounts of textual data, use advanced natural language processing techniques to generate coherent and contextually relevant responses [12]. Their ability to provide immediate answers, generate accessible explanations, and deliver individualized information has positioned AI as a valuable resource for searching and understanding medical information [13,14]. Recent studies have shown that online access to health-related information is more common among young people, women and those with higher levels of education [15], which has encouraged the use of digital platforms and virtual assistants to retrieve medical information quickly and efficiently [13,16]. This access to personalized information promotes greater autonomy in making informed healthcare decisions [14,17].

Beyond information retrieval, LLMs have begun to play an active role in multiple clinical applications in radiology, including decision support [18–20], clinical data analysis [19], medical report generation and simplification [19,21–26], workflow optimization [18] and promotion of responsible use of radiology services [18]. However, their use in medical image interpretation remains challenging, as professional supervision is required to prevent errors and ensure safe and effective use [27]. In dentistry, their implementation is still in its early stages, but their potential to transform AI-assisted diagnosis is significant [28–31].

Although deep learning architectures such as convolutional neural networks (CNNs) have demonstrated high performance in various image analysis tasks, their operation relies on large amounts of well-curated data and specific training protocols [32–36]. These characteristics differ considerably from those of LLMs, which were originally designed for natural language understanding and generation rather than image analysis [20]. However, the emergence of multimodal models such as ChatGPT-4o has extended these capabilities, as its transformer-based architecture enables the simultaneous integration of textual and visual information. This allows image interpretation with performance that in some cases approaches human-level accuracy [37].

While LLMs can interpret, generate and process multiple types of information, including text and images [38], their use in dentistry presents specific challenges. The lack of clinical context and the absence of professional judgement limit their ability to provide accurate interpretations tailored to specific situations. Furthermore, many of the limitations of these models are not fully understood by patients, increasing the risk of their use for non-optimized tasks. Despite these challenges, the acceptance of these models in clinical settings is largely dependent on perceptions of their reliability [39]. ChatGPT's ability to generate personalized explanations in clear and persuasive language may enhance this perception and improve the patient experience [12,23,39], but it may also create a false sense of accuracy. This highlights the need for professional oversight to ensure their safe use in healthcare settings.

In light of the above, this study aims to evaluate the accuracy and reproducibility of ChatGPT-4o, an LLM, in generating responses to enquiries about lower third molar radiographs, using a prompt that simulates a patient's question about the need to extract these teeth. Through this research, we aim to identify the capabilities and limitations of ChatGPT-4o in interpreting third molar orthopantomograms in dentistry to improve clinical practice and optimize the patient experience.

## 2. Materials and methods

### 2.1. Ethics committee

The study was approved by the Ethics Committee of the Universidad Europea de Madrid (approval code 2025–44) and was conducted in accordance with the ethical principles of the Declaration of Helsinki and current international regulations for research involving human subjects.

### 2.2. Radiographic images

The thirty OPGs used in this study were obtained from two private clinics. Patients gave written informed consent for their radiographs to be used in the analysis, after being fully informed of the aims and procedures of the study.

Data collection and processing was carried out in accordance with Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of personal data and the free movement of such data.

Radiographs were included in the study if they met the following criteria: i) presence of visible lower third molars on the radiograph, ii) adequate image quality, iii) absence of identifiable patient information such as name, surname or date of birth on the radiograph.

### 2.3. Generating responses in ChatGPT-4o

Two different Plus accounts were used to generate responses using the ChatGPT-4 model. Each of the 30 OPGs was presented with the following prompt: *"I have had an x-ray to look at my lower wisdom teeth in case I need to have them removed. In a short one-paragraph answer, tell me what you see."*

As ChatGPT's probabilistic algorithm can generate different responses to the same prompt [40], 30 responses were collected for each prompt and its corresponding x-ray. The responses were collected at three different times of the day (morning, afternoon and evening), always starting a 'new chat'. The entire data collection process was carried out from 9 December to 19 December 2024.

To ensure data confidentiality and anonymity, as well as the ethical and safe processing of radiographic images, the temporary chat feature in ChatGPT-4o was used for all sessions. This feature ensured that conversations were not permanently stored [41], thereby minimizing the risk of exposure or unauthorized access to personal or medical information.

All responses were stored in an Excel spreadsheet (Microsoft, Redmond, Washington, USA).

### 2.4. Expert assessment of responses

Two experts in oral surgery (S.A. and A.H.) independently scored the 900 responses using a 3-point Likert scale [42,43]. Table 1. A score of 2 was assigned to correct answers, 1 to partially correct or incomplete answers, and 0 to incorrect answers. In cases of disagreement between the raters, a third expert (A.S.) acted as arbitrator to resolve the discrepancy. Figs. 1–3.

### 2.5. Statistical analysis

All responses were analyzed using the statistical software STATA

**Table 1**  
Rubric used to assess the answers.

Experts' grading	Definition
Incorrect (0)	The answer is completely incorrect or unrelated to the question. It does not show adequate knowledge or understanding of the topic.
partially correct or Incomplete (1)	The answer reflects some knowledge or understanding of the topic, but contains significant errors or some essential elements are missing. It is not entirely incorrect, but the information is not sufficiently accurate or complete to be considered reliable or appropriate.
Correct (2)	The answer is completely correct and demonstrates a sound and accurate understanding of the topic. All key elements are addressed fully and accurately.

version BE 14 (StataCorp, Texas, EE. UU.).

For each of the 30 questions, the absolute (n) and relative (%) frequencies of the ratings assigned by the panel of experts were reported: 0 (incorrect), 1 (partially correct or incomplete) and 2 (correct). To assess the accuracy of ChatGPT-4o responses, the proportion of correct responses (rating of 2) out of the total number of questions was determined, along with its 95 % confidence interval, calculated using the Wald binomial method.

In terms of repeatability, the consistency of assigned ratings across replicates was examined using weighted agreement analysis for ordinal categories and multiple replicates. Both Percent Agreement and Gwet's Coefficient of Agreement were calculated, along with their respective 95 % confidence intervals. The calculated coefficients were categorized according to Gwet's benchmarking scale [44]: <0.0: poor, 0.0–0.2: slight, 0.2–0.4: fair, 0.4–0.6: moderate, 0.6–0.8: Substantial and 0.8–1.0: Almost perfect [42,43].

The level of agreement between the two expert reviewers' responses was assessed using the percentage of agreement and Gwet's AC coefficient (inter-rater reliability). Similarly, the level of agreement across the 30 repetitions of each image was assessed for each expert (intra-rater reliability).

3. Results

A total of 900 responses were obtained. The percentage of correct responses varied between 0 % and 100 % depending on the OPG. Table 2.

When the results were analyzed, the accuracy of ChatGPT-4o in image analysis was found to be 38.44 %, with a 95 % confidence interval ranging from 35.27 % to 41.62 %.

The percentage agreement between the responses across the 30 repetitions of each question was 82.7 %, indicating an almost perfect level of repeatability. However, Gwet's coefficient of agreement was 60.4 %, suggesting a moderate level of repeatability (Table 3).

The ratings given by Expert #1 and Expert #2 for the 900 responses generated by ChatGPT (30 questions + 30 repetitions) showed a percentage of agreement of 80.1 % (Gwet's AC1 = 49.9 %).

The ratings of the responses generated by ChatGPT for the 30 repetitions of the same image showed a percentage of agreement of 79.6 % for Expert #1 (Gwet's AC1 = 48.2 %) and 80.0 % for Expert #2 (Gwet's AC1 = 52.1 %).

4. Discussion

Large language models (LLMs), such as ChatGPT, have attracted increasing interest in medicine [45,46], particularly in radiology, due to their potential to support decision-making, data analysis and the generation of patient-friendly clinical reports [12,18,19,21–25], among other applications. However, although these models have primarily been developed for text processing [47], their potential application in medical image interpretation remains an evolving field, with several limitations identified in recent studies [48–51]. The development of ChatGPT has led to advances in reasoning and visual comprehension [52], opening up new opportunities for its application in clinical settings.

When comparing the results with previous studies, such as that of Silva et al. [48], who evaluated the ability of ChatGPT-3.5 to describe and suggest differential diagnoses for radiolucent lesions in OPGs, variable performance was observed. These authors, together with Wang

Chat temporal



I have had an x-ray to look at my lower wisdom teeth in case I need to have them removed. In a short one-paragraph answer, tell me what you see.

The X-ray shows evidence of multiple dental restorations and implants in both the upper and lower jaws. The lower wisdom teeth appear to be partially impacted, which could lead to complications such as infection, pain, or crowding. The bone structure around these teeth should be carefully assessed to determine the best approach for removal if necessary. A consultation with an oral surgeon is recommended for further evaluation and treatment planning.

Speech icons

**Fig. 1.** Screenshot of the ChatGPT-4o interface showing the model's response, which the expert panel scores as 0 due to incorrect information, as it incorrectly considers the third molars to be partially impacted. The model identifies the presence of dental restorations and implants in the patient.

Chat temporal ⓘ



I have had an x-ray to look at my lower wisdom teeth in case I need to have them removed. In a short one-paragraph answer, tell me what you see.

The panoramic dental X-ray reveals that the lower wisdom teeth (third molars) are impacted. On both the left and right sides, these wisdom teeth are positioned either vertically or at an angle, exerting pressure against the second molars. Additionally, the roots of the wisdom teeth are close to the mandibular nerve, which is an important factor to consider for any planned extraction. Additionally, the image shows wires, indicating ongoing orthodontic treatment, which may influence the decision for wisdom tooth extraction. Consulting an oral surgeon is recommended.

🔊 📄 ✎ ↺

**Fig. 2.** Screenshot of the ChatGPT-4o interface showing the model's response, which the expert panel scores as 1 for providing partially correct or incomplete information, stating that the third molars are in a vertical or angled position without specifying the correct one. The model is again able to detect previous orthodontic treatment.

et al. [51], suggest that the accuracy of these models in clinical tasks is influenced by their processing capacity and level of sophistication. For this reason, the present study used a more advanced model, ChatGPT-4o, which has enhanced image analysis capabilities [52]. Despite the improved architecture, consistent accuracy in radiographic interpretation of third molars cannot be guaranteed.

Thus, although ChatGPT-4o provided correct answers in a significant percentage of cases (38.44 %), its performance is still not robust enough for unsupervised clinical implementation. Furthermore, in cases where the error rate exceeded the accuracy rate, several factors were identified that may have affected the accuracy of the model, such as overlap with second molars or lack of root development. In other cases, although the model correctly identified the need for referral to a surgeon, it made errors in describing the position of the third molar, even though it was not specifically asked to do so. This suggests that in certain situations the model may be fabricated or making up responses [42,43,53], providing incorrect information that is not based on the analysed image data.

This high incidence of these made-up responses was also reported in a recent study evaluating the diagnostic accuracy of ChatGPT in chest radiographs, where the model achieved a diagnostic accuracy of 40.5 %–47 %. [49]. This reinforces the idea that ChatGPT-4o can provide general interpretations but still requires significant human supervision to prevent errors in clinical contexts, specially in complex cases or those with subtle radiological features.

It is important to highlight that although this study focused on evaluating the accuracy of ChatGPT-4o in individual responses, it did not take into account a decisive aspect of patient interaction: the conversational nature of the model. As ChatGPT is designed to encourage dialogue and continuous interaction, it is likely that a patient would continue to ask questions and provide additional information even if the initial response is only partially correct. On one hand, this

could increase the risk of confusion if inconsistent or incorrect responses are generated, but on the other hand, it may also allow for the collection of more patient data, potentially refining the response. However, this phenomenon was not analysed in the present study and is an important aspect for future research.

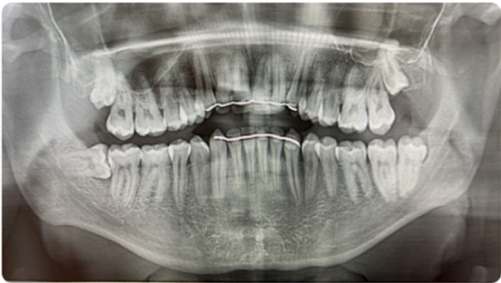
These results are consistent with those of Jeong et al. [54] who compared the accuracy of different chatbots (ChatGPT, ChatGPT Plus, Bard and Bing Chat) with that of dental students. Their results showed that chatbots achieved accuracy rates below 35 % in image interpretation. However, it is worth noting that this study was conducted without the use of actual images, relying solely on textual descriptions. In this study, although the model did not always achieve higher scores due to meandering responses - leading to partial inaccuracies - it demonstrated strong capabilities in correctly identifying the presence of implants, restorations and orthodontic treatments. This suggests that its ability to recognize specific dental treatments is more reliable than its ability to interpret the position of third molars.

A unique aspect of this study is that the prompt used simulated a real patient query about the need for extraction of lower third molars. In contrast to previous studies [48,54] that assessed ChatGPT's ability to interpret medical images based on descriptions or technical questions, in the present case the AI had to generate a response based on a query formulated in non-specialist terms. This approach is relevant as it reflects a potential application in interactions with patients seeking initial guidance on their treatment. However, to date, no similar studies have been identified that evaluate the performance of ChatGPT in this specific context, highlighting the need for further research to explore the impact of prompt formulation on the accuracy of generated responses.

In contrast to LLMs such as ChatGPT, which are not specifically designed for image analysis, deep learning models and convolutional neural networks (CNNs) can achieve levels of accuracy comparable to or



Chat temporal



I have had an x-ray to look at my lower wisdom teeth in case I need to have them removed. In a short one-paragraph answer, tell me what you see

This panoramic X-ray reveals the full dentition, including the lower wisdom teeth. Additionally, metallic hardware or wires are present, indicating previous dental or orthodontic treatments. The lower wisdom teeth (third molars) appear to be in close proximity to the mandibular nerve canal, which is visible as a faint linear structure near the roots of these teeth. The left lower wisdom tooth (on the right side of the image) has a horizontal angulation, while the right lower wisdom tooth appears to be vertical. Additionally, there appears to be some bone coverage over these teeth, which could complicate their extraction. A consultation with a dentist or oral surgeon is recommended for a more detailed evaluation and appropriate planning.

🔊 📄 ✎ 🔄

**Fig. 3.** Screenshot of the ChatGPT-4o interface showing the model’s response, which the expert panel scored as 2 for providing correct information. The model is again able to detect previous orthodontic treatment.

**Table 2**  
Expert assessment of ChatGPT-4o responses (30 repetitions per question).

OPG	Incorrect		Partially correct or Incomplete		Correct	
	n	%	n	%	n	%
1	14	46.67	16	53.33	0	0.00
2	1	3.33	4	13.33	25	83.33
3	0	0.00	7	23.33	23	76.67
4	1	3.33	18	60.00	11	36.67
5	10	33.33	19	63.33	1	3.33
6	10	33.33	10	33.33	10	33.33
7	6	20.00	19	63.33	5	16.67
8	0	0.00	25	83.33	5	16.67
9	0	0.00	7	23.33	23	76.67
10	1	3.33	12	40.00	17	56.67
11	0	0.00	1	3.33	29	96.67
12	1	3.33	2	6.67	27	90.00
13	8	26.67	18	60.00	4	13.33
14	0	0.00	0	0.00	30	100.00
15	11	36.67	2	6.67	17	56.67
16	2	6.67	14	46.67	14	46.67
17	4	13.33	22	73.33	4	13.33
18	4	13.33	19	63.33	7	23.33
19	8	26.67	22	73.33	0	0.00
20	2	6.67	21	70.00	7	23.33
21	2	6.67	26	86.67	2	6.67
22	2	6.67	26	86.67	2	6.67
23	8	26.67	22	73.33	0	0.00
24	1	3.33	5	16.67	24	80.00
25	6	20.00	10	33.33	14	46.67
26	3	10.00	27	90.00	0	0.00
27	1	3.33	21	70.00	8	26.67
28	6	20.00	21	70.00	3	10.00
29	1	3.33	16	53.33	13	43.33
30	0	0.00	9	30.00	21	70.00

**Table 3**  
Agreement between repetitions of answers. <0.0: poor, 0.0–0.2: slight, 0.2–0.4: fair, 0.4–0.6: moderate, 0.6–0.8: Substantial and 0.8–1.0: Almost perfect.

	Coefficient	95 % Confidence Interval		Benchmark scale
Percent Agreement	0.827	0.790	0.8634	Almost Perfect
Gwet’s AC	0.604	0.503	0.705	Moderate

better than human specialists in certain diagnostic tasks [32–36]. In the field of oral surgery, several studies have shown that CNNs can match the performance of oral and maxillofacial surgeons in detecting peri-apical disease [33], odontogenic lesions [34], and identifying ameloblastomas and keratocysts [35]. However, these studies consistently indicate that although CNNs have shown promising performance, their practical effectiveness depends largely on the availability of large, well-structured and accurately labelled clinical datasets [32–36]. However, this remains a significant challenge, as collecting and organizing such datasets is difficult due to the heterogeneity of clinical cases, variations in image quality, and legal restrictions that limit access to properly labelled medical data.

Among AI studies based on CNNs for third molar analysis, models designed to assess the need for odontosection and osteotomy in third molar extractions stand out, as they have achieved high sensitivity but have also shown a tendency towards overdiagnosis [36]. In addition, models that can accurately classify the position of the third molar based on angulation and stage of eruption have been developed, demonstrating that their performance is more effective in specific tasks rather than multitasking approaches [32].

Similarly, in this study, ChatGPT-4o struggled to analyze

radiographs containing multiple elements, often providing answers that mixed correct data with incorrect information in almost 50 % of cases. This suggests that, like CNNs in complex tasks, its accuracy is variable and may be affected by the interaction between different features within the image.

In this context, recent efforts such as the validation of the Quality Analysis of Medical Artificial Intelligence (QAMAI) tool highlight the importance of having a standardised framework for assessing not only the diagnostic accuracy, but also the overall quality and reliability of health information generated by AI systems such as ChatGPT [55].

To overcome these limitations and improve the applicability of AI in dentistry, it is essential to optimize data collection and develop models trained on large and well-structured dental databases. However, this remains a major challenge for the clinical implementation of these technologies [56].

Future studies will include a larger and more diverse set of radiographs to further explore the performance of ChatGPT-4o in a wider range of clinical scenarios.

## 5. Conclusion

Although ChatGPT-4o has shown promising performance in interpreting orthopantomograms, its limitations in accuracy and occasional generation of incorrect information prevent its autonomous clinical use. Expert supervision remains essential to avoid diagnostic errors and ensure safe and reliable interpretations.

This study's findings reinforce the notion that while ChatGPT-4o appears capable of identifying certain dental treatments, its ability to respond correctly to patient queries remains inconsistent. Training the model with more specialized datasets could potentially improve its performance.

Given that this study simulated a real patient query about the need for a lower third molar extraction, we believe it is essential for professional associations, scientific societies and health authorities to promote awareness campaigns aimed at both healthcare professionals and the general public. These initiatives should emphasize that this technology cannot currently replace specialist assessment, particularly in contexts where radiological interpretation is vital for diagnosis and treatment planning.

## CRedit authorship contribution statement

**Castillo Varón Ana Isabel:** Writing – review & editing, Formal analysis. **Díaz-Flores García Victor:** Writing – review & editing, Resources, Methodology, Investigation. **Suárez Ana:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Arena Stefania:** Writing – review & editing, Investigation. **Herranz Calzada Alberto:** Writing – review & editing, Investigation. **Freire Yolanda:** Writing – review & editing, Resources, Methodology, Investigation.

## References

- [1] Rushton MN, Rushton VE. A study to determine the added value of 740 screening panoramic radiographs compared to intraoral radiography in the management of adult (>18 years) dentate patients in a primary care setting. *J Dent* 2012;40:661–9. <https://doi.org/10.1016/j.jdent.2012.04.018>.
- [2] Jerjes W, El-Maaytah M, Swinson B, et al. Inferior alveolar nerve injury and surgical difficulty prediction in third molar surgery: the role of dental panoramic tomography. *J Clin Dent* 2006;17:122–30.
- [3] Cardoso GB, Savegnago GDO, Hirsch WDB, et al. Pathologic conditions associated with impacted third molars: a retrospective study of panoramic radiographs in a Southern Brazilian population. *Imaging Sci Dent* 2023;53:303. <https://doi.org/10.5624/isd.20230036>.
- [4] Gupta S, Bhowate RR, Nigam N, Saxena S. Evaluation of impacted mandibular third molars by panoramic radiography. *ISRN Dent* 2011;2011:1–8. <https://doi.org/10.5402/2011/406714>.
- [5] Matzen LH, Wenzel A. Efficacy of CBCT for assessment of impacted mandibular third molars: a review – based on a hierarchical model of evidence. *Dentomaxillofac Radiol* 2015;44:20140189. <https://doi.org/10.1259/dmfr.20140189>.
- [6] Izzetti R, Nisi M, Aringhieri G, et al. Basic knowledge and new advances in panoramic radiography imaging techniques: a narrative review on what dentists and radiologists should know. *Appl Sci* 2021;11:7858. <https://doi.org/10.3390/app1177858>.
- [7] 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. <https://doi.org/10.1177/1090198115579415>.
- [8] Piras EM, Zanutto A. Prescriptions, X-rays and grocery lists. designing a personal health record to support (The Invisible Work Of) health information management in the household. *Comput Support Coop Work (CSCW)* 2010;19:585–613. <https://doi.org/10.1007/s10606-010-9128-5>.
- [9] Wang L, Wang J, Wang M, et al. Using internet search engines to obtain medical information: a comparative study. *J Med Internet Res* 2012;14:e74. <https://doi.org/10.2196/jmir.1943>.
- [10] Ellenbogen AL, Patrie JT, Gaskin CM. Improving patient access to medical images by integrating an imaging portal with the electronic health record patient portal. *J Am Coll Radiol* 2021;18:864–7. <https://doi.org/10.1016/j.jacr.2020.12.028>.
- [11] Hickmann E, Richter P, Schlieter H. All together now – patient engagement, patient empowerment, and associated terms in personal healthcare. *BMC Health Serv Res* 2022;22:1116. <https://doi.org/10.1186/s12913-022-08501-5>.
- [12] Akinci D'Antonoli T, Stanzone A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn Interv Radiol* 2024;30:80–90. <https://doi.org/10.4274/dir.2023.232417>.
- [13] Nov O, Singh N, Mann DM. Putting ChatGPT's medical advice to the (Turing) test. *JMIR Med Educ* 2023;9:e46939. <https://doi.org/10.1101/2023.01.23.23284735>.
- [14] Neha F, Bhati D, Shukla DK, Amiruzzaman M. ChatGPT: transforming healthcare with AI. *AI* 2024;5:2618–50. <https://doi.org/10.3390/ai5040126>.
- [15] Wang X, Shi J, Kong H. Online health information seeking: a review and meta-analysis. *Health Commun* 2021;36:1163–75. <https://doi.org/10.1080/10410236.2020.1748829>.
- [16] Onyeaka HK, Zambrano J, Longley RM, et al. Use of digital health tools for health promotion in cancer survivors. *Psychooncology* 2021;30:1302–10. <https://doi.org/10.1002/pon.5677>.
- [17] Xu T, Weng H, Liu F, et al. Current status of ChatGPT use in medical education: potentials, challenges, and strategies. *J Med Internet Res* 2024;26:e57896. <https://doi.org/10.2196/57896>.
- [18] Rao A, Kim J, Kamineni M, et al. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 Versus GPT-3.5 in a breast imaging pilot. *J Am Coll Radiol* 2023;20:990–7. <https://doi.org/10.1016/j.jacr.2023.05.003>.
- [19] Lecler A, Duron L, Soyer P. Revolutionizing radiology with GPT-based models: Current applications, future possibilities and limitations of ChatGPT. *Diagn Inter Imaging* 2023;104:269–74. <https://doi.org/10.1016/j.diii.2023.02.003>.
- [20] Frosolini A, Catarzi L, Benedetti S, et al. The role of large language models (LLMs) in providing triage for maxillofacial trauma cases: a preliminary study. *Diagnostics* 2024;14:839. <https://doi.org/10.3390/diagnostics14080839>.
- [21] Lyu Q, Tan J, Zapadka ME, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. *Vis Comput Ind Biomed Art* 2023;6:9. <https://doi.org/10.1186/s42492-023-00136-5>.
- [22] Chung EM, Zhang SC, Nguyen AT, et al. Feasibility and acceptability of ChatGPT generated radiology report summaries for cancer patients. *Digit Health* 2023;9. <https://doi.org/10.1177/20552076231221620>.
- [23] Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *Eur Radio* 2023;34:2817–25. <https://doi.org/10.1007/s00330-023-10213-1>.
- [24] Naik HR, Prather AD, Gurda GT. Synchronous bilateral breast cancer: a case report piloting and evaluating the implementation of the ai-powered large language model (LLM) ChatGPT. *Cureus* 2023. <https://doi.org/10.7759/cureus.37587>.
- [25] Bosbach WA, Senge JF, Nemeth B, et al. Ability of ChatGPT to generate competent radiology reports for distal radius fracture by use of RSNA template items and integrated AO classifier. *Curr Probl Diagn Radio* 2024;53:102–10. <https://doi.org/10.1067/j.cpradiol.2023.04.001>.
- [26] Amedu C, Ohene-Botwe B. Harnessing the benefits of ChatGPT for radiography education: a discussion paper. *Radiography* 2024;30:209–16. <https://doi.org/10.1016/j.radi.2023.11.009>.
- [27] Tian D, Jiang S, Zhang L, et al. The role of large language models in medical image processing: a narrative review. *Quant Imaging Med Surg* 2024;14:1108–21. <https://doi.org/10.21037/qims-23-892>.
- [28] Maniati A, Fakhry N, Chiesa-Estomba C, et al. Synergizing ChatGPT and general AI for enhanced medical diagnostic processes in head and neck imaging. *Eur Arch Oto-Rhino-Laryngol* 2024;281:3297–8. <https://doi.org/10.1007/s00405-024-08511-5>.
- [29] Bajaj S, Gandhi D, Nayar D. Potential applications and impact of ChatGPT in radiology. *Acad Radio* 2024;31:1256–61. <https://doi.org/10.1016/j.acra.2023.08.039>.
- [30] Koga S, Du W. From text to image: challenges in integrating vision into ChatGPT for medical image interpretation. *Neural Regen Res* 2025;20:487–8. <https://doi.org/10.4103/NRR.NRR-D-24-00165>.
- [31] Gokdeniz ST, Kamburoğlu K. Artificial intelligence in dentomaxillofacial radiology. *World J Radio* 2022;14:55–9. <https://doi.org/10.4329/wjr.v14.i3.55>.
- [32] Sukegawa S, Matsuyama T, Tanaka F, et al. Evaluation of multi-task learning in deep learning-based positioning classification of mandibular third molars. *Sci Rep* 2022;12:684. <https://doi.org/10.1038/s41598-021-04603-y>.

- [33] Endres MG, Hillen F, Salloumis M, et al. Development of a deep learning algorithm for periapical disease detection in dental radiographs. *Diagnostics* 2020;10:430. <https://doi.org/10.3390/diagnostics10060430>.
- [34] Yang H, Jo E, Kim HJ, et al. Deep learning for automated detection of cyst and tumors of the jaw in panoramic radiographs. *J Clin Med* 2020;9:1839. <https://doi.org/10.3390/jcm9061839>.
- [35] Poedjiastoeti W, Suebnukarn S. Application of convolutional neural network in the diagnosis of jaw tumors. *Health Inf Res* 2018;24:236. <https://doi.org/10.4258/hir.2018.24.3.236>.
- [36] Danjo A, Kuwada C, Aijima R, et al. Limitations of panoramic radiographs in predicting mandibular wisdom tooth extraction and the potential of deep learning models to overcome them. *Sci Rep* 2024;14:30806. <https://doi.org/10.1038/s41598-024-81153-z>.
- [37] OpenAI. 2023. GPT-4 Technical Report. arXiv preprint arXiv 2303.08774 .
- [38] Liu Y, Han T, Ma S, et al. Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiol* 2023;1:100017. <https://doi.org/10.1016/j.metrad.2023.100017>.
- [39] Alanezi F. Factors influencing patients' engagement with ChatGPT for accessing health-related information. *Crit Public Health* 2024;34:1–20. <https://doi.org/10.1080/09581596.2024.2348164>.
- [40] Xie Y, Seth I, Hunter-Smith DJ, et al. Aesthetic surgery advice and counseling from artificial intelligence: a rhinoplasty consultation with ChatGPT. *Aesthetic Plast Surg* 2023. <https://doi.org/10.1007/s00266-023-03338-7>.
- [41] OpenAI Temporary Chat FAQ. In: ([https://help.openai.com/en/articles/8914046-temporary-chat-faq?utm\\_source=chatgpt.com](https://help.openai.com/en/articles/8914046-temporary-chat-faq?utm_source=chatgpt.com)).
- [42] Suárez A, Jiménez J, de Pedro ML, et al. Beyond the scalpel: assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery. *Comput Struct Biotechnol J* 2023. <https://doi.org/10.1016/j.csbj.2023.11.058>.
- [43] Freire Y, Santamaría Laorden A, Orejas Pérez J, et al. ChatGPT performance in prosthodontics: assessment of accuracy and repeatability in answer generation. *J Prosthet Dent* 2024;131:659.e1–6. <https://doi.org/10.1016/j.prosdent.2024.01.018>.
- [44] Gwet K.L. 2014 Handbook of Inter-Rater Reliability. Maryland.
- [45] Chow JCL, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. *Front Artif Intell* 2023;6. <https://doi.org/10.3389/frai.2023.1166014>.
- [46] Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 2023;388:1233–9. <https://doi.org/10.1056/NEJMs2214184>.
- [47] Hu M., Pan S., Li Y., Yang X. 2023. . Advancing Medical Imaging with Language Models: A Journey from N-grams to ChatGPT. ArXiv 20232304.04920.
- [48] Silva TP, Andrade-Bortoletto MFS, Ocampo TSC, et al. Performance of a commercially available generative pre-trained transformer (GPT) in describing radiolucent lesions in panoramic radiographs and establishing differential diagnoses. *Clin Oral Invest* 2024;28:204. <https://doi.org/10.1007/s00784-024-05587-5>.
- [49] Lee KH, Lee RW, Kwon YE. Validation of a deep learning chest X-ray interpretation model: integrating large-scale AI and large language models for comparative analysis with ChatGPT. *Diagnostics* 2023;14:90. <https://doi.org/10.3390/diagnostics14010090>.
- [50] Srivastav S, Chandrakar R, Gupta S, et al. ChatGPT in radiology: the advantages and limitations of artificial intelligence for medical imaging diagnosis. *Cureus* 2023. <https://doi.org/10.7759/cureus.41435>.
- [51] Wang S, Zhao Z, Ouyang X, et al. Interactive computer-aided diagnosis on medical image using large language models. *Commun Eng* 2024;3:133. <https://doi.org/10.1038/s44172-024-00271-8>.
- [52] OpenAI. 2024. Announcing GPT-4o in the API! In: (<https://community.openai.com/t/announcing-gpt-4o-in-the-api/744700>).
- [53] Suárez A, Díaz-Flores García V, Algar J, et al. Unveiling the ChatGPT phenomenon: Evaluating the consistency and accuracy of endodontic question answers. *Int Endod J* 2023. <https://doi.org/10.1111/iej.13985>.
- [54] Jeong H, Han S-S, Yu Y, et al. How well do large language model-based chatbots perform in oral and maxillofacial radiology? *Dentomaxillofac Radiol* 2024;53: 390–5. <https://doi.org/10.1093/dmfr/twae021>.
- [55] Vaira LA, Lechien JR, Abbate V, et al. Validation of the quality analysis of medical artificial intelligence (QAMAI) tool: a new tool to assess the quality of health information provided by AI platforms. *Eur Arch Oto-Rhino-Laryngol* 2024;281: 6123–31. <https://doi.org/10.1007/s00405-024-08710-0>.
- [56] Uribe SE, Issa J, Sohrabniya F, et al. Publicly available dental image datasets for artificial intelligence. *J Dent Res* 2024;103:1365–74. <https://doi.org/10.1177/00220345241272052>.