p-ISSN: 2184-2205 e-ISSN: 2184-3414 ISBN: 978-989-35106-6-7 © 2024

EVALUATING CHATGPT'S DIAGNOSTIC CAPABILITIES FOR MENTAL HEALTH DISORDERS

Asaf Wishnia[1], Eyal Rosenstreich[1,2], & Uzi Levi[1]
[1]*Department of Behavioral Sciences, Peres Academic Center (Israel)*
[2]*School of Human Movement & Sport, Levinsky-Wingate Academic Center (Israel)*

Abstract

The field of artificial intelligence (AI) has seen significant advancements in recent years, making it a notable technological achievement in various aspects of daily life. In this study, we sought to investigate the feasibility of employing AI in the realm of mental health. Specifically, we assessed the efficacy of ChatGPT as a diagnostic tool for mental health disorders. To this end, 25 vignettes depicting common mental disorders were presented to ChatGPT, and its diagnostic accuracy was evaluated across three experimental conditions (the original vignette, the vignette with gender switch, and a shortened version of the vignette). The results showed high accuracy rate, surpassing random guessing, and highlighted ChatGPT's adherence to specific diagnostic criteria. This accuracy persisted even when the vignettes depicted rare mental disorders. These findings are discussed with an emphasis on potential gender biases, the risks tied to self-diagnosis, and the pressing need for further validation and ethical considerations. The study concludes by addressing the potential for incorporating ChatGPT into the broader realm of mental health in the future.

Keywords: *ChatGPT, mental health, self-diagnose, artificial intelligence.*

1. Introduction and objectives

The internet has become a pivotal source for diagnosing physical and mental health conditions. It offers rapid access to medical information but also poses risks like misdiagnosis and increased anxiety (Bessiere et al., 2010). This dual nature is exemplified by the mixed accuracy of online symptom checkers, with Semigran et al. (2015) reporting an average correctness rate of around 49.67%.

In mental health, the challenge is more pronounced due to the subjective nature of diagnoses and lack of standardization (cf. Lupton, 2014). To address this, AI technologies like ChatGPT are being explored for their potential in improving diagnosis accuracy and efficiency. Since its 2022 release, ChatGPT has shown promise in various fields, suggesting its utility in standardizing mental health diagnoses.

The present study assesses ChatGPT's efficacy as a tool for mental health professionals and for individual self-diagnosis. We aim to evaluate its diagnostic accuracy across a spectrum of mental disorders and its ability to consider gender-specific symptoms and interpret brief symptom descriptions. This research is significant in understanding ChatGPT's potential and limitations as a mental health diagnostic aid.

2. Method

2.1. Materials and procedure

This study utilized 25 vignettes depicting four common mental disorders: Depression, Anxiety, PTSD, and OCD, sourced from scholarly literature, mainly case studies with diagnoses by mental health professionals. Resources included Google Scholar and PubMed, focusing on studies in reputable academic journals with specific diagnoses and patient-reported symptoms. The Psychodynamic Diagnostic Manual (PDM) aided in identifying relevant studies, emphasizing patient experiences over professional opinions. Due to ChatGPT's limitations, vignette quotations were adapted into first-person narratives. Additionally, four rare disorders (Cotard's Syndrome, Capgras Syndrome, Alien Hand Syndrome, and Pica) were included to test ChatGPT's diagnostic range, defining rarity as less than 1% diagnosis rate by professionals.

Conducted from February to April 2023, the study involved ChatGPT V3.5 and V4, and consisted of three conditions. The first condition involved submitting the adjusted original vignettes into ChatGPT V3 to assess its diagnostic accuracy. The second condition altered patient genders in the vignettes to explore

ChatGPT's gender-sensitive diagnostic capability. The third condition, coinciding with ChatGPT-4's release, involved recalibrating the same vignettes with critical symptoms removed, aligning with the DSM-V, to test diagnostic accuracy with incomplete information. A separate, fourth, condition focused on rare mental health cases, utilizing ChatGPT-4 exclusively, with cases sourced from research on PubMed, to evaluate ChatGPT's detection capabilities in rare disorders.

3. Results

In the first condition, ChatGPT achieved a 96% accuracy in diagnosing 25 original vignettes, $\chi^2(1) = 21.160$, $p < .001$, indicating significantly better performance than chance. In the second condition (gender-switched vignettes), ChatGPT maintained a high accuracy of 95.23% on 21 vignettes, $\chi^2(1) = 17.190$, $p < .001$. In the third condition (shortened vignettes), accuracy was 86.95%, $\chi^2(1) = 12.565$, $p < .001$.

ChatGPT also correctly identified all rare mental health conditions, but no statistical analysis was conducted due to the small sample size. Chi-square tests comparing accuracy in the original and the altered vignettes, revealed no significant differences for the gender switched vignettes, $\chi^2(1) = .053$, $p = .819$, as well as for the shortened vignettes, $\chi^2(1) = 1.57$, $p = .692$.

4. Discussion and conclusions

The study aimed to evaluate ChatGPT as a diagnostic tool for mental health disorders across different scenarios. ChatGPT showed high diagnostic accuracy across the three experimental conditions: original, gender-switch, and symptom elaboration. This aligns with previous research on AI in mental health (e.g., Davenport & Kalakota, 2019), highlighting AI's potential but also the need for validation and ethical considerations. The study also examined gender biases in diagnoses, revealing discrepancies such as in Autism Spectrum Disorder diagnoses between genders, reflecting biases in clinical practice. Additionally, the study tested ChatGPT on rare mental illnesses like Cotard's and Capgras Syndromes, where it correctly diagnosed all cases. However, the study is limited by the small sample size and potential biases from using vignettes from research articles. Future research should focus on broader number of vignettes collected from various sources.

References

Bessiere, K., Pressman, S., Kiesler, S., & Kraut, R. (2010). Effects of internet use on health and depression: A longitudinal study. *Journal of Medical Internet Research, 12*(1), e6.

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal, 6*(2), 94-98.

Lupton, D. (2014). Critical perspectives on digital health technologies. *Sociology Compass, 8*(12), 1344-1359.

Semigran, H. L., Linder, J. A., Gidengil, C., & Mehrotra, A. (2015). Evaluation of symptom checkers for self-diagnosis and triage: Audit study. *BMJ, 351*, h3480.