EuropeanJournal of
Human Health

ORIGINAL ARTICLE

ANALYZING AI CHATBOT REACTIONS TO TOP SEARCHES ON NEUROMYELITIS OPTICA SPECTRUM DISORDERS (NMOSD): A NEED FOR FURTHER EXPLORATION

Emine Rabia KOC [1]

[1] Department of Neurology, Uludag University, Faculty of Medicine, Bursa, Turkey.

ABSTRACT

Aim: Artificial intelligence chatbots (AICs) in the health field have become increasingly common, allowing individuals to gather information about their health conditions. However, concerns remain regarding the accuracy, reliability, ethics, and security of these AICs' information in medicine. This study aims to assess the reliability and quality of the information provided by AICs for NMOSD disease, symptoms, and treatment options while discussing potential benefits and disadvantages for NMOSD patients.

Methods: The study aimed to evaluate the responses of three AICs, ChatGPT, Gemini, and Perplexity when asked about the most frequently searched keyword for NMOSD via Google Trends. The responses were assessed by three separate examiners for readability, understandability, actionability, reliability, and transparency.

Results: Based on the Coleman-Liau index, the responses were challenging to read and suitable for professionals. Perplexity PEMAT-P scored highest in understandability (50%) compared to Gemini(40%) and ChatGPT(40%). Regarding PEMAT-P actionability scores, Gemini scored the highest (48%), while ChatGPT obtained the lowest(37%). The reliability of responses varied from poor to fair. The treatment information quality was assessed using the DISCERN score, and it was found that ChatGPT received the lowest score while Perplexity received the highest. None of the AI chatbots addressed the side effects of treatment, potential consequences of not undergoing treatment, effects on quality of life, or shared decision-making.

Conclusions: It is important to address the accuracy and reliability of these technologies before full integration into the medical field. Patients should critically evaluate information from AI chatbots and be cautious about relying solely on them for health-related decisions.

Keywords: Artificial Intelligence, chatbots, Quality Assessment, NMOSD

Corresponding Author: Emine Rabia Koc erabiakoc@uludag.edu.tr

Received: ***November 2, 2024;*** *Accepted:* ***November 11, 2024;*** *Published Online:* ***November 12, 2024***

Cite this article as: Koc, E., R. (2024). Analyzing AI Chatbot Reactions to Top Searches on Neuromyelitis Optica Spectrum Disorders (NMOSD): A Need for Further Exploration. European Journal of Human Health 4(4),98-107.

INTRODUCTION

NMOSD is a rare autoimmune disorder that leads to recurring inflammatory attacks affecting the optic nerve, spinal cord, and brain[1]. AQP4-IgG antibodies are a diagnostic indicator of the condition and substantially impact its pathogenesis**.** The frequency of NMOSD varies globally and by region, with rates between 0.5 and 10 per 100,000 people in most populations.[2–4]. NMOSD usually follows a clinical pattern characterized by recurrent attacks leading to cumulative neurological disabilities over time. Because of the recurring nature of the attacks and the fact that they cause permanent disability in the long term, they should be recognized early by physicians and treatment should be started early. Even among us clinicians, recognizing the symptoms and signs of NMOSD, diagnosing the disease, and starting treatment can sometimes be delayed [5].

People use artificial intelligence tools quite frequently today, both at the first symptom stage and at the diagnosis and treatment stage, to learn about the accuracy of the diagnosis and treatment or whether there are different treatment options [6]. The Internet is an invaluable tool for patients to independently research and gather information about their health conditions. Patients must receive accurate and sufficient information from reliable sources during this process. Artificial intelligence chatbots (AIC) have made significant progress in the field of health as in every area of life, and have provided people with data on which tests should be performed at the time of diagnosis and treatment methods from the symptom-based diagnosis stage. However, there are still doubts about the accuracy, reliability, ethics, and security of the information provided by these artificial intelligence robots in medicine [7–9]. There is a lack of literature assessing the quality, understandability, readability or actionability of information on NMOSD provided by AI chatbots.

Therefore, before relying on AICs in the symptom diagnosis and treatment process, it is crucial to evaluate the quality of the information provided by AICs comprehensively. This study aims to assess the reliability and quality of the information provided by AICs for NMOSD disease, symptoms, and treatment options and to discuss the potential benefits and disadvantages of using these AICs by NMOSD patients.

Material and Methods

We used Google Trends to find the most popular search queries about NMOSD worldwide from January 1, 2021, to January 1, 2024. We then converted them into questions and asked three different AICs (Perplexity, ChatGPT, Gemini) these questions. Our first question was “What is NMOSD?”, the second was “What are the main symptoms of NMOSD?”, and the last question was “What are the treatment options for NMOSD?”. The study design is summarized in figure 1. Three different examiners recorded the response outputs from the AIC on three separate computers.

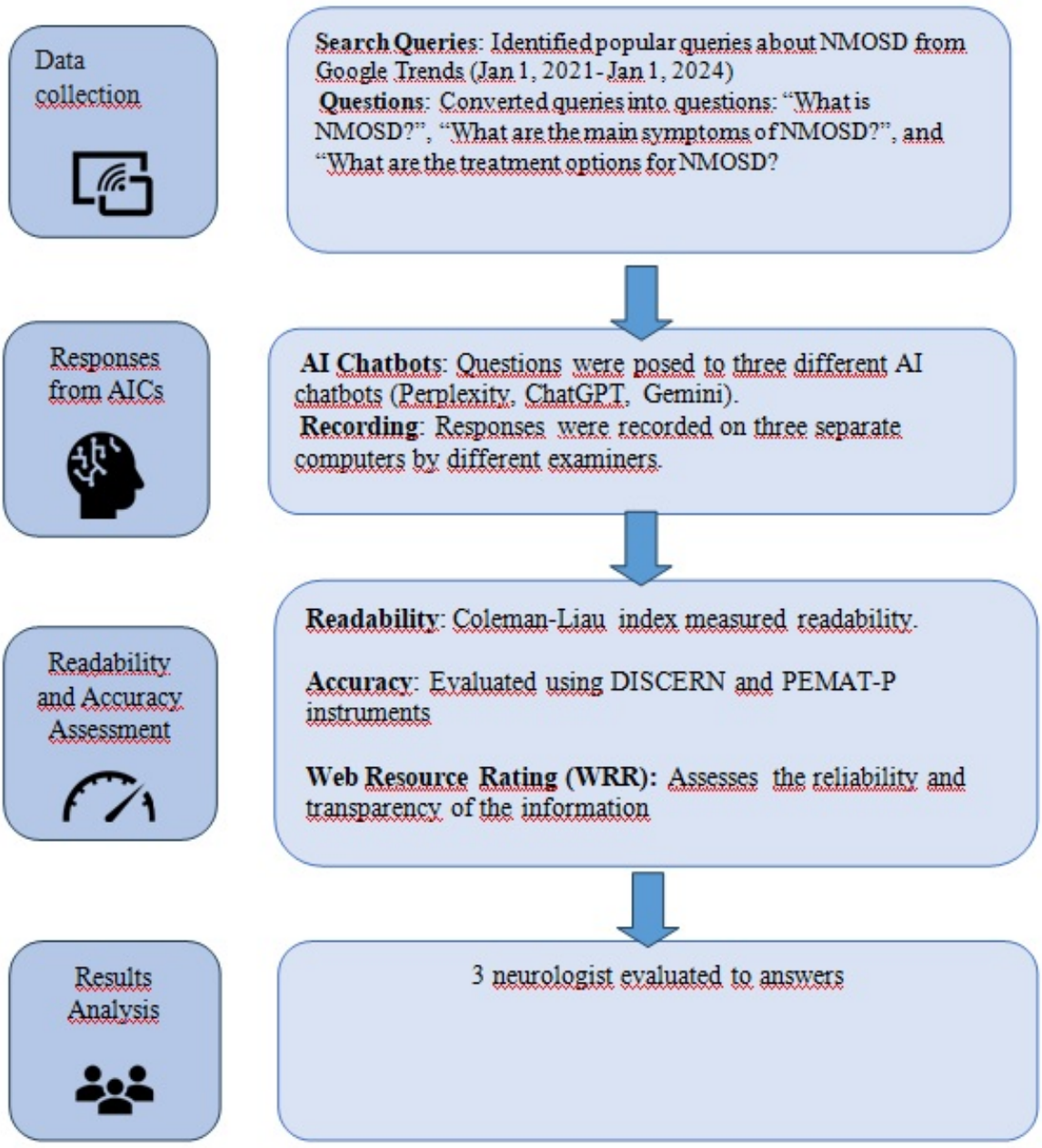

Figure 1: Study Design 1

The Coleman-Liau index was utilized to measure the readability of the text; values above 11 indicate that reading is challenging and suggest that at least a college-level education is required [10]. To assess the accuracy of healthcare details supplied by each chatbot DISCERN and PEMAT-P instruments were used. The examiners then computed the DISCERN score, which evaluates the reliability and quality of medical information based on specific criteria related to treatment choices (Questions 1-8) and specific details of treatment options (Questions 9-15). This scoring system consists of 16 questions from 1 to 5. The final question provides an overall quality rating. DISCERN scores are interpreted as follows: 16-26 indicates poor quality, 27-38 indicates low quality, 39-50 indicates average quality, 51-62 indicates good quality, and 63-75 indicates excellent quality [11]. An Intraclass Correlation Coefficient (ICC) of 0.962 (0.958-0.986) with a significance level of $p< .001$ was computed for the assessment of DISCERN scores. The understandability and actionability of answers were assessed with PEMAT-P (scored as a percentage). The higher the score, the more understandable or actionable the material [12]. Additionally, we used the Web Resource Rating (WRR) scale developed to measure the reliability and transparency of internet-derived information; this was also scored between 0-100% [13]. The median values (minimum-maximum) were reported for the number of words, the Coleman-Liau index, and DISCERN results. The data were analyzed using IBM SPSS V23 and IBM SPSS AMOS V24 (IBM Corp, Armonk, NY, USA).

Results

Response outputs given from 3 different AICs were evaluated by three different examiners by using "Update on the diagnosis and treatment of neuromyelitis optica spectrum disorders (NMOSD)- revised recommendations of the Neuromyelitis Optica Study Group (NEMOS) Part I: Diagnosis and differential diagnosis guidelines were used to evaluate the answers, and part II: Attack therapy and long-term management" papers [1].

Based on the Coleman-Liau index, it was noted that the responses were challenging to read and were at a level suitable for professionals. In the evaluation of chatbot PEMAT-P understandability, it was found that Perplexity (50%) scored higher in understandability compared to Gemini (40%) and ChatGBT (40%). When AICs were assessed for PEMAT-P actionability scores, Gemini achieved the highest score (48%), while ChatGBT obtained the lowest (37%). Additionally, according to the WRR scale, the responses were determined to have reliability ranging from poor to fair (27%-58.2%). These findings are summarized in Table 1.

Table 1: Evaluation of the results of readability, understandability, and reliability of chatbots

Chatbots	ChatGPT	Gemini	Perplexity
Word Count	176 (94-284)	137 (135-139)	255 (73-378)
Coleman Liau Index	15.7 (12.1-19.3)	14.2 (13.5-14.5)	18.2 (15.4-20.8)
PEMAT-P Understandability, %	40 (27.3-47)	40 (30-55.5)	50 (33.3-68.3)
PEMAT-P Actionability, %	37 (27-60)	48 (40-60)	42.9 (20-60)
WRR, %	27 (17.8-34.6)	49.1 (47.5-64.2)	58.2 (50-75)

PEMAT-P: The Patient Education Materials Assessment Tool-Printable materials, WRR: Web Resource Rating

Table 2 displays the median values of responses for each question in the DISCERN score, which assesses the quality of treatment information. Upon evaluating the total DISCERN score, it was found that ChatGPT received the lowest score while Perplexity received the highest. Based on this scoring system, ChatGPT and Gemini were rated as providing low-quality information, whereas Perplexity was deemed to offer average quality. Amongst the chatbots assessed, ChatGPT consistently scored lowest in certain questions due to its failure to specify references. None of the AI chatbots addressed the side effects of treatment or potential consequences of not undergoing treatment, effects on quality of life, or shared decision-making; thus, they received a score of 1 for each question. In contrast, Perplexity excelled in these aspects because it provided clear information sources used in compiling publications and details about when this information was produced and it was up-to-date, as well as easy accessibility to additional support and information sources.

Table 2: Median score per DISCERN question amongst all AI C

	DISCERN question	ChatGPT	Gemini	Perplexity
2	Does it achieve its aims?	5(5-5)	3(3-4)	5(2-5)
3	Is it relevant?	4(4-4)	3(3-4)	3(2-4)
4	Is it clear what sources of information were used to compile the publication (other than the author or producer)?	1(1-1)	3(2-4)	4(4-5)
5	Is it clear when the information used or reported in the publication was produced?	1(1-1)	3(2-4)	4(4-5)
6	Is it balanced and unbiased?	5(5-5)	4(4-5)	5(4-5)
7	Does it provide details of additional sources of support and information?	1(1-1)	4(4-5)	4(4-5)
8	Does it refer to areas of uncertainty?	1(1-1)	1(1-1)	1(1-3)
9	Does it describe how each treatment works?	1(1-4)	1(1-2)	4(1-5)
10	Does it describe the benefits of each treatment?	1(1-4)	1(1-1)	1(1-3)
11	Does it describe the risks of each treatment?	1(1-1)	1(1-1)	1(1-1)
12	Does it describe what would happen if no treatment is used?	1(1-1)	1(1-1)	1(1-1)
13	Does it describe how the treatment choices affect overall quality of life?	1(1-1)	1(1-1)	1(1-3)
14	Is it clear that there may be more than one possible treatment choice?	1(1-4)	1(1-4)	1(1-5)
15	Does it provide support for shared decision making?	1(1-3)	1(1-1)	1(1-1)
16	Based on the answers to all of the above questions, rate the overall quality of the publication as a source of information about treatment choices?	2(2-3)	3(2-3)	3(2-4)
1-15	Total DISCERN score	31(30-38)	35(34-36)	46(33-48)

Discussions

Evaluating the quality of consumer health information provided by AI chatbots on NMOSD is a novel aspect addressed in this study. Given the increasing usage of these platforms, it is crucial to assess the accuracy and reliability of health information available through AI chatbots.The COVID-19 pandemic has markedly boosted the use of health-oriented chatbots, which now serve various functions such as answering questions, assessing symptoms, recommending care options, and facilitating tasks like booking appointments. These digital assistants offer several benefits, including the ability to deliver medical information drawn from extensive health datasets, 24/7 accessibility, and immediate information retrieval. They are also useful for postoperative education, personalized health monitoring, and assisting with diagnostic and treatment decisions. However, their use comes with notable drawbacks, such as the potential decrease in human interaction, concerns about data security, challenges with accuracy and reliability, ethical considerations that need to be carefully managed [14–16].

Patients with chronic diseases, in particular, are looking for a second opinion, support, or additional information about their illness [17, 18]. A study conducted on MS patients, one of the demyelinating diseases of the central nervous system, showed that MS patients are more likely to explore online resources and often turn to the Internet as their primary source of information [19]. There is no study has been conducted on this subject in the NMOSD patient group. NMOSD is not a disorder that is well-known and easily diagnosed also by branches other than neurology physicians. Since NMOSD is among the rare diseases, it is likely that a person will seek more comprehensive information in this area when faced with NMOSD diagnosis. As a result, chatbots are likely to appeal to this group as a valuable tool for obtaining information and support. At this point, accessing accurate, reliable, and understandable information is of great importance.

The quality and accuracy of information provided by healthcare chatbots can vary significantly. Our study revealed that the information from AI chatbots generally ranged from low to moderate quality, with poor to moderate reliability and understandability. On the other hand, studies have shown that AI chatbots may have difficulty conveying information about NMOSD in a way that the average person can easily understand [20, 21]. This finding suggests that AI chatbots use medical terminology that may not be useful to the lay audience. Furthermore, 2 out of 3 chatbots may have difficulty conveying complex medical information, as they do not adequately use visual aids such as figures and tables. Some chatbots lacked scientifically based references and occasionally generated non-existent references when asked. Additionally, many references were outdated, failing to reflect recent developments in the healthcare field. Moreover, none of the AI chatbots addressed important aspects such as treatment side effects, potential consequences of

forgoing treatment, impacts on quality of life, or shared decision-making. To enhance the applicability and reliability of information from AI chatbots, it would be useful to present up-to-date, scientifically based content in clearer, more understandable language and to include features that support shared decision-making in new versions.

Since artificial intelligence chatbots have limited demographic, clinical, laboratory, and radiological information about users, more personal data may need to be shared for correct diagnosis and treatment, which will bring about some ethical and security problems [16].

The limitation of our study is that, as the chatbots' own search trends are not publicly accessible, the research examined AI chatbot reactions to commonly searched topics on Google Trends. More investigation is required to assess ongoing conversations with more intricate queries, information accuracy from paid AI chatbots, and the impact of using different words and sentence structures on response quality.

In conclusion, AI chatbots should be considered a supplementary resource rather than a substitute for healthcare professionals lacking human empathy and experience. Before their full integration into the medical field, it is essential to address the accuracy and reliability of these technologies. Users must critically evaluate sources from AI chatbots. While AICs offer easy access to information, they often provide unclear or error-prone responses. Patients should be aware that AICs may not offer specific disease treatment options and instead present low-moderate-quality, unreliable information. It is crucial that patients do not base treatment decisions solely on information from these sources.

Competing interests: There are no conflicts of interest

Funding: Not Applicable

References:

1. Wingerchuk DM, Banwell B, Bennett JL, et al (2015) International consensus diagnostic criteria for neuromyelitis optica spectrum disorders. Neurology 85:177–89. https://doi.org/10.1212/WNL.0000000000001729
2. Pandit L, Asgari N, Apiwattanakul M, et al (2015) Demographic and clinical features of neuromyelitis optica: A review. Mult Scler 21:845–53. https://doi.org/10.1177/1352458515572406
3. Flanagan EP, Cabre P, Weinshenker BG, et al (2016) Epidemiology of aquaporin-4 autoimmunity and neuromyelitis optica spectrum. Ann Neurol 79:775–783. https://doi.org/10.1002/ana.24617
4. Bukhari W, Khalilidehkordi E, Mason DF, et al (2022) NMOSD and MS prevalence in the Indigenous populations of Australia and New Zealand. J Neurol 269:836–845. https://doi.org/10.1007/s00415-021-10665-9
5. Delgado-Garcia G, Lapidus S, Talero R, Levy M (2022) The patient journey with NMOSD: From initial diagnosis to chronic condition. Front Neurol 13:966428. https://doi.org/10.3389/fneur.2022.966428
6. https://trends.google.com/trends/explore?cat =45&q=NMOSD&hl=en
7. Safdar NM, Banja JD, Meltzer CC (2020) Ethical considerations in artificial intelligence. Eur J Radiol 122:108768. https://doi.org/10.1016/j.ejrad.2019.108768
8. Keskinbora KH (2019) Medical ethics considerations on artificial intelligence. Journal of Clinical Neuroscience 64:277–282. https://doi.org/10.1016/j.jocn.2019.03.001
9. London AJ, Karlawish J, Largent EA, et al (2024) Algorithmic identification of persons with dementia for research recruitment: ethical considerations. Inform Health Soc Care 49:28–41. https://doi.org/10.1080/17538157.2023.2299881
10. Coleman M, Liau TL (1975) A computer readability formula designed for machine scoring. 1975;60:283. Journal of Applied Psychology 60:283
11. Charnock D, Shepperd S, Needham G, Gann R (1999) DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health (1978) 53:105–11. https://doi.org/10.1136/jech.53.2.105
12. https://www.ahrq.gov/health-literacy/patient-education/pemat-p.html
13. Dobbins M, Watson S, Read K, et al (2018) A Tool That Assesses the Evidence, Transparency, and Usability of Online Health Information: Development and Reliability Assessment. JMIR Aging 1:e3. https://doi.org/10.2196/aging.9216
14. Parviainen J, Rantala J (2022) Chatbot breakthrough in the 2020s? An ethical reflection on the trend of automated consultations in health care. Med Health Care Philos 25:61–71. https://doi.org/10.1007/s11019-021-10049-w
15. Garcia Valencia OA, Suppadungsuk S, Thongprayoon C, et al (2023) Ethical Implications of Chatbot Utilization in Nephrology. J Pers Med 13:1363. https://doi.org/10.3390/jpm13091363
16. Laymouna M, Ma Y, Lessard D, et al (2024) Roles, Users, Benefits, and Limitations of Chatbots in Health Care: Rapid Review. J Med Internet Res 26:e56930. https://doi.org/10.2196/56930
17. Askin A, Sengul L, Tosun A (2020) YouTube as a Source of Information for Transcranial Magnetic Stimulation in Stroke: A Quality, Reliability and Accuracy Analysis. J Stroke Cerebrovasc Dis 29:105309. https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.105309
18. Ruiz-Roca JA, Martínez-Izquierdo A, Mengual-Pujante D, et al (2020) Is YouTube a useful tool for oral care in patients with Parkinson's disease? Special Care in Dentistry 40:464–469. https://doi.org/10.1111/scd.12489

19. Altunisik E, Firat YE, Kiyak Keceli Y (2022) Content and quality analysis of videos about multiple sclerosis on social media: The case of YouTube. Mult Scler Relat Disord 65:104024. https://doi.org/10.1016/j.msard.2022.104024
20. Pan A, Musheyev D, Bockelman D, et al (2023) Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries About Cancer. JAMA Oncol 9:1437–1440. https://doi.org/10.1001/jamaoncol.2023.2947
21. Musheyev D, Pan A, Loeb S, Kabarriti AE (2024) How Well Do Artificial Intelligence Chatbots Respond to the Top Search Queries About Urological Malignancies? Eur Urol 85:13–16. https://doi.org/10.1016/j.eururo.2023.07.004