International Journal of Otorhinolaryngology and Head and Neck Surgery
Poojary SC et al. Int J Otorhinolaryngol Head Neck Surg. 2026 Aug;12(4):608-612
https://www.ijorl.com
pISSN 2454-5929 | eISSN 2454-5937

DOI: https://dx.doi.org/10.18203/issn.2454-5929.ijohns20262383

Original Research Article

Factual accuracy versus the health literacy barrier: a blinded neurotological assessment of ChatGPT in vertigo education

Shwetha C. Poojary[1]*, Prasanna Nayak[2]

[1]Department of Otorhinolaryngology, Karwar Institute of Medical Sciences, Uttara Kannada, Karnataka, India
[2]Department of Paediatrics, Father Muller Medical College, Mangalore, Karnataka, India

Received: 06 July 2026
Accepted: 21 July 2026

***Correspondence:**
Dr. Shwetha C. Poojary,
E-mail: shwethapadukone@gmail.com

ABSTRACT

Background: Patients increasingly utilize artificial intelligence (AI) models like ChatGPT to seek medical information regarding complex symptoms such as vertigo. The reliability of these AI platforms in providing accurate, comprehensible, and safe clinical information remains a subject of ongoing investigation. Objectives were to evaluate the clinical accuracy, comprehensiveness, and readability of ChatGPT's responses to common patient inquiries regarding vertigo and vestibular disorders.
Methods: Fifty frequently asked questions (FAQs) concerning vertigo were curated from high-traffic patient forums, clinical encounters, and search engine trends. These questions were categorized into four domains: Diagnosis, treatment, home exercises/lifestyle, and prognosis. The FAQs were inputted into ChatGPT-4. Responses were independently evaluated using a 5-point Likert scale for accuracy and comprehensiveness. Readability was assessed using the Flesch-Kincaid grade level (FKGL) and Flesch reading ease (FRE) scores.
Results: ChatGPT demonstrated high overall accuracy (mean score 4.38±0.62) and comprehensiveness (4.21±0.74). The model performed best in the home exercises/lifestyle category (accuracy 4.65±0.45) and scored lowest in the diagnosis domain (accuracy 4.05±0.82), where it occasionally failed to distinguish atypical features of central vs. peripheral vertigo. The mean FKGL was 12.4±1.8, and the mean FRE was 41.2±10.5, indicating a reading level appropriate for college students but potentially difficult for the general public.
Conclusions: ChatGPT-4 generates highly accurate and comprehensive answers to common patient questions about vertigo, making it a valuable supplementary educational tool. However, its sophisticated vocabulary and occasional diagnostic ambiguity emphasize the necessity for physician oversight.

Keywords: ChatGPT, Artificial intelligence, Vertigo, Patient education, Otolaryngology, Vestibular disorders

INTRODUCTION

Vertigo and dizziness are among the most common presenting complaints in clinics, affecting approximately 15% to 20% of adults yearly. The etiology of vertigo spans a broad spectrum-from peripheral vestibular disorders like benign paroxysmal positional vertigo (BPPV) and Ménière's disease to life-threatening central pathologies. Because vestibular symptoms are complex and often highly distressing, patients frequently experience anxiety regarding their underlying diagnosis. Consequently, there is a growing trend of patients consulting digital health platforms and the internet for self-triage and symptom management prior to formal medical appointments.

The introduction of generative pre-trained transformer models, particularly OpenAI's ChatGPT, has revolutionized how the public accesses health information.[1] Built upon large language models (LLMs), ChatGPT can simulate human-like conversations and synthesize vast amounts of medical literature. Recent

studies have demonstrated ChatGPT's impressive capabilities in passing standardized medical licensing examinations and answering specialty-specific clinical questions across various disciplines.[2,3]

Within otolaryngology, ChatGPT has shown considerable promise in synthesizing treatment alternatives for common procedures and offering perioperative guidance.[4]

However, a rigorous evaluation of its adherence to established otolaryngology clinical practice guidelines has highlighted variable performance. This variability underscores the critical risk of AI "hallucinations"-instances where the model generates highly plausible but factually incorrect or unverified medical statements, which can mislead patients.[5]

Vertigo presents a particularly unique diagnostic and educational challenge for AI. Achieving high diagnostic precision in vestibular disorders relies heavily on eliciting nuanced historical details, such as precise symptom duration, specific positional triggers, and the presence or absence of associated audiological and neurological signs.

A recent analysis indicated that while ChatGPT can generally distinguish between vestibular and non-vestibular etiologies, its diagnostic accuracy for specific, complex clinical cases remains inferior to the evidence-based reasoning of experienced clinicians.[6]

Furthermore, evaluating the utility of patient-facing digital health tools involves assessing not only accuracy but also accessibility. Studies comparing AI models against traditional search results suggest that while AI-generated information is often more comprehensive, it is frequently rendered at reading levels too difficult for the average patient to comprehend safely.[7]

Assessing AI-generated text in otolaryngology publications remains an active area of study to ensure patient safety and comprehension.[8]

Therefore, as patients increasingly integrate these advanced LLMs into their healthcare journey, clinicians must understand the exact strengths and limitations of conversational AI. The primary objective of this study was to critically evaluate the clinical accuracy, comprehensiveness, and readability of ChatGPT-4 when addressing a diverse array of common patient questions about vertigo.

METHODS

Study design and ethical approval

This cross-sectional, observational study evaluated the quality of AI-generated medical information.

Ethical approval and accordance

Institutional review board (IRB) approval was waived, and the requirement for informed consent was exempted, as no human subjects, animal subjects, or protected health information were involved in the generation or evaluation of this data.

Data collection and question curation

A set of 50 FAQs regarding vertigo was compiled using Google Trends data (search term: "vertigo") and a review of highly trafficked patient support forums.

The FAQs were categorized into four domains: Diagnosis and etiology, treatment and interventions, home exercises and lifestyle and prognosis and expectations.

In March 2026, all fifty questions were inputted individually into ChatGPT-4 (OpenAI, San Francisco, CA, USA) via a "New Chat" session to minimize contextual bias.

Evaluation metrics

The responses were independently graded by evaluating clinicians using a 5-point Likert scale for two primary dimensions:

Accuracy

It was like (1=Completely inaccurate/dangerous to 5=completely accurate and consistent with current medical guidelines.

Comprehensiveness

It was like (1=Severely incomplete or missing key information to 5=fully comprehensive and covering all nuances of the question).

Readability assessment

Each response was analyzed using two standard readability formulas: Flesch-Kincaid grade level (FKGL) and Flesch reading ease (FRE).

Statistical analysis

Descriptive statistics (means and standard deviations) were calculated. Inter-rater reliability was assessed using the intraclass correlation coefficient (ICC).

A one-way analysis of variance (ANOVA) was utilized to identify statistical differences across the four question categories ($p<0.05$) as shown in the Figure 1 below.

The diagram outlines the systematic track splitting between blind clinician-driven qualitative scoring and standardized readability calculations for the 50 vertigo inquiries.

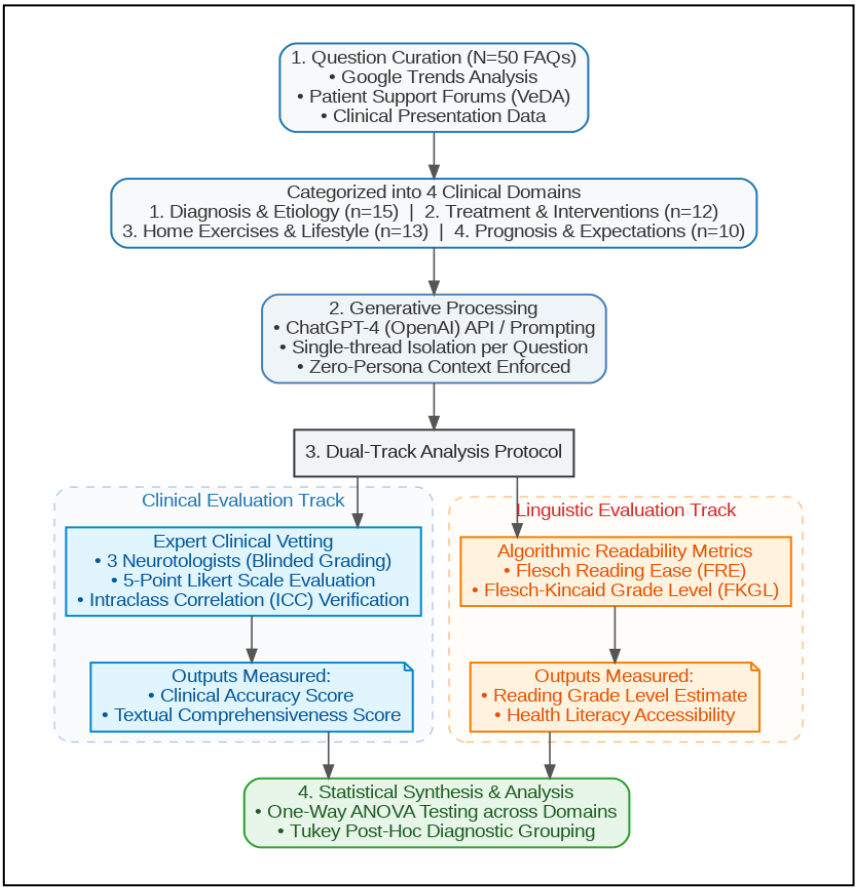

Figure 1: Methodological workflow of the cross-sectional evaluation paradigm.

RESULTS

Clinical accuracy and comprehensiveness

Across all 50 questions, the mean accuracy score was 4.38 (SD=0.62) out of 5, indicating a high level of factual correctness. None of the responses were rated as a '1'. The mean comprehensiveness score was 4.21 (SD=0.74) (Table 1). Additionally, the generated ChatGPT-4 responses had a mean word count of 315 (SD=48) words.

Table 1: Mean accuracy and comprehensiveness scores by question category.

Question category	Number of questions	Mean accuracy (SD)	Mean comprehensiveness (SD)
Diagnosis and etiology	15	4.05 (0.82)	3.95 (0.88)
Treatment and interventions	12	4.41 (0.55)	4.25 (0.68)
Home exercises and lifestyle	13	4.65 (0.45)	4.50 (0.52)
Prognosis and expectations	10	4.52 (0.51)	4.18 (0.75)
Overall	50	4.38 (0.62)	4.21 (0.74)

An ANOVA revealed a statistically significant difference in accuracy scores between the domains (p=0.024). ChatGPT performed significantly better in the home exercises and lifestyle category compared to the diagnosis and etiology category.

Readability assessment

The mean FKGL was 12.4 (SD=1.8), corresponding to a 12th-grade reading level. The mean FRE was 41.2 (SD=10.5), categorized as "difficult to read."

DISCUSSION

Overall, findings demonstrate that ChatGPT is highly accurate and comprehensive when disseminating vestibular health information, though its utility is constrained by a high required reading level and occasional diagnostic oversimplifications.[6,7]

Clinical accuracy and its limitations

Our reviewers rated the model's accuracy at 4.38 out of 5. ChatGPT excelled in the home exercises and lifestyle

category, accurately describing the Epley maneuver or dietary restrictions for Ménière's disease.[8] This demonstrates that for established, protocol-driven vestibular care, generative AI can serve as a dependable digital health asset, reinforcing home-based interventions. However, evaluating the reliability of AI as a standalone triage tool reveals persistent vulnerabilities. In a recent study by Hochstetler et al ChatGPT accurately answered only 45.5% of otolaryngology queries based on clinical practice guidelines, highlighting that while AI is conversational, its capacity for strict guideline adherence remains flawed.[9] Furthermore, when comparing ChatGPT directly against otolaryngology consultants on case-based questions, Hoch et al found that although AI generated significantly longer and highly cohesive responses, its actual medical adequacy and conciseness were significantly inferior to human experts.[10] This limitation is particularly evident in our diagnosis and etiology results. While the AI successfully identified common peripheral conditions like BPPV, it occasionally failed to emphasize crucial "red flags" (e.g., diplopia, dysarthria, sudden-onset tinnitus) that differentiate central nervous system pathologies from benign peripheral vestibular conditions.[11] Unlike an experienced clinician who dynamically triages a patient based on emergent symptoms, an LLM processes text statically, meaning it cannot execute clinical intuition or account for sudden clinical shifts. The omission of central warning signs poses a theoretical triage risk, reinforcing principle that AI should not replace professional clinical reasoning.[5]

Furthermore, because vestibular disorders frequently require objective diagnostic confirmation-such as pure-tone audiometry, videonystagmography (VNG), or imaging-the model's inability to interpret raw clinical data limits its adherence to strict evidence-based practice.

The readability challenge

Despite its factual correctness, ChatGPT failed to deliver information at a reading level accessible to the average patient (mean FKGL 12.4), which corroborates findings from other contemporary literature on the subject.[7] The American Medical Association recommends that patient-facing educational materials be written at or below an 8th-grade level to accommodate varying levels of health literacy. This readability gap is not isolated to vestibular disorders; it is a systemic limitation of generative AI across the otolaryngology spectrum. Soon et al evaluated ChatGPT's advice on ENT emergencies (e. g., quinsy, acute mastoiditis) and discovered that while the AI provided safe and accurate advice, the linguistic complexity of the text far exceeded the reading capabilities of the general public.[11] Similarly, Nielsen et al evaluated GPT-4 as a patient information source across various otolaryngology sub-specialties, noting that while the chatbot excelled in providing "highly relevant" answers, its sophisticated output often masked contextual depth required for true patient comprehension.[12] AI models generate text optimized for statistical linguistic probability, which in medical contexts results in dense polysyllabic terminology (e.g., "otoconia," "vestibulocochlear nerve," "endolymphatic hydrops"). In a high-anxiety clinical scenario like an acute vertigo attack, presenting a patient with overly complex jargon can inadvertently exacerbate psychological distress and reduce adherence to recommended lifestyle modifications. Physicians recommending AI as a supplementary resource must advise patients to explicitly prompt the AI to simplify its explanations. Research in AI prompt engineering suggests that iterative prompting can successfully lower the reading grade level of outputted medical text, bridging the health literacy gap.

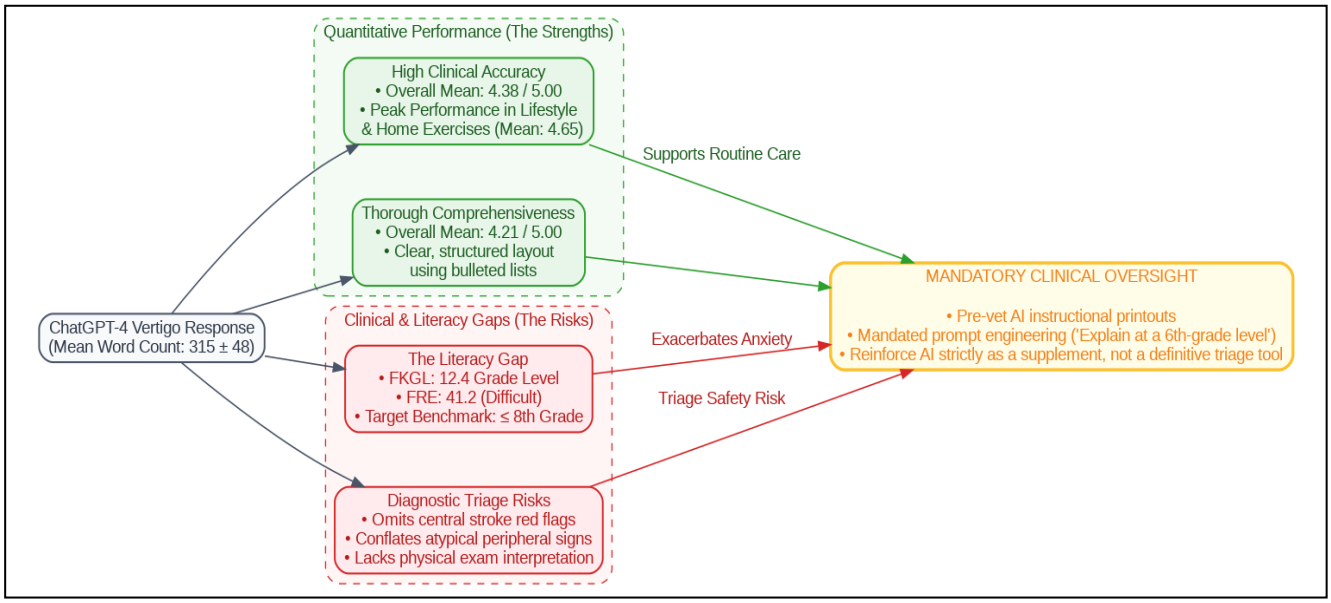

Figure 2: Conceptual schematic demonstrating the digital medical divergence.

*While ChatGPT-4 demonstrates robust guidelines-based alignment for stable lifestyle interventions, its advanced reading difficulty (12.4 grade level) and lack of central stroke triage limits autonomous clinical placement without peer oversight.

Implications for otolaryngologists and pediatricians

As AI-generated content increasingly populates both patient web searches and peer-reviewed literature, specialists must pivot from viewing AI as a disruptive force to leveraging it as a structured clinical ally.[8] Otolaryngologists and pediatricians can utilize LLMs to draft clear post-operative instructions, generate customized vestibular rehabilitation schedules, or translate complex audiogram results into patient-facing educational notes. However, given the model's variable adherence to specific clinical practice guidelines, all AI-generated content requires rigorous physician vetting and professional oversight prior to patient distribution.[5] Collaborative oversight ensures that pediatric-specific vestibular presentations-which often present differently than adult vertigo-are not overlooked by generic algorithmic models.

Limitations

This study has several limitations. First, ChatGPT's LLM is dynamically updated; therefore, responses generated in March 2026 may differ from future iterations due to continuous reinforcement learning. Second, the Likert scale grading system is inherently subjective, despite strong inter-rater reliability. Third, our curated questions were distinct, isolated prompts. In reality, a patient's interaction with a chatbot is conversational, and the AI's diagnostic accuracy might change upon consecutive follow-up prompting in a single thread. Future studies should evaluate multi-turn conversational AI interactions with actual vertigo patients to assess real-time anxiety reduction and understanding.

CONCLUSION

ChatGPT-4 serves as a highly accurate and comprehensive repository of information for common patient questions regarding vertigo and vestibular disorders. It performs exceptionally well in describing lifestyle modifications and established treatments. Otolaryngologists and pediatricians should acknowledge patients' use of these platforms and actively guide them on how to critically interpret this data.

Declaration

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

Funding: No funding sources
Conflict of interest: None declared
Ethical approval: Not required

REFERENCES

1. Lechien JR, Rameau A. Applications of ChatGPT in Otolaryngology-Head Neck Surgery: A State of the art Review. Otolaryngol Head Neck Surg. 2024;171(3):667-77.
2. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on Medical Questions in the National Medical Licensing Examination in Japan: Evaluation Study. JMIR Form Res. 2023;7:e48023.
3. Zalzal HG, Abraham A, Cheng J, Shah RK. Can ChatGPT help patients answer their otolaryngology questions? Laryngoscope Investig Otolaryngol. 2023;9(1):e1193.
4. Langlie J, Kamrava B, Pasick LJ, Mei C, Hoffer ME. Artificial intelligence and ChatGPT: An otolaryngology patient's ally or foe? Am J Otolaryngol. 2024;45(3):104220.
5. Tessler I, Wolfovitz A, Alon EE, Gecel NA, Livneh N, Zimlichman E, et al. ChatGPT's adherence to otolaryngology clinical practice guidelines. Eur Arch Otorhinolaryngol. 2024;281(7):3829-34.
6. Liu X, Shi S, Zhang X, Qianwen G, Wuqing W. The role of ChatGPT-4o in differential diagnosis and management of vertigo-related disorders. Sci Rep. 2025;15(1):18688.
7. Bellinger JR, De La Chapa JS, Kwak MW, Ramos GA, Morrison D, Kesser BW. BPPV Information on Google Versus AI (ChatGPT). Otolaryngol Head Neck Surg. 2024;170(6):1504-11.
8. Carnino JM, Chong NYK, Bayly H, Salvati LR, Tiwana HS, Levi JR. AI-generated text in otolaryngology publications: a comparative analysis before and after the release of ChatGPT. Eur Arch Otorhinolaryngol. 2024;281(11):6141-6.
9. Maksimoski M, Noble AR, Smith DF. Does ChatGPT Answer Otolaryngology Questions Accurately? Laryngoscope. 2024;134(9):4011-5.
10. Buhr CR, Smith H, Huppertz T, Bahr-Hamm K, Matthias C, Blaikie A, et al. ChatGPT Versus Consultants: Blinded Evaluation on Answering Otorhinolaryngology Case-Based Questions. JMIR Med Educ. 2023;9:e49183.
11. Soon S, Perry B. Paging Dr. ChatGPT: safety, accuracy and readability of ChatGPT in ENT emergencies. Aust J Otolaryngol. 2025;8:8.
12. Nielsen JPS, von Buchwald C, Grønhøj C. Validity of the large language model ChatGPT (GPT4) as a patient information source in otolaryngology by a variety of doctors in a tertiary otorhinolaryngology department. Acta Otolaryngol. 2023;143(9):779-82.

Cite this article as: Poojary SC, Nayak P. Factual accuracy versus the health literacy barrier: a blinded neurotological assessment of ChatGPT in vertigo education. Int J Otorhinolaryngol Head Neck Surg 2026;12:608-12.