Commentary

ChatGPT's Inconsistency in the Diagnosis of Alzheimer's Disease

ArunSundar MohanaSundaram[a], Bhushan Patil[b,c] and Domenico Praticò[d,*]
[a] *School of Pharmacy, Sathyabama Institute of Science and Technology, Chennai, India*
[b] *MannSparsh Neuropsychiatric Hospital, Kalyan, India*
[c] *Manasa Rehabilitation and De-Addiction Center, Titwala, India*
[d] *Alzheimer's Center at Temple, Lewis Katz School of Medicine, Temple University, Philadelphia, PA, USA*

Received 11 April 2024
Accepted 4 May 2024
Published 27 May 2024

Abstract. A recent article by El Haj et al. provided evidence that ChatGPT could be a potential tool that complements the clinical diagnosis of various stages of Alzheimer's Disease (AD) as well as mild cognitive impairment (MCI). To reassess the accuracy and reproducibility of ChatGPT in the diagnosis of AD and MCI, we used the same prompt used by the authors. Surprisingly, we found that some of the responses of ChatGPT in the diagnoses of various stages of AD and MCI were different. In this commentary we discuss the possible reasons for these different results and propose strategies for future studies.

Keywords: Alzheimer's disease, mild cognitive impairment, brain aging, diagnosis, artificial intelligence, ChatGPT

We have read with great interest, the article "ChatGPT as a Diagnostic Aid in Alzheimer's Disease: An Exploratory Study" written by El Haj et al. [1], in which the authors provide evidence that ChatGPT could be a potential tool that complements the clinical diagnosis of various stages of Alzheimer's disease (AD) as well as mild cognitive impairment (MCI). To reassess the accuracy and reproducibility of ChatGPT in the diagnosis of AD and MCI, we used the same prompt used by the authors [1]. Surprisingly, we found that the responses of ChatGPT in the diagnosis of various stages of AD and MCI were different. In our study, case 1 (which was diagnosed as "Mild AD" in the study by El Haj et al. [1]) was classified as "early to moderate AD". When the diagnostic responses to case-2 & 4 were re-analyzed, we obtained "moderate to severe AD" and "normal ageing or possibly MCI" instead of moderate AD and MCI, respectively. However, the diagnosis for "advanced stage of AD" we obtained was consistent with that of their findings [1].

The transcript of the ChatGPT's diagnosis of AD and MCI obtained in our study is provided in the Supplementary Material. It is important to note that for this exercise we have used ChatGPT 3.5. However, El Haj et al. have not specified the exact version of ChatGPT used for their study [1]. One plausible reason for the different results is that the authors might have used an advanced version (i.e., ChatGPT-4). Irrespective of the version of ChatGPT employed in both studies, inconsistency in the diagnostic response of AD/MCI (or any other condition) is without doubt a timely and at the same time pressing concern that needs some further consideration.

In our study, the diagnosis of case-1 as "early to moderate AD" was due to the following flawed

[*] Correspondence to: Domenico Praticò, MD, FCPP, Alzheimer's Center at Temple, Lewis Katz School of Medicine, Temple University, Philadelphia, PA 19140, USA. E-mail: praticod@temple.edu.

assumptions by ChatGPT: 1) deficits in multiple cognitive domains represent moderate AD, 2) substantial temporo-parietal atrophy, chiefly in the hippocampus, and 3) increased levels of tau and phospho-tau protein bolster the diagnosis of moderate AD. Our ChatGPT's diagnosis of case-2 as "moderate to severe AD" was based on the consideration of the following conditions as moderate to severe AD: 1) memory-related cognitive dysfunctions and inability to perform basic daily activities and 2) remarkable temporo-parietal atrophy, chiefly in the hippocampus.

Although the categorization of case-3 as "severe AD" is consistent with our own study result, a few nuances are still missing in the ChatGPT response obtained in our study. Based on the analysis of the blood test, ChatGPT failed to rule out any acute medical issue that might underlie the advanced stage of this AD case. Finally, in our study case-4 was impeccably categorized as "normal ageing or possibly MCI", instead of the categorization as MCI presented in the study by El Haj et al. [1] We reason that this diagnosis of ChatGPT in our study is justifiable because the various markers are not dramatically different between MCI and age-matched control subjects.

Physicians and other healthcare providers, patients, caregivers, and even healthy adults who are utilizing ChatGPT as a diagnostic tool for AD/MCI should be very cautious and should not consider ChatGPT as a definitive and fully reliable diagnostic aid, at least the ChatGPT-3.5 version.

Interestingly, a recent study by Huang et al. demonstrated that the ChatGPT, although showed accuracy in identifying the myths associated with AD, was not 100% aligned with the clinician opinion on the cases [2]. Another paper by Cao et al. demonstrated that the responses provided by ChatGPT in the context of hepatocellular carcinoma surveillance and diagnosis are inaccurate and unreliable [3]. Besides, analysis of ChatGPT for the self-diagnosis potential in common orthopedic diseases showed that this AI chatbot is inconsistent in terms of accuracy and reproducibility [4]. ChatGPT failed to reliably and accurately assess the efficacy of an array of drugs used for the prevention of episodic and chronic migraine [5]. The reason underpinning this problem was attributed to the "AI hallucinations" wherein a remarkably high amount (66%) of fake and non-existent sources were used for the assessment. Furthermore, the use of some predatory and hijacked journals as standard references might also contribute to its inaccurate and inconsistent responses [6].

Taken together, the outcome of our study underscores the need to exercise great caution in the utility of ChatGPT as a reliable diagnostic tool in any stage of AD or MCI. However, we should not forget that generative AI technologies like ChatGPT are somewhat still at their early stages of development. We believe that future development and advancements in the field and the implementation of more appropriate data training will improve the capability of ChatGPT to diagnose AD/MCI and other health conditions accurately and consistently.

AUTHOR CONTRIBUTIONS

ArunSundar MohanaSundaram (Conceptualization; Data curation; Writing – original draft; Writing – review & editing); Bhushan Patil (Conceptualization; Data curation; Writing – original draft; Writing – review & editing); Domenico Pratico (Conceptualization; Writing – original draft; Writing – review & editing).

ACKNOWLEDGMENTS

The authors have no acknowledgments to report.

FUNDING

The authors have no funding to report.

CONFLICT OF INTEREST

The authors have no conflict of interest to report.

SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: https://dx.doi.org/10.3233/ADR-240069.

REFERENCES

[1] El Haj M, Boutoleau-Bretonnière C, Gallouj K, Wagemann N, Antoine P, Kapogiannis D, Chapelet G (2024) ChatGPT as a diagnostic aid in Alzheimer's disease: An exploratory study. *J Alzheimers Dis Rep* **8**, 495-500.

[2] Huang SS, Song Q, Beiting KJ, Duggan MC, Hines K, Murff H, Leung V, Powers J, Harvey TS, Malin B, Yin Z (2023) Fact check: Assessing the response of ChatGPT to Alzheimer's disease statements with varying degrees of misinformation. *medRxiv*, doi: https://doi.org/10.1101/2023.09.04.23294917 [Preprint]. Posted September 07, 2023.

[3] Cao JJ, Kwon DH, Ghaziani TT, Kwo P, Tse G, Kesselman A, Kamaya A, Tse JR (2023) Accuracy of information provided by ChatGPT regarding liver cancer surveillance and diagnosis. *AJR Am J Roentgenol* **221**, 556-559.

[4] Kuroiwa T, Sarcon A, Ibara T, Yamada E, Yamamoto A, Tsukamoto K, Fujita K (2023) The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: Exploratory study. *J Med Internet Res* **25**, e47621.

[5] Moskatel LS, Zhang N (2023) The utility of ChatGPT in the assessment of literature on the prevention of migraine: An observational, qualitative study. *Front Neurol* **14**, 1225223.

[6] Dadkhah M, Oermann MH, Hegedüs M, Raman R, Dávid LD (2024) Diagnosis unreliability of ChatGPT for journal evaluation. *Adv Pharm Bull* **14**, 1-4.