Original Article

Check for updates

pISSN 2234-7518 • eISSN 2005-372X
https://doi.org/10.4041/kjod24.106
Korean J Orthod 2025;55(2):131-141

THE KOREAN JOURNAL of ORTHODONTICS
KJO

Artificial intelligence solutions for temporomandibular joint disorders: Contributions and future potential of ChatGPT

Betul Kula[a]
Ahmet Kula[b]
Fatih Bagcier[c]
Bulent Alyanak[d]

[a]Department of Orthodontics, Istanbul Galata University, Istanbul, Türkiye
[b]Department of Prosthodontics, Uskudar University, Istanbul, Türkiye
[c]Physical Medicine and Rehabilitation Clinic, Basaksehir Cam and Sakura City Hospital, Istanbul, Türkiye
[d]Department of Physical Medicine and Rehabilitation, Golcuk Necati Celik State Hospital, Kocaeli, Türkiye

Objective: This study aimed to evaluate the reliability and usefulness of information generated by Chat Generative Pre-Trained Transformer (ChatGPT) on temporomandibular joint disorders (TMD). **Methods:** We asked ChatGPT about the diseases specified in the TMD classification and scored the responses using Likert reliability and usefulness scales, the modified DISCERN (mDISCERN) scale, and the Global Quality Scale (GQS). **Results:** The highest Likert scores for both reliability and usefulness were for masticatory muscle disorders (mean ± standard deviation [SD]: 6.0 ± 0), and the lowest scores were for inflammatory disorders of the temporomandibular joint (mean ± SD: 4.3 ± 0.6 for reliability, 4.0 ± 0 for usefulness). The median Likert reliability score indicates that the responses are highly reliable. The median Likert usefulness score was 5 (4–6), indicating that the responses were moderately useful. A comparative analysis was performed, and no statistically significant differences were found in any subject for either reliability or usefulness (P = 0.083–1.000). The median mDISCERN score was 4 (3–5) for the two raters. A statistically significant difference was observed in the mean mDISCERN scores between the two raters (P = 0.046). The GQS scores indicated a moderate to high quality (mean ± SD: 3.8 ± 0.8 for rater 1, 4.0 ± 0.5 for rater 2). No statistically significant correlation was found between mDISCERN and GQS scores (r = −0.006, P = 0.980). **Conclusions:** Although ChatGPT-4 has significant potential, it can be used as an additional source of information regarding TMD for patients and clinicians.

Key words: Artificial intelligence, ChatGPT, Temporomandibular joint disorders

Received June 12, 2024; Revised October 25, 2024; Accepted December 9, 2024.

Corresponding author: Betul Kula.
Assistant Professor, Department of Orthodontics, Istanbul Galata University, Evliya Celebi District, Mesrutiyet Avenue No: 62, Beyoglu, Istanbul 34440, Türkiye.
Tel +90-5543505571 **e-mail** betul.kula@galata.edu.tr

How to cite this article: Kula B, Kula A, Bagcier F, Alyanak B. Artificial intelligence solutions for temporomandibular joint disorders: Contributions and future potential of ChatGPT. Korean J Orthod 2025;55(2):131-141. https://doi.org/10.4041/kjod24.106

INTRODUCTION

Temporomandibular joint disorders (TMD) represent a significant public health concern, affecting an estimated 5–12% of the general population. They refer to a group of conditions characterized by pain and loss of function and are associated with headaches, restricted mouth opening, hypermobility syndromes, fibromyalgia, anxiety, and sleep disorders.[1-4] Given the impact of TMD on daily life, patients seek prompt solutions. In recent years, several innovations have emerged in various domains, including treatment modalities and patient management strategies for TMD. However, notable challenges remain for health professionals due to the complex nature of these diseases and their etiologies. Studies show that 50–80% of patients obtain information about their disease online before visiting a doctor.[5]

Artificial intelligence (AI) is a technology that can solve complex problems like human thinking and mimic human cognitive processes.[6] In recent years, the development of large language models (LLMs) has revolutionized the field of AI.[7,8] A multitude of chatbots such as Woebot, Your.MD, HealthTap, Cancer Chatbot, VitaminBot, Babylon Health, Safedrugbot, Microsoft Bing, and Google Bard are utilized for diverse purposes within the realm of health.[9-11] In dentistry, AI is used in scheduling appointments, clinical diagnosis and treatment planning, malocclusion detection in orthodontics, automatic classification of restorations in panoramic radiographs, periodontal diseases, root caries, and detection of maxillofacial abnormalities.[12]

The most discussed and used chatbot is Chat Generative Pre-Trained Transformer (ChatGPT), developed by the San Francisco-based company OpenAI and released in November 2022. The chatbot was trained on a large dataset and functions in a manner analogous to that of a human in dialogue with users. ChatGPT uses AI to respond to natural language queries, such as those of humans.[13-15] Its popularity is due to its detail, speed, and convenience. ChatGPT can provide numerous services to healthcare professionals in dentistry and healthcare.[16-18] ChatGPT has been subjected to three examinations by the United States Medical Licensing Examination, and its capabilities have been demonstrated.[19]

In addition, it is important to note that AI models, particularly LLMs, can produce varying results depending on the phrasing of the input and quality of the datasets on which they were trained. These models may occasionally demonstrate inconsistencies in their responses and, in certain instances, generate false or incomplete information, a phenomenon known as hallucinations. Hallucinations are more prevalent when a model lacks sufficient or consistent data on a given topic, particularly in cases involving obscure or controversial subjects.[20,21]

Several articles on AI and LLMs have recently been published in various fields.[8-19] These publications evaluated the competence and reliability of ChatGPT, focusing on the views of health professionals. Despite a comprehensive literature review, no studies have been conducted on TMD. This study evaluated the reliability and usefulness of the ChatGPT-4 responses to TMD keywords. This study assessed ChatGPT's efficacy in informing patients and professionals about TMD, a growing concern, and its reliability and utility.

We propose the following hypothesis: ChatGPT, with its extensive training on multiple topics, provides reliable and useful information for patients with TMD. This study is an effective supplementary resource for understanding TMD symptoms, management options, and self-care strategies, increasing patient knowledge and improving patient outcomes.

MATERIALS AND METHODS

This study was conducted following the principles of the Declaration of Helsinki. Ethical committee approval was not required because no human or animal data were used in this study.

Subheadings with various etiologies have been identified for TMD. The most recent guideline from the American Academy of Orofacial Pain and the International Headache Society, as modified by Okeson, was used as keywords.[22] These are "masticatory muscle disorders," "disc-condyle complex irregularities," "structural disorders of the articular surfaces," "inflammatory disorders of temporomandibular joint (TMJ)," "chronic mandibular hypermobility," "ankyloses," and "growth disorders of TMJ." A dialogue was initiated with the AI service (ChatGPT-4) utilizing the entry "TMD." A detailed discussion of the specific etiology of ChatGPT has been conducted. To obtain comprehensive information, we first asked ChatGPT about the disease. We then asked about causes, symptoms, and treatment. We created a new account to ensure that ChatGPT provides impartial answers. Each keyword was initiated as a new conversation and recorded for analysis. ChatGPT responses were obtained from the version released on October 12. The wording of the ChatGPT questions was standardized and consistent. Responses were examined for inconsistencies and hallucinations (Supplementary data).

The content of each response was evaluated for reliability using the ChatGPT reliability score.[14] The ChatGPT reliability score is a Likert-type scale with scores ranging from 1 to 7. Responses from medical and scientific sources were verified to ensure they were free of incomplete or error-prone information. High scores indicate high usefulness (Table 1).

The ChatGPT usefulness score was used to assess the

https://doi.org/10.4041/kjod24.106

Table 1. Reliability score scale, usefulness score scale, modified DISCERN scale, Global Quality Scale

Reliability score scale (Likert-type)
1. Completely unreliable: None of the information provided can be verified with medical sources, contains inaccurate or missing information.
2. Very unreliable: Much of the information cannot be verified with medical sources, is partially accurate but contains significantly inaccurate or missing information.
3. Less reliable: Most of the information provided can be verified with medical scientific sources but there is some important inaccurate or missing information.
4. Reliable: Most of the information provided can be verified with medical scientific sources, but there is some minor inaccurate or missing information.
5. Relatively reliable: Most of the information provided can be verified with medical scientific sources and there is very little inaccurate or missing information.
6. Very reliable: Most of the information provided can be verified with medical scientific sources and there is virtually no inaccurate or missing information.
7. Completely reliable: All the information provided can be verified with medical scientific sources and there is no inaccurate or missing information.
Usefulness score scale (Likert-type)
1. Not useful at all: Unintelligible language, contradictory information and missing important information. Not useful for patients.
2. Very little useful: Partially intelligible language. Some important information is missing or inaccurate. Limited use for patients.
3. Less useful: Intelligible language. The most important information is covered, but some important information is missing or inaccurate. Useful for patients.
4. Partially useful: Intelligible language. Some important information is missing or inaccurate, but the most important information is covered. Somewhat useful for patients.
5. Moderately useful: Intelligible language and the most important information is covered, but some important information is still missing or inaccurate. Useful for patients.
6. Very useful: Intelligible language. All important information is covered, but some unimportant information or details are also covered. Very useful for patients.
7. Extremely useful: Intelligible language and all-important information is covered. Additional information and resources are also provided, which are extremely useful for patients.
Modified DISCERN scale
1. Are the aims clear and achieved?
2. Are reliable sources of information used?
3. Is the information presented balanced and unbiased?
4. Are additional sources of information listed for patient reference?
5. Are areas of uncertainty mentioned?
Modified Global Quality Scale*
1. The information exhibits poor quality, lacks a smooth flow, and is missing significant information. It does not offer any value to patients.
2. The information is generally of low quality with an unstructured flow. While some information is correctly presented, it omits many important topics, making it of very limited use for patients.
3. The information demonstrates moderate quality with suboptimal flow. It adequately covers some crucial information but poorly addresses others, making it useful for patients.
4. The information is of good quality and generally flows well. Most relevant information is included, but some topics are still missing. It is considered useful for patients.
5. The information showcases excellent quality and flows seamlessly, making it highly beneficial for patients. It covers a comprehensive range of essential information and is very valuable.

*Data from the article of Bernard et al. (Am J Gastroenterol 2007;102:2070-7).[23]

content of each response to determine its utility in patients.[17] The scale is a Likert-type scale with values ranging from 1 to 7. A high score indicated high usefulness (Table 1).

The DISCERN scale was designed to evaluate the reliability of the ChatGPT responses. The modified DISCERN (mDISCERN) scale comprised five questions with yes/no answers (Table 1). According to the mDISCERN criteria, a score of 1 was negative and 5 was positive. Scores of 2, 3, and 4 indicated partial applicability. A score of "partially" is preferred for limited applicability. The total scores range from 0 to 5, with higher scores indicating greater reliability.

The Global Quality Scale (GQS) was modified with a specific emphasis on assessing the accuracy and utility of the information initially presented by Bernard. The responses generated by ChatGPT were evaluated using a GQS (Table 1).[23]

Two specialists of independent fields—an orthodontist and a physical medicine specialist—were involved in evaluating the screenshots to avoid bias. If the raters disagreed, a third independent rater assessed the data and made the final decision.

Statistical analysis

Analyses were performed using MedCalc® Statistical Software version 19.7.2 (MedCalc Software Ltd., Ostend, Belgium; https://www.medcalc.org). Descriptive statistics were used to describe the continuous variables (mean, standard deviation [SD], minimum, median, and maximum). Quantitative data were summarized as the mean, SD, and median (minimum to maximum). The Wilcoxon signed-rank test determined relationships between two dependent variables that did not conform to a normal distribution. The inter-rater agreement was analyzed using Cronbach's α. The intraclass correlation coefficient results indicate that positive values ranging from 0 to 0.2 represent poor agreement, 0.2 to 0.4 represent fair agreement, 0.4 to 0.6 represent moderate agreement, 0.6 to 0.8 represent good agreement, and 0.8 to 1.0 represent very good agreement. The level of statistical significance was set at $P < 0.05$.

RESULTS

Two raters with extensive experience and independence evaluated the responses. The results are presented in the following tables and figure.

The responses are shown in Table 2, as indicated by a Likert-type reliability scale. The median Likert reliability scores were 6 (4–6) and 5 (4–6), respectively, indicating that the responses were highly reliable and reproducible for raters 1 and 2. The median reliability scores showed no statistically significant differences between the groups ($P > 0.05$; Table 3). The Inter-rater reliability of the scoring process was evaluated for the entire content. A high level of agreement was observed between the two raters (Cronbach α: 0.829).

Table 3 presents the distribution of the responses to the Likert-type usefulness scale. The median Likert usefulness score was 5 (4–6) for two raters, indicating that the responses were moderately helpful. The agreement of usefulness scoring between the two raters was examined for all content. Moderate agreement was observed between the two raters (Cronbach α: 0.693). The median usefulness scores showed no statistically significant differences between the groups ($P > 0.05$; Table 3).

The highest Likert scores for the reliability and usefulness of the information provided by the ChatGPT were for masticatory muscle disorders (mean: 6.0 ± 0). The lowest scores in terms of both reliability and usefulness were for inflammatory disorders of the TMJ (mean: 4.3 ± 0.6 for reliability, mean: 4.0 ± 0 for usefulness). The scores the two raters gave ranged between four and six ($P = 1.000–0.317$, respectively).

A comparative analysis was performed on all topics' total reliability and usefulness scores. No statistically significant difference was found in any subject for the reliability or usefulness total scores ($P = 0.083$-1.000; Table 3).

Table 2 presents the distribution of ChatGPT response scores according to the mDISCERN scale. No statistically significant difference was identified between the median mDISCERN values of the groups ($P > 0.05$). Regarding reliability, as determined using the mDISCERN tool, Rater 1 identified very high reliability; however, Rater 2 identified high reliability (mean ± SD: 4.1 ± 0.7, 3.8 ± 0.6, respectively). The median mDISCERN score of the responses was 4 (3–5) for two raters. A statistically significant difference was observed in the mean mDISCERN scores between raters 1 and 2 ($P = 0.046$; Table 3). There was no statistically significant correlation between DISCERN and GQS scores ($r = -0.006$, $P = 0.980$).

The mean GQS scores for rater 1 and rater 2 were 3.8 ± 0.8 and 4.0 ± 0.5, respectively. The raters assessed the responses as of moderate to high quality. As demonstrated in Table 3, the GQS scores indicate a low level of agreement between the two raters (Cronbach α: 0.452). No statistically significant differences were observed between the groups ($P > 0.05$).

The reliability and usefulness indices were calculated based on the mean scores assigned by the raters. The resulting distribution is shown in Figure 1. The highest mean value for reliability was observed in masticatory muscle disorders and ankylosis (mean ± SD: 6.0 ± 0). In contrast, the lowest mean value was noted in chronic mandibular hypermobility (mean ± SD: 4.0 ± 0), as determined by the separate scoring of two independent

https://doi.org/10.4041/kjod24.106

Table 2. Inter-rater reliability, usefulness, mDISCERN and GQS

	Reliability (Likert-type)		Usefulness (Likert-type)		mDISCERN		GQS	
	Rater 1	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2
Masticatory muscle disorders								
Causes	6	6	6	5	4	5	3	4
Symptoms	6	6	6	6	4	4	4	5
Treatment	6	6	6	6	5	4	2	4
Disc-condyle complex derangement								
Causes	5	4	6	4	3	4	3	4
Symptoms	6	5	6	5	5	4	4	4
Treatment	6	6	6	6	4	4	3	4
Structural disorders of the articular surfaces								
Causes	5	5	6	5	3	4	5	5
Symptoms	6	5	6	5	5	4	3	5
Treatment	5	4	5	4	4	3	4	4
Inflammatory disorders of TMJ								
Causes	4	5	4	5	3	4	3	5
Symptoms	5	4	4	4	4	3	5	5
Treatment	4	4	4	4	4	4	4	5
Chronic mandibular hypermobility								
Causes	5	4	4	4	5	4	4	5
Symptoms	5	4	5	4	4	3	5	5
Treatment	4	4	5	4	3	3	4	4
Ankylosis								
Causes	6	5	6	5	5	4	4	5
Symptoms	6	6	5	4	5	5	5	5
Treatment	6	5	5	4	4	3	4	5
Growth disorders of TMJ								
Causes	6	5	5	5	4	3	3	4
Symptoms	5	4	5	4	4	3	4	5
Treatment	6	6	4	5	5	4	3	5
Mean ± SD	5.4 ± 0.7	4.9 ± 0.8	5.2 ± 0.8	4.7 ± 0.7	4.1 ± 0.7	3.8 ± 0.6	3.8 ± 0.8	4.0 ± 0.5
Median (min–max)	6 (4–6)	5 (4–6)	5 (4–6)	5 (4–6)	4 (3–5)	4 (3–5)	4 (2–5)	4 (3–5)

Wilcoxon signed rank test.
mDISCERN, modified DISCERN; GQS, Global Quality Scale; TMJ, temporomandibular joint; SD, standard deviation.

raters. For usefulness, the highest mean value was seen in masticatory muscle disorders and disc-condyle complex derangement (mean ± SD: 6.0 ± 0), whereas the lowest mean value was seen in chronic mandibular hypermobility (mean ± SD: 4.0 ± 0). The raters had no significant differences in the reliability or usefulness total scores.

DISCUSSION

Recently, AI has brought about innovative tools such as ChatGPT, which are designed to engage in conversations and provide information on various topics. ChatGPT can be used in various areas of dentistry, such as symptom assessment and temporary treatment suggestions, appointment scheduling and reminders, the planning and assessment of treatments, the analysis and di-

Table 3. Inter-rater differences in Likert-type reliability, Likert-type usefulness, mDISCERN and GQS score of the subject

	Masticatory muscle disorders	Disc-condyle complex derangement	Structural disorders of the articular surfaces	Inflammatory disorders of TMJ	Chronic mandibular hypermobility	Ankylosis	Growth disorders of TMJ
Reliability (Likert-type)							
Rater 1							
Mean ± SD	6.0 ± 0	5.7 ± 0.6	5.3 ± 0.6	4.3 ± 0.6	4.7 ± 0.6	6.0 ± 0	5.7 ± 0.6
Median (min–max)	6 (6–6)	6 (5–6)	5 (5–6)	4 (4–5)	5 (4–5)	6 (6–6)	6 (5–6)
Rater 2							
Mean ± SD	6.0 ± 0	5.0 ± 1.0	4.7 ± 0.6	4.3 ± 0.6	4.0 ± 0	5.3 ± 0.6	5.0 ± 1.0
Median (min–max)	6 (6–6)	5 (4–6)	5 (4–5)	4 (4–5)	4 (4–4)	5 (5–6)	5 (4–6)
Usefulness (Likert-type)							
Rater 1							
Mean ± SD	6.0 ± 0	6.0 ± 0	5.7 ± 0.6	4.0 ± 0	4.7 ± 0.6	5.3 ± 0.6	4.7 ± 0.6
Median (min–max)	6 (6–6)	6 (6–6)	6 (5–6)	4 (4–4)	5 (4–5)	5 (5–6)	5 (4–5)
Rater 2							
Mean ± SD	5.7 ± 0.6	5.0 ± 1.0	4.7 ± 0.6	4.3 ± 0.6	4.0 ± 0	4.3 ± 0.6	4.7 ± 0.6
Median (min–max)	6 (5–6)	5 (4–6)	5 (4–5)	4 (4–5)	4 (4–4)	4 (4–5)	5 (4–5)
mDISCERN							
Rater 1							
Mean ± SD	4.3 ± 0.6	4.0 ± 1.0	4.0 ± 1.0	3.7 ± 0.6	4.0 ± 1.0	4.7 ± 0.6	4.3 ± 0.6
Median (min–max)	4 (4–5)	4 (3–5)	4 (3–5)	4 (3–4)	4 (3–5)	5 (4–5)	4 (4–5)
Rater 2							
Mean ± SD	4.3 ± 0.6	4.0 ± 0	3.7 ± 0.6	3.7 ± 0.6	3.3 ± 0.6	4.0 ± 1.0	3.3 ± 0.6
Median (min–max)	4 (4–5)	4 (4–4)	4 (3–4)	4 (3–4)	3 (3–4)	4 (3–5)	3 (3–4)
GQS							
Rater 1							
Mean ± SD	3.0 ± 1.0	3.3 ± 0.6	3.3 ± 0.6	4.7 ± 0.6	4.3 ± 0.6	4.3 ± 0.6	3.3 ± 0.6
Median (min–max)	3 (2–4)	3 (3–4)	3 (3–4)	5 (4–5)	4 (4–5)	4 (4–5)	3 (3–4)
Rater 2							
Mean ± SD	4.3 ± 0.6	4.0 ± 0	4.3 ± 0.6	3.3 ± 0.6	4.0 ± 0	3.7 ± 0.6	4.0 ± 0
Median (min–max)	4 (4–5)	4 (4–4)	4 (4–5)	3 (3–4)	4 (4–4)	4 (3–4)	4 (4–4)
Reliability*	1.000	0.157	0.157	1.000	0.157	0.157	0.157
Usefulness*	0.317	0.180	0.083	0.317	0.157	0.083	1.000
mDISCERN*	1.000	1.000	0.564	1.000	0.157	0.157	0.083
GQS*	0.102	0.157	0.180	0.157	0.317	0.317	0.157

*Wilcoxon signed rank test.
mDISCERN, modified DISCERN; TMJ, temporomandibular joint; SD, standard deviation; GQS, Global Quality Scale.

agnostics of images, the monitoring of patients and the subsequent follow-up, and the conducting of literature reviews.[14,24]

The effects of AI on dentistry span various clinical and administrative domains. Clinically, AI is primarily used for diagnostic purposes, enhancing the accuracy and efficiency of identifying oral conditions, such as dental caries and periodontal diseases.[25] Moreover, AI-driven imaging technologies have advanced the detection capabilities of dental radiographs, enabling the early and accurate diagnosis of complex oral issues. AI contributes to practice management by streamlining processes, such as appointment scheduling, patient communication, and claims processing, allowing dental professionals to

https://doi.org/10.4041/kjod24.106

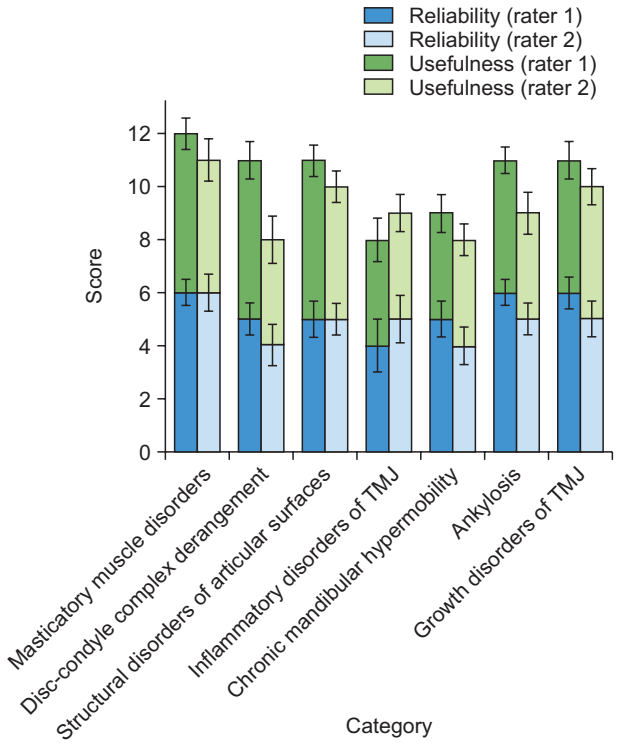

Figure 1. Assessment of average reliability and usefulness scores of each disease (interquartile range error bars). TMJ, temporomandibular joint.

devote more time to direct patient care.[26,27] Innovations like AI-powered live coaching systems and computer vision technologies for pediatric and orthodontic care are anticipated to improve patient outcomes and practice efficiency.[28]

Despite these benefits, the integration of AI into dentistry is challenging. Ethical concerns, data privacy issues, and potential errors are the prominent controversies surrounding the adoption of AI in clinical practice.[29]

In addition, the quality of training data heavily influences the efficacy of AI tools, necessitating continuous updates. These challenges highlight the importance of adopting a cautious approach emphasizing the transparency and reliability of AI applications. As AI technology matures, its potential to transform dental-care delivery and patient engagement will continue to grow, fostering a new era of precision and enhanced patient-centered care.[30] Health-related misinformation can lead to misdiagnosis, treatment errors, and serious consequences.[31]

Given the above scientific data, we aimed to assess whether ChatGPT displays a high level of knowledge on specific topics. To this end, we examined the reliability and usefulness of the information provided by ChatGPT concerning the various subtopics related to TMD affecting the daily lives of patients.

This study had several strengths. The authors employed a combination of objective and validated tools to assess reliability. These included Likert scales, GQS, and mDISCERN criteria, mainly compared with previously published descriptive studies.[32,33] The Likert scale is an effective tool for evaluating the reliability and usefulness of generated responses, as it allows for a clear distinction between the levels of accuracy. This scale enables discernible differentiation among the levels of accuracy and is commonly used in the existing literature for comparable assessments.[34-36]

Discern is a validated instrument recognized as an effective method for assessing the reliability and quality of written health information. It is intended for health professionals and laypeople (Institute of Health Sciences, University of Oxford, Oxford, UK).[37] In the present study, as suggested by previous studies, the mDISCERN scale was used to assess the reliability of the information to focus on specific criteria through a detailed analysis of TMD.[38]

TMJ problems require multidisciplinary treatment involving the cooperation of many specialists. Therefore, our study evaluated the responses separately by an orthodontist and a physical therapy practitioner. This can provide valuable insights into the ChatGPT platform's strengths and limitations in addressing the diverse aspects of TMJ problems and help ensure a more reliable and holistic evaluation.

Our results indicate that masticatory muscle diseases exhibited the highest reliability and usefulness scores. The responses to masticatory muscle disorder treatment were evaluated, with a mean score of 6.0 ± 0. After reviewing the responses, a wide range of information was provided regarding treatment options for masticatory muscle diseases in the literature. These options vary from conservative treatments such as occlusal splints and medicines to more interventional approaches such as Botox injections and surgical procedures.[39] ChatGPT's treatment options were comprehensive and accurate in the context of masticatory muscle disorders. However, other studies require additional details. Accurate information on conservative treatments should be highlighted. However, superficial information on advanced surgical options is a potential limitation. The model's information was incomplete and lacked guidance in complex areas like TMJ. This is a limitation of ChatGPT. The lowest mean usefulness and reliability scores were associated with inflammatory TMJ diseases. The reasons for this lower score could vary, including the complexity of the subject matter, quality of the available information, and presentation of the content.

However, ChatGPT can provide helpful information, and it is vital to highlight the significance of seeking guidance from qualified healthcare experts and reliable

medical sources when making decisions regarding TMD. It is necessary to consider ChatGPT as an additional source of information. However, it does not replace the expertise of actual doctors. Previous research noted that ChatGPT frequently recommended contacting an orthodontist, dentist, or doctor at the end of responses. This result is consistent with our findings.[31]

Although the raters agreed on the usefulness and reality scores, they disagreed on the GQS or mDISCERN scores. Therefore, a third evaluator was included, and the scores were incompatible. The mDISCERN grading of the responses revealed that the level of agreement between the two evaluators is not particularly strong, and the observed agreement may be attributed to chance (Cronbach α: 0.452). In particular, there were differences in the scores related to patient treatment. This may be because the raters treated patients using different clinical practices.

Balel[40] evaluated the answers generated by ChatGPT regarding maxillofacial surgery by physicians using the GQS scale; the results were of moderate quality. In their current study, Kılınç and Mansız[41] used the Flesch-Kincaid and DISCERN tools to evaluate the reliability and readability of the information about orthodontics generated by ChatGPT-4. The mean DISCERN value was reported to be 2.96 ± 0.05 for general questions, 3.04 ± 0.06, 2.38 ± 0.27 and 2.82 ± 0.31 for treatment-related questions. The mDISCERN tool indicated high reliability for both raters (mean ± SD: 4.1 ± 0.7 to 3.8 ± 0.6, respectively). It is thought that in our study, the discrepancy in evaluations may be attributed to the different specialties of the evaluators.

Dursun and Bilici Geçer[35] used the Likert scale, mDISCERN, GQS, and Flesch Reading Ease Score to assess ChatGPT-3.5, ChatGPT-4, Gemini, and Copilot AI. Responses were moderately reliable and of good quality. ChatGPT-4's mean GQS score was 3.8 ± 0.62, indicating high quality. These results confirmed our findings. In our study, ChatGPT-4's responses for TMD were moderate to high quality. The highest mean Likert score for the ChatGPT-4 responses was 4.5 ± 0.61. We found the mean Likert scores 5.4 ± 0.7 and 4.9 ± 0.8 for reliability and usefulness.

Tanaka et al.[42] evaluated the reliability of ChatGPT-4 in answering questions on clear aligners, temporary anchorage devices, and digital imaging in orthodontics using a Likert scale. The study found a slight agreement and discrepancy in ratings between assessors, with a combined Fleiss' κ value of 0.004 ($P < 0.001$). Our study demonstrated a high level of agreement regarding inter-rater reliability (Cronbach α: 0.829).

Khanagar et al.[43] noted that in the context of AI applications for diagnosis and treatment planning in orthodontics, AI achieved accuracy and precision comparable to that of trained examiners. Abu Arqub et al.[44] reported that ChatGPT generated responses to questions about clear aligner therapy in orthodontics with less-than-optimal accuracy. However, none of these studies referred to TMD.

Response variability and hallucination risk in AI models

LLM's accuracy depends on the questions and training data despite their large datasets. Some inconsistencies are due to the model's answers to the questions. The formulation of questions can result in discrepancies in model responses. Hallucinations, a prevalent issue in language models, must be considered. The generation of false or fabricated information by the model presents a significant risk, particularly in clinical applications.[45]

ChatGPT can generate coherent responses but may fabricate details or provide outdated information. A systematic review found that 96.7% of studies were concerned about the accuracy of ChatGPT. These concerns highlight the risks of ChatGPT, including misinformation, lack of originality, and hallucinations.[46] ChatGPT's effectiveness depends on the training data. It struggles with rare or novel medical conditions but shows potential in certain contexts, including clinical documents and healthcare efficiency. Additionally, studies have reported remarkable accuracy in specific medical examinations, suggesting that it could be useful in medical education and decision-making.[47]

In this study, some of the responses from the model were either incomplete or incorrect. ChatGPT responses to complex topics, such as inflammatory disorders of the TMJ, were incomplete or lacked detail. This demonstrates the need to rigorously evaluate language models when providing health information. Similarly, the aforementioned studies highlighted the phenomenon of hallucinations in ChatGPT responses.

Our study has several limitations. The ChatGPT 4.0 model can provide information on many subjects. ChatGPT 4.0 is based on an updated knowledge base until April 2023. However, they may not be aware of the recent developments. However, ChatGPT cannot verify or update the accuracy of the information provided. ChatGPT does not assess individual health or private medical histories. ChatGPT provides general information rather than personalized medical advice. To evaluate individual health conditions accurately, it is essential to consult qualified healthcare professionals. The absence of a published study on this topic prevented us from discussing it in detail.[48]

A previous study on TMD on YouTube found that information was inadequate. It is important to evaluate sources, as they may not provide accurate information, particularly for complex medical conditions.[49] ChatGPT's main advantage is its conversational interactivity.[50] Stud-

https://doi.org/10.4041/kjod24.106

ies are needed to compare the information provided by ChatGPT with YouTube and other patient information platforms. This may highlight areas of improvement and help patients make informed decisions.

CONCLUSIONS

Although AI has immense potential in healthcare, it should always be used with human expertise. Future research should focus on analyzing various LLMs. The proliferation of misinformation has become a major concern in the digital age. Academic institutions are important in ensuring that patients with TMD are directed toward the correct diagnostic and therapeutic processes. AI models have limitations, including variability in responses and the risk of hallucinations. However, the information provided by these models provide must be verified by experts in clinical applications.

AUTHOR CONTRIBUTIONS

Conceptualization: All authors. Data curation: All authors. Formal analysis: All authors. Investigation: All authors. Methodology: All authors. Project administration: All authors. Resources: All authors. Software: All authors. Supervision: All authors. Validation: All authors. Visualization: All authors. Writing–original draft: All authors. Writing–review & editing: All authors.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

FUNDING

None to declare.

SUPPLEMENTARY MATERIAL

Supplementary data is available at https://doi.org/10.4041/kjod24.106

REFERENCES

1. Andre A, Kang J, Dym H. Pharmacologic treatment for temporomandibular and temporomandibular joint disorders. Oral Maxillofac Surg Clin North Am 2022;34:49-59. https://doi.org/10.1016/j.coms.2021.08.001
2. Shaffer SM, Brismée JM, Sizer PS, Courtney CA. Temporomandibular disorders. Part 1: anatomy and examination/diagnosis. J Man Manip Ther 2014;22:2-12. https://doi.org/10.1179/2042618613y.0000000060
3. Thomas DC, Khan J, Manfredini D, Ailani J. Temporomandibular joint disorder comorbidities. Dent Clin North Am 2023;67:379-92. https://doi.org/10.1016/j.cden.2022.10.005
4. Valesan LF, Da-Cas CD, Réus JC, Denardin ACS, Garanhani RR, Bonotto D, et al. Prevalence of temporomandibular joint disorders: a systematic review and meta-analysis. Clin Oral Investig 2021;25:441-53. https://doi.org/10.1007/s00784-020-03710-w
5. AlGhamdi KM, Moussa NA. Internet use by the public to search for health-related information. Int J Med Inform 2012;81:363-73. https://doi.org/10.1016/j.ijmedinf.2011.12.004
6. Mintz Y, Brodie R. Introduction to artificial intelligence in medicine. Minim Invasive Ther Allied Technol 2019;28:73-81. https://doi.org/10.1080/13645706.2019.1575882
7. Kung JE, Marshall C, Gauthier C, Gonzalez TA, Jackson JB 3rd. Evaluating ChatGPT performance on the orthopaedic in-training examination. JB JS Open Access 2023;8:e23.00056. https://doi.org/10.2106/jbjs.oa.23.00056
8. Sharma S, Pajai S, Prasad R, Wanjari MB, Munjewar PK, Sharma R, et al. A critical review of ChatGPT as a potential substitute for diabetes educators. Cureus 2023;15:e38380. https://doi.org/10.7759/cureus.38380
9. Palanica A, Flaschner P, Thommandram A, Li M, Fossat Y. Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. J Med Internet Res 2019;21:e12887. https://doi.org/10.2196/12887
10. Acar AH. Can natural language processing serve as a consultant in oral surgery? J Stomatol Oral Maxillofac Surg 2024;125:101724. https://doi.org/10.1016/j.jormas.2023.101724
11. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019;6:94-8. https://doi.org/10.7861/futurehosp.6-2-94
12. Ahmed N, Abbasi MS, Zuberi F, Qamar W, Halim MSB, Maqsood A, et al. Artificial intelligence techniques: analysis, application, and outcome in dentistry-a systematic review. Biomed Res Int 2021;2021:9751564. https://doi.org/10.1155/2021/9751564
13. Alhaidry HM, Fatani B, Alrayes JO, Almana AM, Alfhaed NK. ChatGPT in dentistry: a comprehensive review. Cureus 2023;15:e38317. https://doi.org/10.7759/cureus.38317
14. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor

Dent 2023;35:1098-102. https://doi.org/10.1111/jerd.13046
15. Strunga M, Urban R, Surovková J, Thurzo A. Artificial intelligence systems assisting in the assessment of the course and retention of orthodontic treatment. Healthcare (Basel) 2023;11:683. https://doi.org/10.3390/healthcare11050683
16. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. J Dent Res 2020;99:769-74. https://doi.org/10.1177/0022034520915714
17. Cankurtaran RE, Polat YH, Aydemir NG, Umay E, Yurekli OT. Reliability and usefulness of ChatGPT for inflammatory bowel diseases: an analysis for patients and healthcare professionals. Cureus 2023;15:e46736. https://doi.org/10.7759/cureus.46736
18. Uz C, Umay E. "Dr ChatGPT": is it a reliable and useful source for common rheumatic diseases? Int J Rheum Dis 2023;26:1343-9. https://doi.org/10.1111/1756-185x.14749
19. Hasnain M, Hayat A, Hussain A. Revolutionizing chronic obstructive pulmonary disease care with the open AI application: ChatGPT. Ann Biomed Eng 2023;51:2100-2. https://doi.org/10.1007/s10439-023-03238-6
20. Chelli M, Descamps J, Lavoué V, Trojani C, Azar M, Deckert M, et al. Hallucination rates and reference accuracy of ChatGPT and bard for systematic reviews: comparative analysis. J Med Internet Res 2024;26:e53164. https://doi.org/10.2196/53164
21. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023;15:e35179. https://doi.org/10.7759/cureus.35179
22. Yaltırık M, Palancıoğlu A, Koray M, Turgut CT. Temporomandibular joint disorders and diagnosis. Yeditepe J Dent 2017;13:43-50. https://doi.org/10.5505/yeditepe.2017.07078
23. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. Am J Gastroenterol 2007;102:2070-7. https://doi.org/10.1111/j.1572-0241.2007.01325.x
24. Agrawal P, Nikhade P. Artificial intelligence in dentistry: past, present, and future. Cureus 2022;14:e27405. https://doi.org/10.7759/cureus.27405
25. Anil S, Sudeep K, Saratchandran S, Sweety VK. Revolutionizing dental caries diagnosis through artificial intelligence. In: Chibinski ACR, ed. Dental caries perspectives - a collection of thoughtful essays. London: IntechOpen; 2023. https://doi.org/10.5772/intechopen.112979
26. Musleh D, Almossaeed H, Balhareth F, Alqahtani G, Alobaidan N, Altalag J, et al. Advancing dental diagnostics: a review of artificial intelligence applications and challenges in dentistry. Big Data Cogn Comput 2024;8:66. https://doi.org/10.3390/bdcc8060066
27. Ghaffari M, Zhu Y, Shrestha A. A review of advancements of artificial intelligence in dentistry. Dent Rev 2024;4:100081. https://doi.org/10.1016/j.dentre.2024.100081
28. Balaban C, Inam W, Kennedy R, Faiella R. The future of dentistry: how AI is transforming dental practices. Compend Contin Educ Dent 2021;42:14-7. https://pubmed.ncbi.nlm.nih.gov/33481621/
29. Xie B, Xu D, Zou XQ, Lu MJ, Peng XL, Wen XJ. Artificial intelligence in dentistry: a bibliometric analysis from 2000 to 2023. J Dent Sci 2024;19:1722-33. https://doi.org/10.1016/j.jds.2023.10.025
30. Kukalakunta Y, Thunki P, Yellu RR. Integrating artificial intelligence in dental healthcare: opportunities and challenges. J Deep Learn Genom Data Anal 2024;4:34-41. https://aithor.com/paper-summary/integrating-artificial-intelligence-in-dental-healthcare-opportunities-and-challenges
31. Kessels RP. Patients' memory for medical information. J R Soc Med 2003;96:219-22. https://doi.org/10.1177/014107680309600504
32. Vinufrancis A, Al Hussein H, Patel HV, Nizami A, Singh A, Nunez B, et al. Assessing the quality and reliability of AI-generated responses to common hypertension queries. Cureus 2024;16:e66041. https://doi.org/10.7759/cureus.66041
33. Onder CE, Koc G, Gokbulut P, Taskaldiran I, Kuskonmaz SM. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. Sci Rep 2024;14:243. https://doi.org/10.1038/s41598-023-50884-w
34. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Seifman MA. Investigating the impact of innovative AI chatbot on post-pandemic medical education and clinical assistance: a comprehensive analysis. ANZ J Surg 2024;94:68-77. https://doi.org/10.1111/ans.18666
35. Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? BMC Med Inform Decis Mak 2024;24:211. https://doi.org/10.1186/s12911-024-02619-8
36. Hatia A, Doldo T, Parrini S, Chisci E, Cipriani L, Montagna L, et al. Accuracy and completeness of ChatGPT-generated information on interceptive orthodontics: a multicenter collaborative study. J Clin Med 2024;13:735. https://doi.org/10.3390/jcm13030735
37. Alan R, Alan BM. Utilizing ChatGPT-4 for providing

information on periodontal disease to patients: a DISCERN quality analysis. Cureus 2023;15:e46213. https://doi.org/10.7759/cureus.46213
38. Zengin O, Onder ME. Educational quality of YouTube videos on musculoskeletal ultrasound. Clin Rheumatol 2021;40:4243-51. https://doi.org/10.1007/s10067-021-05793-6
39. Wadhwa S, Kapila S. TMJ disorders: future innovations in diagnostics and therapeutics. J Dent Educ 2008;72:930-47. https://pubmed.ncbi.nlm.nih.gov/18676802/
40. Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? J Stomatol Oral Maxillofac Surg 2023;124:101471. https://doi.org/10.1016/j.jormas.2023.101471
41. Kılınç DD, Mansız D. Examination of the reliability and readability of Chatbot Generative Pretrained Transformer's (ChatGPT) responses to questions about orthodontics and the evolution of these responses in an updated version. Am J Orthod Dentofacial Orthop 2024;165:546-55. https://doi.org/10.1016/j.ajodo.2023.11.012
42. Tanaka OM, Gasparello GG, Hartmann GC, Casagrande FA, Pithon MM. Assessing the reliability of ChatGPT: a content analysis of self-generated and self-answered questions on clear aligners, TADs and digital imaging. Dental Press J Orthod 2023;28:e2323183. https://doi.org/10.1590/2177-6709.28.5.e2323183.oar
43. Khanagar SB, Al-Ehaideb A, Vishwanathaiah S, Maganur PC, Patil S, Naik S, et al. Scope and performance of artificial intelligence technology in orthodontic diagnosis, treatment planning, and clinical decision-making: a systematic review. J Dent Sci 2021;16:482-92. https://doi.org/10.1016/j.jds.2020.06.018
44. Abu Arqub S, Al-Moghrabi D, Allareddy V, Upadhyay M, Vaid N, Yadav S. Content analysis of AI-generated (ChatGPT) responses concerning orthodontic clear aligners. Angle Orthod 2024;94:263-72. https://doi.org/10.2319/062623-472.1
45. Siontis KC, Attia ZI. ChatGPT hallucinating: can it get any more humanlike? Eur Heart J 2024;45:321-3. https://doi.org/10.1093/eurheartj/ehad548
46. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023;11:887. https://doi.org/10.3390/healthcare11060887
47. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2:e0000198. https://doi.org/10.1371/journal.pdig.0000198
48. Fatima A, Shafi I, Afzal H, Díez IT, Lourdes DRM, Breñosa J, et al. Advancements in dentistry with artificial intelligence: current clinical applications and future perspectives. Healthcare (Basel) 2022;10:2188. https://doi.org/10.3390/healthcare10112188
49. Vaira LA, Sergnese S, Salzano G, Maglitto F, Arena A, Carraturo E, et al. Are YouTube videos a useful and reliable source of information for patients with temporomandibular joint disorders? J Clin Med 2023;12:817. https://doi.org/10.3390/jcm12030817
50. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312. https://doi.org/10.2196/45312