Evaluation of Quality of AI Chatbots in Patient Education Guidelines on Dental Implant Maintenance

Rejina Shrestha[1], Amar Bhochhibhoya[2], Sahara Shrestha[3]
[1]Dental surgeon, Kanti Children's Hospital, Maharajgunj, Kathmandu, Nepal
[2]Assistant Professor, Maharajgunj Medical Campus, Maharajgunj, Kathmandu. Nepal
[3]Assistant Professor, Kirtipur Ayurveda Hospital, Kirtipur, Nepal

ABSTRACT

Introduction: The use of Artificial intelligence (AI) chatbots in medical science has been ever increasing. Patients rely on the information provided by these chatbots. The assessment of these chatbots is of absolute necessity for the detection of accuracy of their responses. The patient education guidelines have to be clear and useful in a structured and understandable sentence pattern, along with proper referencing.

Methods: Patient education guideline on dental implant was generated. Two investigators scored the chatbots based on Quality Analysis of Medical Artificial Intelligence (QAMAI) scores of two chatbots: Chat GPT and Claude. The dimensions detected were accuracy, clarity, relevance, completeness, references and usefulness. Statistical analysis were carried out by unpaired t test.

Results: The QAMAI scores for both Chat GPT and Claude were good. There were no statistically different differences among the QAMAI score of the two chatbots. (p=0.73)

Conclusion: The patient education guide provided by Chat GPT and Claude are satisfactory. There is, however, room for improvement in increasing the quality of the AI chatbots.

Key words: Artificial intelligence, Dental implant, Maintenance, QAMAI score

INTRODUCTION

The placement of dental implant for the replacement of the missing tooth is a reliable option in terms of longevity and stability.[1,2] It has revolutionized the field of dentistry offering remarkable clinical success. With the high rise in the use of dental implants, patient education on the dental implant maintenance forms the core component in its success. The education on prevention and intervention of peri-implant diseases must be provided to each and every implant patient as it is a long-term commitment which requires effort and continuous motivation.

The two main components of implant maintenance include patient-performed home care and professional interventions. Maintaining excellent oral hygiene is critical and includes common hygiene practices like brushing, flossing, use of interdental brushes and mouthwashes. Professional maintenance includes Supportive Peri-implant Care (SPIC). A thorough examination and recording are done followed by mechanical implant cleaning with titanium or plastic-curettes, ultrasonics or air polishing. These strategies prevent the onset of peri-implant diseases, which are peri-implant mucositis and peri-implantitis.[3]

The easily accessible information forums nowadays for the patients are AI chatbots. These large language models (LLMs) are not only

Conflict of Interest: None

***Corresponding Author**
Dr. Amar Bhochhibhoya,
Lecturer, Department of Prosthodontics,
Maharajgunj Medical Campus, Maharajgunj,
Kathmandu, Nepal.
Phone No.: 9804320719
E-mail: amarbhochhibhoya@gmail.com

limited to disseminating information but also showcase advanced reasoning.[4,5] Among the many chatbots available, Chat GPT and Claude are extensively used for their multimodal interactions and versatility. ChatGPT was developed by OpenAI and Claude was created by Anthropic. Both the chatbots have been claimed to be excellent in terms of logic, reasoning and image analysis.

The reliability of these chatbots should be thoroughly checked, specially when being concerned about sensitive issues, primarily health.[6,7,8] With AI hallucinations being highly prevalent, the performance of AI chatbots become questioning for health focused questions.[9] The present study was conducted to explore the reliability of AI chatbots based on QAMAI tools.

METHODS

This is an original cross-sectional study designed to evaluate quality of AI-generated responses to patient education guides related to dental implant maintenance. The duration of study was from January, 2025 to March, 2025. The data was entirely derived from two different AI tools: ChatGPT (GPT-4o) and Claude by Anthropic by a single data extractor (S.S). Two investigators (R.S and A.B) scored the quality of the chatbots based on QAMAI tool.[10] The QAMAI tool consists of 6 parameters, accuracy, clarity, relevance, completeness, references and usefulness. It is based on 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree). The total QAMAI score was interpreted according to its classification (Table No. 1)

Data collection

The prompt was presented to two free version of AI programs, ChatGPT (version 4.o) and Claude by Anthropic on the same day. Any previous prompts were deleted to avoid the influence of the previous prompts on the response. Both the AI tools were asked to produce patient education guides for the four diseases in their default and standard settings, without any fine-tuning or modifications, using the same prompts to ensure consistency in the inputs: “Write a patient education guide for dental implant maintenance”. The generated responses were recorded in Microsoft Word documents. The outputs were subsequently assessed with the Quality Analysis of Medical Artificial Intelligence (QAMAI) score. Two observers assessed the QAMAI score independently with good agreement (Cohen’s Kappa=0.85). The disagreements were settled by consensus decision before the final analysis.

Statistical analysis

The collected data was imported into Microsoft Excel and analyzed using SPSS version 21. The inter-observer agreement for the QAMAI scores was assessed using Cohen’s Kappa coefficient. An unpaired t-test was used to assess the quality of the AI-generated responses from the two tools, with a p-value < 0.05 being considered statistically significant.

RESULTS

Among all the domains examined, the average score of accuracy, clarity, relevance and usefulness was found to be higher in ChatGPT whereas, the average score of completeness and sources of information was found to be higher in Claude. A graphical comparison of parameters for the patient education guides created by Claude and ChatGPT is presented in Figure 1.

Among the 6 parameters of QAMAI score, Chat GPT showed higher performance in 4 parameters. When comparing the QAMAI score of both the chatbots, there was no statistically significant difference among the QAMAI scores of the 2 chatbots. Both the chatbots were classified under good quality AI. This means that the AI system provides mostly reliable and complete information, but there may be some areas for refinement.

Table 1: Classification of QAMAI Score.

Score	Classification	Description
6–11 points	Poor quality	The AI system provides information that is largely unreliable or incomplete. Immediate improvement is required
12–17 points	Fair quality	The AI system provides some useful information, but there are significant areas for improvement
18–23 points	Good quality	The AI system provides mostly reliable and complete information, but there may be some areas for refinement
24–29 points	Very good quality	The AI system provides reliable and complete information in most areas. There are minor areas for improvement
30 points	Excellent quality	The AI system provides highly reliable and complete information

Table 2: QAMAI Scores produced by ChatGPT and Claude.

S. No.	AI chatbot	Quality	Score	p-value
1	Chat GPT	Good	21.5	0.73
2	Claude	Good	22	

*p-values <0.05 are considered statistically significant.

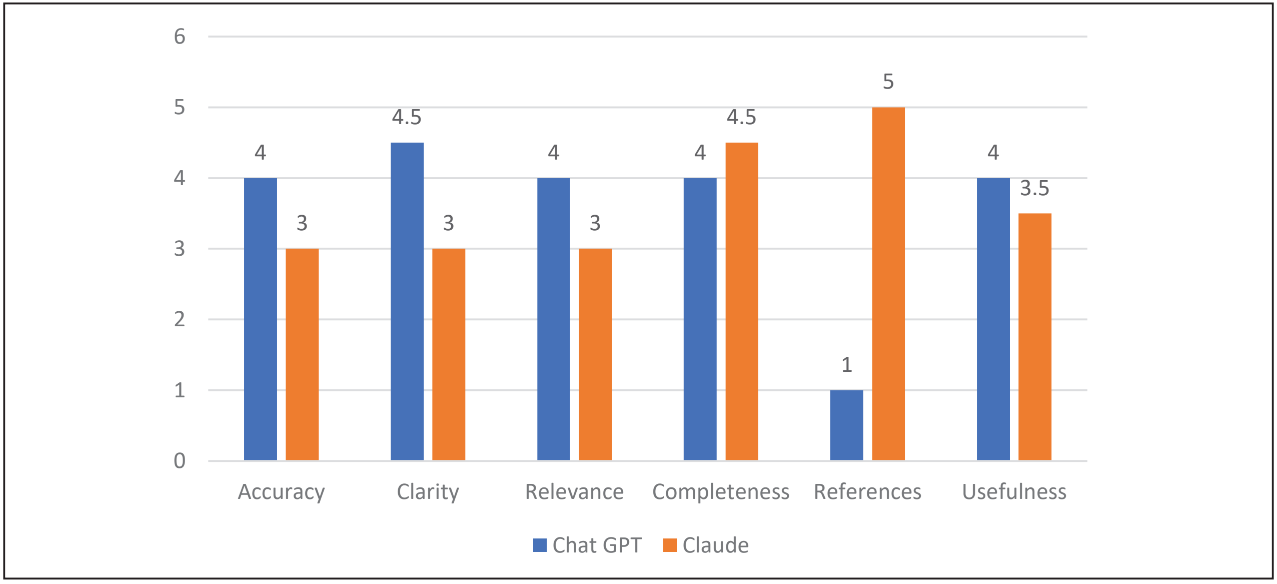

***Figure 1: Comparison** of the accuracy, clarity, relevance, completeness, references and usefulness for the patient education guides produced by ChatGPT and Claude.*

DISCUSSION

This cross-sectional study compared responses generated by Claude and ChatGPT for patient education guides on dental implant maintenance guidelines. There were no statistically significant differences in score among the two chatbots. (p-value= 0.73)

QAMAI (Quality Assessment of Medical Artificial Intelligence) is a validated tool designed to assess the quality of health-related information generated by AI systems. The instrument evaluates six parameters, with a minimum score of 6 and maximum score of 30.[10] Both Chat GPT and Claude showed good quality of information. This shows that there is room for enhancement. Other popular tool for the assessment of quality of AI tools is the DISCERN tool.[11]

Among all the parameters, the referencing of Chat GPT was found to be very weak. Artificial hallucination is the term used in AI in which the chatbots confidently generate plausible but incorrect information.[9] This fact decreases the credibility of the AI chatbots. Thus, artificial intelligence may assist as a supplementary tool but cannot substitute humans on the decisive roles.[12]

It is important for the dental clinician to be well acquainted with the information being provided to the patients. The extravagant use of AI chatbots should not be undermined. The online extraction of information is being popular due to easy accessibility, the availability on the internet without additional cost and the ability to remain anonymous.[13] Caution should be practiced from both the clinician and the patient on determining the degree of dependability on this information.

Based on the dental implant maintenance, personal and professional thorough cleaning, early diagnosis and treatment of peri-implant diseases, reduction of risk factors and the maintenance of the prostheses forms the foundation of healthy peri-implant tissue and stable periimplantitis. Every effort should be applied to halt the disease progression at its reversible condition, which is peri-implant mucositis.[14] Once bone loss is evident and peri-implantitis occurs, therapy should be aggressive.[15] Proper maintenance of the dental implant helps the patient to institute healthy peri-implant tissue and reduce the financial, physical and mental burden associated with the progression of the disease.

This study shows that the quality of medical artificial intelligence needs further improvement. The study was limited to two AI tools, ChatGPT 4.o and Claude. Furthermore, the analysis was limited to English-language content. Investigations must be done including different chatbots to determine the chatbot with the most advanced and superior education material quality.

Moreover, the study highlights that AI-generated educational materials often are unable to meet ideal standards for patient education due to factual inaccuracies, inconsistencies, and potential to misinformation—risks that are especially concerning in medical contexts.[16] Therefore, relying solely on AI tools for medical content creation is inappropriate; human oversight is essential to ensure accuracy and relevance to current clinical practice. The authors recommend that future research should center on boosting the trustworthiness of AI tools and establishing robust evaluation frameworks to guide their medical use. Addressing these issues is vital to ensuring AI-generated medical content meets ethical and educational standards.

CONCLUSION

This study set out to compare two AI platforms—ChatGPT 4.0 and Claude—in their ability to produce patient education guides on implant maintenance. The comparison covered aspects such as accuracy, clarity, relevance, completeness, referenced and usefulness in content. The findings revealed minimal differences between the two tools when measured across the QAMAI score. However, as the research only evaluated these two specific AI models, its conclusions aren't necessarily applicable to the broader field. It is critical to expand such studies to include a wider range of AI systems to enhance validity and generalizability.

Acknowledgement: None.

REFERENCES

1. Hjalmarsson L, Gheisarifar M, Jemt T. A Systematic Review of Survival of Single Implants as Presented in Longitudinal Studies with a Follow-Up of at Least 10 Years. European

Journal of Oral Implantology 2016;9: S155–S162.

2. Jemt T. Single-Implant Survival: More Than 30 Years of Clinical Experience. International Journal of Prosthodontics 2016;6: 551–558.
3. Herrera D, Berglundh T, Schwarz F, Chapple I, Jepsen S, Sculean A, Kebschull M, Papapanou PN, Tonetti MS, Sanz M, EFP Workshop Participants and Methodological Consultant. Prevention and treatment of peri-implant diseases—The EFP S3 level clinical practice guideline. Journal of Clinical Periodontology. 2023;50:4-76.
4. *Flaharty KA, Hu P, Hanchard SL, et al. Evaluating large language models on medical, lay-language, and self-reported descriptions of genetic conditions. Am J Hum Genet. 2024 Sep 5;111(9):1819–1833. doi: 10.1016/j.ajhg.2024.07.011. doi. Medline.* [DOI] [PMC free article] [PubMed] [Google Scholar]
5. *Rengers TA, Thiels CA, Salehinejad H. Academic Surgery in the Era of Large Language Models: A Review. JAMA Surg. 2024 Apr 1;159(4):445–450. doi: 10.1001/jamasurg.2023.6496. doi. Medline.*
6. Liu Y, Li H, Ouyang J, Xue Z, Wang M, He H, Song B, Zheng X, Gan W. Evaluating Large Language Models for Preoperative Patient Education in Superior Capsular Reconstruction: Comparative Study of Claude, GPT, and Gemini. JMIR Perioper Med. 2025 Jun 12;8:e70047. doi: 10.2196/70047. PMID: 40505086; PMCID: PMC12178570.
7. *Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries About Cancer. JAMA Oncol. 2023 Oct 1;9(10):1437–1440. doi: 10.1001/jamaoncol.2023.2947. doi. Medline.* [DOI] [PMC free article] [PubMed] [Google Scholar]
8. *Xue Z, Zhang Y, Gan W, Wang H, She G, Zheng X. Quality and Dependability of ChatGPT and DingXiangYuan Forums for Remote Orthopedic Consultations: Comparative Analysis. J Med Internet Res. 2024 Mar 14;26:e50882. doi: 10.2196/50882. doi. Medline.* [DOI] [PMC free article] [PubMed] [Google Scholar]
9. Chelli M, Descamps J, Lavoué V, et al. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. J Med Internet Res. 2024 May 22;26:e53164. doi: 10.2196/53164. doi. Medline. [DOI] [PMC free article] [PubMed] [Google Scholar]
10. Vaira LA, Lechien JR, Abbate V, Allevi F, Audino G, 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 Otorhinolaryngol. 2024 Nov;281(11):6123-6131. doi: 10.1007/s00405-024-08710-0. Epub 2024 May 4. PMID: 38703195; PMCID: PMC11512889.
11. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health. 1999 Feb;53(2):105-11. doi: 10.1136/jech.53.2.105. PMID: 10396471; PMCID: PMC1756830.
12. 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 May 7;32(1):45. doi: 10.1186/s12929-025-01131-z. PMID: 40335969; PMCID: PMC12057020.
13. Banerjee P, Puri A, Mathur R. Enhancing mental health referrals through AI-enabled chatbots. J Psychol Clin Psychiatry. 2024;15(6):299-301.
14. Heitz-Mayfield LJA, Salvi GE. Peri-implant mucositis. J Clin Periodontol. 2018 Jun;45 Suppl 20:S237-S245. doi: 10.1111/jcpe.12953. PMID: 29926488.
15. Schwarz F, Derks J, Monje A, Wang HL. Peri-implantitis. J Periodontol. 2018 Jun;89 Suppl 1:S267-S290. doi: 10.1002/JPER.16-0350. PMID: 29926957.
16. Saji JG, Balagangatharan A, Bajaj S, Swarnkar V, Unni D, Dileep A. Analysis of patient education guides generated by ChatGPT and Gemini on common anti-diabetic drugs: A cross-sectional study. Cureus. 2025 Mar 25;17(3).