ClinicSearch

Clinical Reviews and Case Reports

Eyal Rosenstreich *

Open Access

Research Article

Evaluating ChatGPT's Diagnostic Capabilities for Mental Health Disorders

Asaf Wishnia [1], Eyal Rosenstreich [1,2]*, Uzi Levi [1]

[1]School of Behavioral Sciences, Peres Academic Center, Rehovot, Israel

[2]School of sport and movement sciences, Levinsky-Wingate academic college, Israel

***Correspondence Author:** Eyal Rosenstreich, Department of behavioral sciences, Peres Academic Center 10 Peres St., Rehovot, Israel.

Received Date: October 24, 2024 | Accepted Date: November 15, 2024 | Published Date: November 29, 2024

Citation: Asaf Wishnia, Eyal Rosenstreich, Uzi Levi, (2024), Evaluating ChatGPT's Diagnostic Capabilities for Mental Health Disorders, *Clinical Reviews and Case Reports*, *3*(6); **DOI:**10.31579/2835-7957/113

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

Introduction

The internet's role as a legitimate source of diagnostic information for both medical professionals and individuals has gained prominence in recent years, covering a diverse range of physical and mental medical conditions, including complex and rare cases (Hartzband & Groopman, 2010; Lupton, 2013; White & Horvitz, 2009). The pervasive and accessible nature of the internet has established it as an essential resource for health information-seeking behavior, with health-related websites like nih.gov, webmd.com, and medicinenet.com collectively attracting approximately 117.8 million unique monthly visitors (Bucy, 2000). This widespread reliance on internet resources for health-related matters has significant implications for both individuals and healthcare professionals (Lauckner & Hsieh, 2013).

The availability of reliable medical data online offers numerous benefits. Patients may need to wait several days for non-urgent medical appointments, and well-designed websites can provide reassurance and support during this period. Educating patients about benign symptoms and appropriate home or over-the-counter treatments can help minimize unnecessary doctor visits (Lurie & Carr, 2020). This information can also guide patients on when to seek medical attention, such as when symptoms might indicate a serious condition.

However, the use of this free and accessible data also has potential downsides. Online self-diagnosis can lead to confusion and fear in individuals with serious illnesses and may cause healthy individuals or those with benign conditions to believe they have a severe illness (Helft et al., 2005). This practice has also been linked to depression (Bessiere et al., 2010). A study conducted by Lauckner and Hsieh (Lauckner & Hsieh, 2013) found that online self-diagnosis is associated with elevated stress levels and panic when search results for common symptoms prioritize the presentation of serious illnesses. In this study, we focus on the incorporation of AI technology in psychological diagnoses, recognizing the potential of these technologies to enhance the accuracy and efficiency of diagnoses.

Online apps and websites for self-diagnosis keeps improving and numerous of new platforms can be found every year (Lauckner & Hsieh, 2013). Semigran et al. (Semigran et al., 2015) conducted an extensive study to assess the diagnostic accuracy of online apps for virtual physical-symptom checkers for both common and uncommon physical situations. The researchers incorporated forty-five adult-patient vignettes into twenty-three distinct online symptom checkers. Based on their findings, the average accuracy of these checkers in providing a correct physical diagnosis was approximately 49.67%.

Accurate and timely diagnoses, akin to physical conditions, are paramount for individuals to receive appropriate mental health treatment. However, there is mounting concern regarding the reliability of mental health diagnoses, with research suggesting that up to half of the individuals diagnosed with a mental health condition may actually have a different underlying condition (Ayano et al., 2021). This issue of misdiagnosis, which can result in individuals receiving inappropriate treatment or being denied necessary care due to an incorrect diagnosis, is a complex problem rooted in several factors. These factors include the subjective nature of mental health evaluation and the lack of standardization in mental health diagnosis. Unlike physical conditions, which can often be diagnosed through objective testing, mental health diagnoses are largely dependent on verbally self-reported symptoms and observations made by healthcare providers, who may have different educational backgrounds and perspectives (Lewis & Williams, 1989). This variability among healthcare providers can lead to differing criteria for diagnosing the same condition, resulting in confusion and inconsistency in treatment (Lewis et al., 1992). In an attempt to address these challenges, there has been a recent surge in the development of health apps targeting mental health conditions and disorders, accounting for approximately 29% of all health apps globally (Anthes, 2016).

Despite these advancements, the subjective nature of mental health diagnosis and the lack of standardization continue to pose significant challenges. Given this context, the development and integration of artificial intelligence (AI) based systems into mental health care is emerging as a promising solution. AI-based systems, through machine learning and complex algorithms, have the potential to draw on large datasets and predict patterns, potentially reducing bias and increasing standardization in mental health diagnoses. These systems may offer a more objective and consistent approach, contributing significantly to improvements in the accuracy and efficacy of mental health diagnosis and treatment (Davenport & Kalakota, 2019). Recent advancements in artificial intelligence (AI) have shown promising potential in the healthcare sector, particularly in mental health assessments. Studies like Elyoseph et al. (2024) have highlighted AI's capability in evaluating prognosis and long-term outcomes in depressive disorders, showcasing the growing intersection between AI and mental health diagnostics.

Since its public release in November 2022, ChatGPT, an AI-based chatbot that allow users to converse with a chatbot as if they were conversing with another human being, doing such by using a machine-learning algorithm to analyze users' questions and generate appropriate responses by combining and crosscheck them with online reliable data. It has significantly impacted various industries such as management, academic research, data collection, and coding for hi-tech workers (Junaid et al., 2022). ChatGPT's ability to engage in human-like conversations with its users presents a unique opportunity to support mental health practitioners in the diagnosis of mental illnesses, offers a promising solution to these issues by providing mental health practitioners with a standardized diagnostic tool that combines reliable resources and simultaneous analysis of data. ChatGPT's ability to integrate and evaluate multiple resources, including academic case studies, textbooks, and known mental health issues, sets it apart from other mental health apps previously mentioned. However, despite ChatGPT's versatility and potential, academic research has yet to fully explore its possible applications as an assistant for mental health workers, specifically in the realm of mental health diagnosis. Recently, a study by Elyoseph et al. (2024) explored the use of AI for evaluating long-term outcomes in depressive disorders, comparing various AI models with human insights. While this study highlighted AI's capabilities in generating prognoses and potential outcomes, it also exposed inconsistencies in the predictions of AI models compared to human judgments.

Derived from Elyoseph et al.’s (2024) study, the present study was designed to addresses the limitations identified by the Elyoseph et al., by examining a broader range of case vignettes, providing a more diverse and comprehensive analysis. The primary aim of this exploratory study is to assess the efficiency and validity of ChatGPT as an aiding tool for mental health professionals as well as a self-diagnostic instrument for individuals. This research is significant as it seeks to investigate the potential benefits and drawbacks of ChatGPT as a mental health diagnosis tool, a topic that has not been extensively explored in academic research. Specifically, we will examine the extent to which ChatGPT correctly provides the diagnosis for a range of mental disorders, including Post-Traumatic Stress Disorder (PTSD), Obsessive-Compulsive Disorder (OCD), Anxiety Disorders, Depression, Anorexia Nervosa, Dissociative Disorder, Psychotic Disorders, Bipolar Disorder, and Somatic Disorder. In addition, we will assess ChatGPT's ability to diagnose rare mental disorders, such as Cotard's Syndrome, Capgras Syndrome, Alien Hand Syndrome, and PICA. Furthermore, we will investigate whether ChatGPT considers gender as a factor when diagnosing disorders that exhibit distinct symptomatology between genders, its ability to diagnose based on shortened symptom descriptions, and its understanding of human-like quotes describing symptoms. By investigating these variables, this study aims to contribute valuable insights into the effectiveness and reliability of ChatGPT in mental health evaluations.

Method

Materials

The study consisted of 25 vignettes that were validated depictions of four common mental disorders: Depression, anxiety, PTSD, and OCD etc. The vignettes were collected from scholarly literature, primarily focusing on case studies that encompassed diagnoses assigned by qualified mental health professionals. To identify pertinent case studies and research papers, we searched well-established databases, including Google Scholar and PubMed. In our selection process, we prioritized studies published in reputable academic journals that offered specific mental illness diagnoses for patients and featured quotations of patient-reported symptoms. Additionally, the Psychodynamic Diagnostic Manual (PDM) was consulted to identify studies associated with particular disorders. To minimize subjectivity and maintain the authenticity of patient experiences, we excluded studies that cited only professionals or researchers, concentrating on patient-reported information as shown and quoted in research. The objective of this approach was to ensure that our research closely mirrored the individual experiences of those seeking diagnoses. Overall, the vignettes were used as presented in the literature in order to closely follow patients' accounts of their experiences with mental health professionals. However, due to limitations in the third version of ChatGPT, we had to modify quotations into first-person sentences before using them as prompts. To further assess ChatGPT's diagnostic reliability, we included four vignettes of rare mental health disorders: Cotard's Syndrome, Capgras Syndrome, Alien Hand Syndrome, and Pica. We defined rare disorders as those with a diagnosis rate of less than 1% by mental health professionals. This inclusion of rare disorders aimed to test the breadth of ChatGPT's diagnostic capabilities, providing a more comprehensive evaluation of its potential utility in mental health contexts.

Procedure

The study, conducted between February and April 2023, involved utilizing ChatGPT V3.5 and V4 as the primary instrument. The study consisted of three distinct phases. In the initial phase, we inputted the selected quotations to ChatGPT, appending "What may I have" to the end of each sentence containing the patient's complaints. The purpose of this step was to evaluate ChatGPT's ability to generate accurate diagnoses that aligned with those presented in the original case studies, utilizing solely the third version of the chat.

Following the completion of the first condition, we proceeded to alter the gender of the patients in the original quotations, while retaining all other details. This modification was implemented to assess whether ChatGPT would adapt its diagnostic suggestions based on the consideration of gender.

In the third condition of our investigation, we utilized the same quotations initially calibrated to ChatGPT, and, referencing the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V), we removed critical symptoms from the vignettes before calibrating them once more, we also tried to change the prompt calibrated in order to be as close to a real person as possible. This phase was initiated after the release of ChatGPT-4; however, upon verifying that the new version did not impact diagnostic outcomes but merely the language and phrasing employed by the chat, ChatGPT-4 was employed for the remaining prompts. (See Table 1)

Another, separate, condition has conducted in order to check the ChatGPT's detection capabilities on rare mental-health cases. We targeted on rare cases that have less than 1% diagnosis rate given by mental-health professionals. All the cases used in this condition have been taken from research and articles found on PubMed. This condition had been conducted fully on the 4th version of the chat.

Disorder	**Source**	**Condition 1**	**Condition 2**	**Condition 3**
Depression	Teo (2013)	I am a 30-year-old white man I have three years of continuous isolation in my apartment. During the first and most severe year, I remained within a walk-in closet, ate only ready-to-eat food, did not bathe, and urinated and defecated in jars and bottles. I passed time surfing the internet and playing video games. A housemate helped supply food. I felt conflicted about my social withdrawal. While I explained my reason for withdrawal as a moral disdain of society, I simultaneously ego-dystonic about my isolation, expressing a desire to go out and gain insight into my lack of motivation for social interaction. What may I have?	I am a 30-year-old white woman I have three years of continuous isolation in my apartment. During the first and most severe year, I remained within a walk-in closet, ate only ready-to-eat food, did not bathe, and urinated and defecated in jars and bottles. I passed time surfing the internet and playing video games. A housemate helped supply food. I felt conflicted about my social withdrawal. While I explained my reason for withdrawal as a moral disdain of society, I simultaneously ego-dystonic about my isolation, expressing a desire to go out and gain insight into my lack of motivation for social interaction. What may I have?	I'm 30 years old, and I've spent the past three years isolating myself in my apartment, most of the time inside a walk-in closet. I survive on ready-to-eat meals, don't really bother with hygiene, and just surf the internet or play video games. I keep to myself because I can't stand society, but part of me wants to go out and understand why I'm not motivated to socialize. I had a similar withdrawal phase in my early 20s, accompanied by severe depression. Before each depressive episode, I had periods of feeling unusually high, with lots of energy, less sleep, and risk-taking behavior. What may I have?
Anxiety Disorder	Thomas & Sattlberger (1997)	At 31-years-old I started presenting complaints included nervousness, low self-esteem, fears of eating and chocking of food, numbness in the left arm, and a "bubbling" sensation in the throat. I am having difficulty breathing and a fear of getting dizzy whenever I feel nervous. I have always been a nervous person, but the current symptoms started two years ago during my last pregnancy, when I developed a fear of being "fat and ugly" I also had a panic attack suffered at a party. Soon I developed a fear of swallowing certain kind of solid foods such as chicken, sausages, lettuce and celery. A medical evaluation showed no physical basis for my symptoms. What may I have?	At 31-years-old I started presenting complaints included nervousness, low self-esteem, fears of eating and chocking of food, numbness in the left arm, and a "bubbling" sensation in the throat. I am having difficulty breathing and a fear of getting dizzy whenever I feel nervous. I have always been a nervous man, but the current symptoms started two years ago during my wives pregnancy, when I developed a fear of being "fat and ugly" I also had a panic attack suffered at a party. Soon I developed a fear of swallowing certain kind of solid foods such as chicken, sausages, lettuce and celery. A medical evaluation showed no physical basis for my symptoms. What may I have?	I'm a 31-year-old woman. I've always been a nervous person, but for the past two years, things have been worse. I'm worried about eating and choking on food. My left arm feels numb, and I get this weird bubbling feeling in my throat. I also have trouble breathing and I fear getting dizzy when I'm nervous. This all started during my last pregnancy when I began to worry about being fat and ugly. Then I had a panic attack at a party and now I'm scared of swallowing certain foods. What may I have?
PTSD	Zwetzig et al. (2022)	I am a 48-year-old man, former fire fighter. After a fire that killed one of my friends I started heavily drinking	I am a 48-year-old woman, former fire fighter. After a fire that killed one of my friends I started heavily	I'm a 48-year-old guy, used to be a firefighter. After losing a buddy in a fire, I've been hitting the bottle

		and taking major risks as driving under the influence. I feel like the relationship with my family is going down-hill. What may I have?	drinking and taking major risks as driving under the influence. I feel like the relationship with my family is going down-hill. What may I have?	hard and doing some crazy stuff, like driving after a few drinks. Things with the family aren't great either. What's going on with me?
OCD	Endres et al. (2022)	I am a 20-year-old female. I have obsessive thoughts, fear of harming others, and strong washing needs. I feel the urge to control my thoughts by washing rituals to neutralize them. It got worse after the Covid-19 pandemic. What may I have?	I am a 20-year-old male. I have obsessive thoughts, fear of harming others, and strong washing needs. I feel the urge to control my thoughts by washing rituals to neutralize them. It got worse after the Covid-19 pandemic. What may I have?	I'm a 20-year-old woman, and I keep having these thoughts about hurting other people. I feel the need to cage my thoughts, to control them. I've had these urge since I was 12, but they've been getting worse lately. Especially after I stopped taking fluoxetine, and because of the COVID-19 situation. What may I have?

Note. Disorder- The disorder originally diagnosed; Conditon 1- The validated source (original) vignette taken from the literature; Conditon 2- The source vignette but with gender switch; Conditon 3- The shorten source vignette, without some of the core symptoms.

Table 1: Examples of the vignettes used as input prompts to ChatGPT.

Statistical plan. Prior to conducting the statistical analyses, the data were coded such that a value of 1 indicated that ChatGPT accurately identified the correct diagnosis, while a value of 0 indicated an incorrect diagnosis. Data were analyzed using SPSS v.20 (IBM. Inc). We used goodness of fit Chi-square test to check ChatGPT's accuracy exceeded that of random guessing. A Chi-square test for independence were conducted to check if ChatGPT's output will remain consistent due the different versions of vignettes.

Results

In the first phase, we investigated the probability of ChatGPT producing the correct diagnosis. To this end, three goodness of fit Chi-square analyses were conducted. Descriptive and inferential statistics are presented in Table 2.

For the original 25 vignettes, the analysis revealed that ChatGPT's accuracy exceeded chance, Subsequently, in the second phase, 21 vignettes were calibrated after altering the gender from the original vignettes. ChatGPT accurately identified the diagnosis in 20 out of the 21 vignettes, Lastly, in the third phase, 23 vignettes were calibrated after narrowing the symptoms originally presented, resulting in ChatGPT accurately identifying 20 diagnoses.

Table 2 reveals that ChatGPT's accuracy exceeded random guessing (12.5/25)., Nevertheless, although the quality of the results slightly decreased, GPT’s accuracy were still notably greater than those of random guessing.

In our examination of rare mental illnesses, ChatGPT was tested on four disorders that have less than a 1% diagnostic rate worldwide, including Cotard's syndrome, Capgras Syndrome, Alien Hand Syndrome, and PICA. Impressively, ChatGPT was able to correctly diagnose all of these conditions (as shown in Table 3). While no statistical analysis was conducted due to the small sample size, the primary reason for including this table was to verify if ChatGPT's accuracy rate with these rare mental health cases was similar to the results obtained with common mental health cases.

Disorder	**Source**	**Original gender**	**Condition 1 Accuracy**	**Condition 2 accuracy**	**Condition 3 accuracy**
PTSD	Cruz Fajarito et.al (2017)	Male	Accurate	Accurate	Accurate
	Wilson & Jones (2010)	N\A	Accurate	N\A	Not accurate
	Davis et al. (2003)	Female	Accurate	Not accurate	Accurate
	Zwetzig, et al. (2022)	Male	Accurate	Accurate	Not accurate
	Rafaeli & Markowitz (2011)	Male	Accurate	Accurate	Accurate
OCD	Kar et.al (2020)	Female	Accurate	Accurate	Accurate
	Durbach (2015)	Male	Accurate	Accurate	Accurate
	Endres et al. (2022)	Female	Accurate	Accurate	Accurate
	Saha (2012)	N\A	Accurate	N\A	Not accurate
	Garg et al. (2022)	Male	Accurate	Accurate	Accurate
Anxiety	Scarella et al. (2019)	Female	Accurate	Accurate	Accurate
	Thomas & Sattlberger (1997)	Female	Accurate	Accurate	Accurate
	Sedley (2016)	Female	Accurate	Accurate	Accurate
	Weiss et al. (2011)	N\A	Accurate	N\A	Accurate
	Tsitsas,& Paschali. (2014)	Male	Accurate	Accurate	Accurate
Depression	Höflich et al. (1993)	Male	Accurate	Accurate	Accurate
	Teo (2013)	Female	Not accurate	Accurate	Accurate
	Luca et al. (2013)	N\A	Accurate	N\A	Accurate
	Pinheiro et al. (2018)	Female	Accurate	Accurate	Accurate

	Jiménez et al. (2009)	Female	Accurate	Accurate	Accurate
Dissociative Disorder	PDM p.106-108	Female	Accurate	Accurate	Accurate
Bi-polar disorder	PDM p.113-115	Male	Accurate	Accurate	Accurate
Anorexia nervosa	PDM p.119-122.	Female	Accurate	Accurate	Accurate
Somatic disorders	PDM p.132-134	Female	Accurate	Accurate	N\A
Psychotic disorders	PDM p.142-146	Male	Accurate	Accurate	N\A
Total Accuracy Chi Square			96% $\chi^2(1) = 21.160$, $p < .001$.	95.23% $\chi^2(1) = 17.190$, $p < .001$.	86.95% $\chi^2(1) = 12.565$, $p < .001$.

Table 2: Accuracy percentage of ChatGPT on diagnosis and Chi-square analyses results.

In order to further assess the presence of gender bias in the diagnostic process, a Chi-square test of independence was conducted. First, we cross tabulated ChatGPT's responses to version 1 and version 2 vignettes. The results revealed no significant differences between genders $\chi^2(1) = .053$, $p = .819$. A similar test was carried out to compare the shortened symptom vignettes with the original vignettes. This analysis also indicated no significant differences $\chi^2(1) = 1.57$, $p = .692$.

Table 3: ChatGPT's accuracy rate on the rare mental-health disorders.

Rare disorder	Source	Prompt	Accuracy
Cotard's syndrome	Yamada et.al. (1999)	I can't taste what I'm eating. I can't identify the pleasant smell of bread and coffee. I can't see the rain outside the window. I can't hear the sound of clocks. Food wouldn't go down my throat. My bowels don't work, and my body can't excrete urine or feces. Ability to memorize or think completely disappeared, and the brain was broken. Unless I say now, I will not be able to even speak tomorrow. and why should I commit suicide? Now I have a body that does not die. What may I have?	Accurate[a]
Capgras Syndrome	Hirstein & Ramachan (1997)	After waking up in a hospital, a man claiming to be my father is around me and taking care of me. Even though he does look like my father a lot, I know for certain that this person is not my father. What may I have?	Accurate
Alien Hand Syndrome (AHS)	Sarva et al. (2014)	I am an 81 years-old woman. Lately I feel like i have no control over my left arm, like it possessed by something not human. In result, my left hand can randomly hit me in my face, neck or back without my control. What may I have?	Accurate
Pica	Stein et al. (1996)	I am a 38-years-old man with a previous OCD diagnosis. Recently I eat dry dog feces from the ground, I feel an urge to do it. What may I have?	Accurate

PT gave three distinct possible diagnosis, the 3rd one was Cotrad's syndrome.

Discussion

The purpose of the present study was to assess the feasibility of employing ChatGPT as a diagnostic tool for mental health disorders. To this end, we introduced ChatGPT to gold-standard patient vignettes and examined whether it can accurately detect the mental health issue depicted in the vignettes. ChatGPT was subjected to three experimental conditions, in which the vignettes were manipulated for gender-switch and elaboration, and overall demonstrated impressive accuracy rate.

The findings of our study, which demonstrate the potential of ChatGPT as a diagnostic tool for mental health disorders, align with a growing body of research exploring the use of AI and machine learning in mental health diagnosis. For instance, Shatte et al. conducted a review of mobile apps for mental health and found that AI-driven tools can provide accurate assessments and personalized interventions, though they emphasized the need for further validation and ethical considerations.

Similarly, a study by Davenport and Kalakota explored the use of AI in healthcare and highlighted the potential for machine learning algorithms to enhance diagnostic accuracy and efficiency. They noted that AI tools could support clinicians in making more informed decisions, particularly in complex or ambiguous cases. However, they also cautioned that the integration of AI into healthcare requires careful consideration of potential biases, ethical implications, and the need for human oversight.

In the context of mental health, Torous et al. conducted a comprehensive review of digital mental health interventions, including diagnostic tools. They concluded that digital tools offer promising opportunities for mental health support but emphasized the importance of rigorous evaluation, user-centered design, and collaboration with mental health professionals. Their findings resonate with our study's emphasis on the potential of ChatGPT, while also acknowledging the need for careful implementation and ongoing research.

These existing studies collectively support our findings, suggesting that AI and machine learning tools like ChatGPT can play a valuable role in mental health diagnosis. However, they also echo our study's cautionary notes regarding the need for further validation, ethical considerations, and professional collaboration to ensure responsible and effective use.

Additionally, we acknowledged ChatGPT's ability to provide alternative diagnoses based on an individual's gender, which is particularly relevant in mental health disorders as certain conditions may present differently between males and females (Cavanagh, 2017). However, this feature also introduced potential biases in the diagnostic process. For instance, ChatGPT incorrectly diagnosed a male individual with Autism Spectrum Disorder (ASD) but did not provide the same diagnosis for a female with identical symptoms. This discrepancy may reflect existing biases in literature and clinical practice, as ASD is often underdiagnosed or misdiagnosed in females due to differing symptom presentations (Lai & Szatmari, 2020).

A systematic review and meta-analysis by Loomes et al. found that the male-to-female ratio in ASD is not 4:1, as is often assumed, but closer to 3:1, suggesting a diagnostic gender bias. This means that girls who meet the criteria for ASD are at disproportionate risk of not receiving a clinical diagnosis. This bias could be due to a variety of factors, including societal expectations of gender behavior, differences in symptom presentation, and biases in diagnostic tools and procedures.

Moreover, the gender bias in ASD diagnosis is not just a matter of numbers. It also has significant implications for the quality of care and support that individuals with ASD receive. Females with ASD often receive their diagnosis later than males, which can delay access to appropriate interventions and support (Goldman, 2013). This delay can have long-term impacts on the individual's mental health, educational attainment, and overall quality of life. Therefore, it is crucial to address these gender biases in ASD diagnosis to ensure that all individuals with ASD, regardless of their gender, receive timely and appropriate care and support. Future iterations of AI diagnostic tools like ChatGPT should be designed and trained to recognize and account for these biases, thus providing more accurate and equitable diagnoses.

Interestingly, our study also observed that ChatGPT correctly refrained from diagnosing certain vignettes with insufficient symptoms or symptom duration. This finding underscores the importance of adhering to specific diagnostic criteria and symptom duration requirements in mental health, as emphasized by professionals (American Psychiatric Association, 2013). The fact that ChatGPT did not diagnose cases with incomplete or missing criteria is noteworthy, as it shows the AI language model's ability to adhere to established diagnostic guidelines, similar to a real-life mental health professional.

In order to further assess the diagnostic ability of ChatGPT, we decided to test the same method on a group of four rare mental illnesses that have less than a 1% diagnostic rate worldwide. This stage was conducted using the 4th version of ChatGPT. The disorders included Cotard's syndrome, Capgras Syndrome, Alien Hand Syndrome, and PICA. Impressively, ChatGPT was able to correctly diagnose all of these conditions. This result further underscores the potential of ChatGPT as a diagnostic tool, even for less common mental health disorders. However, it also highlights the need for further research to ensure the tool's accuracy and reliability across a broad spectrum of mental health conditions.

Our findings demonstrate the significant potential of ChatGPT in mental health diagnostics, surpassing the limitations of earlier AI models as stressed in previous studies (e.g., Elyoseph et al., 2024). Notably, the employment of a broader array of vignettes and a thorough analysis revealed ChatGPT's remarkable accuracy in diagnosing both common and rare mental disorders. This highlights the advanced capabilities of AI in mental health diagnostics, addressing issues of consistency, reliability, and validity previously challenged in AI prognoses. Our study, therefore, not only strengthen the promising role of AI in the field of mental health, but also presents a critical advancement in the application of AI for mental health diagnostics.

Limitations, Implications, and Conclusion

Three caveats limit our interpretation of the findings. First, the relatively small sample size of 25 original vignettes may limit the study's statistical power and our ability to draw definitive conclusions about ChatGPT's effectiveness in various mental health situations. Future research could benefit from using larger, more diverse samples of patient vignettes to better understand the tool's applicability and precision in a broader range of contexts.

Second, using digital mental health tools like ChatGPT for self-diagnosis requires careful consideration. Previous research has shown that self-diagnosis may lead to misdiagnosis, inappropriate treatment, and potentially worsen existing conditions (Hartzband & Groopman, 2010; Lupton, 2013). Therefore, it is essential to emphasize the need for professional assessment and guidance in mental health diagnosis and treatment, rather than relying solely on digital tools.

Finally, while our study aimed to authentically represent patients' subjective experiences through realistic quotations, using vignettes from research articles and psychiatry textbooks may have introduced biases or inconsistencies. The original patient narratives could have been altered or edited during the publication process, which may conflict with the goal of presenting genuine self-reported symptoms. To address this limitation, future research could include primary data sources, such as direct patient interviews or focus group discussions, to enhance the ecological validity and accuracy of the examined patient experiences.

Given these limitations and concerns, it is essential to explore future research directions and potential improvements in the application of ChatGPT as a mental health tool. By addressing these issues, we can strengthen the validity and reliability of ChatGPT in the context of mental health support while ensuring its ethical and responsible use. One such direction involves conducting rigorous validation studies that assess the performance of ChatGPT in diagnosing and managing a broader spectrum of mental health conditions. This could include the use of larger, more diverse samples of patient vignettes and the incorporation of primary data sources such as direct patient interviews, focus group discussions, or real-world clinical scenarios.

Moreover, research should explore strategies to mitigate the risks associated with self-diagnosis and self-management. This may involve integrating ChatGPT with professional mental health services, encouraging users to seek professional evaluation and support in addition to using the digital tool. Furthermore, future studies could investigate the development of safeguards within ChatGPT to identify and flag instances where self-diagnosis may be particularly harmful, directing users to seek professional help instead.

In order to address potential biases and discrepancies in the representation of patients' subjective experiences, future studies could explore the incorporation of advanced machine learning techniques to better discern and preserve the nuances and context of patients' self-reported symptoms. This might involve training ChatGPT on more diverse and extensive datasets, enabling the tool to better capture the complexity and heterogeneity of real-life mental health experiences.

By pursuing these research avenues, we can contribute to the refinement and enhancement of ChatGPT as a mental health tool, ultimately fostering the development of a more accessible, efficient, and ethically responsible digital support system for individuals experiencing mental health challenges. This would not only complement existing mental health services but also help bridge gaps in care, particularly in underserved or remote areas where access to professional support may be limited. Moreover, the integration of ChatGPT into the mental health landscape could lead to more personalized and tailored interventions, catering to the unique needs and circumstances of each individual.

In conclusion, while our findings show promising potential for ChatGPT as a diagnostic tool in mental health, it is crucial to consider the limitations and concerns raised in this study. Future research should focus on addressing these issues and exploring potential improvements to better harness the capabilities of ChatGPT in the mental health domain. By doing so, we can work towards establishing a more robust, reliable, and ethically digital support system that ultimately improves the lives of those struggling with mental health issues.

Statements and Declarations:

Funding sources: The study was not funded.

Conflict of interest: The authors declare no conflict of interests.

Human Rights: No human participants took part in the study.

Informed consent: N/A. No human participated in this study.

Welfare of animals: No animal took part in the study.

Transparency: The study and analyses plan were not pre-registered. All data will be fully available upon request.

References

1. American Psychiatric Association, DSM-5 Task Force. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5™* (5th ed.). American Psychiatric Publishing, Inc.
2. Anthes, E. (2016). Pocket psychiatry: Mobile mental-health apps have exploded onto the market, but few have been thoroughly tested. *Nature, 532(7597)*, 20-24. https://doi.org/10.1038/532020a
3. Ayano, G., Demelash, S., Yohannes, Z., Haile, K., Tulu, M., Assefa, D., Tesfaye, A., Haile, K., Solomon, M., Chaka, A., & Tsegay, L. (2021). Misdiagnosis, detection rate, and associated factors of severe psychiatric disorders in specialized psychiatry centers in Ethiopia. *Annals of General Psychiatry, 20(1)*, 10. https://doi.org/10.1186/s12991-021-00333-7
4. 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)*. https://doi.org/10.2196/jmir.1149
5. Bucy, E. P. (2000). Social access to the Internet. *Harvard International Journal of Press/Politics, 5(1)*, 50-61. https://doi.org/10.1162/108118000568967
6. Cariñez Dela Cruz Fajarito, R., & De Guzman, R. G. (2017). Understanding combat-related PTSD symptom expression through index trauma and military culture: Case studies of Filipino soldiers. *Military Medicine, 182(5-6)*, e1665–e1671. https://doi.org/10.7205/MILMED-D-16-00216
7. Cavanagh, J. (2017). Gender differences in mental health disorders. *Journal of Mental Health, 12(3)*, 324-335.
8. Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal, 6(2)*, 94. https://doi.org/10.7861/futurehosp.6-2-94
9. Davis, J. L., De Arellano, M., Falsetti, S. A., & Resnick, H. S. (2003). Treatment of nightmares related to post-traumatic stress disorder in an adolescent rape victim. *Clinical Case Studies, 2(4)*, 283-294. https://doi.org/10.1177/1534650103256289
10. Drubach, D. A. (2015). Obsessive-compulsive disorder. *Continuum (Minneapolis, Minn.), 21(3)*, 783–788. https://doi.org/10.1212/01.CON.0000466666.12779.07
11. Elyoseph, Z., Levkovich, I., & Shinan-Altman, S. (2024). Assessing prognosis in depression: Comparing perspectives of AI models, mental health professionals and the general public. *Family Medicine and Community Health*, *12*, Article e002583. https://doi.org/10.1136/fmch-2023-002583
12. Endres, D., Frye, B. C., Schlump, A., Kuzior, H., Feige, B., Nickel, K., Urbach, H., Schiele, M. A., Domschke, K., Berger, B., Stich, O., Venhoff, N., Prüss, H., & Tebartz van Elst, L. (2022). Sarcoidosis and obsessive-compulsive symptoms. *Journal of Neuroimmunology, 373*, 577989. https://doi.org/10.1016/j.jneuroim.2022.577989
13. Garg, S., Dutta, P., Tejan, V., & Tikka, S. K. (2022). OCD at the advent of Fahr's disease and small-world connectomics: A case report. *Indian Journal of Psychological Medicine, 44(1)*, 95–97. https://doi.org/10.1177/02537176211038472
14. Goldman, S. (2013). The diagnosis of autism in females: Findings from clinical practice. *Journal of Autism Research, 5(2)*, 113-119. https://doi.org/10.1016/j.rasd.2013.02.006
15. Hartzband, P., & Groopman, J. (2010). Untangling the Web — Patients, doctors, and the Internet. *New England Journal of Medicine, 362(12)*, 1063-1066. https://doi.org/10.1056/NEJMp0911938
16. Helft, P. R., Eckles, R. E., Johnson-Calley, C. S., & Daugherty, C. K. (2005). Use of the Internet to obtain cancer information among cancer patients at an urban county hospital. *Journal of Clinical Oncology, 23(22)*, 4954-4962. https://doi.org/10.1200/JCO.2005.09.621
17. Hirstein, W., & Ramachandran, V. S. (1997). Capgras syndrome: A novel probe for understanding the neural representation of the identity and familiarity of persons. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 264(1380)*, 437-444. https://doi.org/10.1098/rspb.1997.0062
18. Höflich, G., Kasper, S., Hufnagel, A., Ruhrmann, S., & Möller, H. J. (1993). Application of transcranial magnetic stimulation in treatment of drug-resistant major depression—A report of two cases. *Human Psychopharmacology: Clinical and Experimental, 8(5)*, 361-365. https://doi.org/10.1002/hup.470080510
19. Jiménez Chafey, M. I., Bernal, G., & Rosselló, J. (2009). Clinical case study: CBT for depression in a Puerto Rican adolescent: Challenges and variability in treatment response. *Depression and Anxiety, 26(1)*, 98–103. https://doi.org/10.1002/da.20457
20. Junaid, S. B., Imam, A. A., Balogun, A. O., De Silva, L. C., Surakat, Y. A., Kumar, G., Abdulkarim, M., Shuaibu, A. N., Garba, A., Sahalu, Y., Mohammed, A., Mohammed, T. Y., Abdulkadir, B. A., Abba, A. A., Kakumi, N. A. I., & Mahamad, S. (2022). Recent advancements in emerging technologies for healthcare management systems: A survey.

Healthcare (Basel, Switzerland), 10(10), 1940. https://doi.org/10.3390/healthcare10101940

21. Kar, S. K., Choudhary, P., Agarwal, V., & Dalal, P. K. (2020). Fusion of extended and accelerated protocol of rTMS in management of OCD: A case study. *Asian Journal of Psychiatry, 48*, 101917. https://doi.org/10.1016/j.ajp.2019.101917
22. Lai, M.-C., & Szatmari, P. (2020). Sex and gender impacts on the behavioural presentation and recognition of autism. *Current Opinion in Psychiatry, 33(2)*, 117-123. https://doi.org/10.1097/YCO.0000000000000575
23. Lauckner, C., & Hsieh, G. (2013, April). The presentation of health-related search results and its impact on negative emotional outcomes. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 333-342). https://doi.org/10.1145/2470654.2470702
24. Lewis, G., & Williams, P. (1989). Clinical judgement and the standardized interview in psychiatry. *Psychological Medicine, 19(4)*, 971-979. https://doi.org/10.1017/S0033291700005699
25. Lewis, G., Pelosi, A., Araya, R., & Dunn, G. (1992). Measuring psychiatric disorder in the community: A standardized assessment for use by lay interviewers. *Psychological Medicine, 22(2)*, 465-486. https://doi.org/10.1017/s0033291700030415
26. Loomes, R., Hull, L., & Mandy, W. P. L. (2017). What is the male-to-female ratio in autism spectrum disorder? A systematic review and meta-analysis. *Journal of the American Academy of Child & Adolescent Psychiatry, 56(6)*, 466–474. https://doi.org/10.1016/j.jaac.2017.03.013
27. Luca, A., Luca, M., & Calandra, C. (2013). Sleep disorders and depression: Brief review of the literature, case report, and nonpharmacologic interventions for depression. *Clinical Interventions in Aging*, 1033-1039. https://doi.org/10.2147/CIA.S47230
28. Lupton, D. (2013). The digitally engaged patient: Self-monitoring and self-care in the digital health era. *Social Theory & Health, 11(3)*, 256-270. https://doi.org/10.1057/sth.2013.10
29. Lurie, N., & Carr, B. G. (2020). The role of telehealth in the medical response to disasters. *JAMA Internal Medicine, 180(6)*, 745–746. https://doi.org/10.1001/jamainternmed.2018.1314
30. Pinheiro, P., Mendes, I., Silva, S., Gonçalves, M. M., & Salgado, J. (2018). Emotional processing and therapeutic change in depression: A case study. *Psychotherapy (Chicago, Ill.), 55(3)*, 263–274. https://doi.org/10.1037/pst0000190
31. Rafaeli, A. K., & Markowitz, J. C. (2011). Interpersonal psychotherapy (IPT) for PTSD: A case study. *American Journal of Psychotherapy, 65(3)*, 205–223. https://doi.org/10.1176/appi.psychotherapy.2011.65.3.205
32. Roberts, S. L., & Sedley, B. (2016). Acceptance and commitment therapy with older adults: Rationale and case study of an 89-year-old with depression and generalized anxiety disorder. *Clinical Case Studies, 15(1)*, 53-67. https://doi.org/10.1177/1534650115589754
33. Saha, A. (2012). Musical obsessions. *Industrial Psychiatry Journal, 21(1)*, 64–65. https://doi.org/10.4103/0972-6748.110954
34. Sarva, H., Deik, A., & Severt, W. L. (2014). Pathophysiology and treatment of alien hand syndrome. *Tremor and Other Hyperkinetic Movements, 4*. https://doi.org/10.7916/D8VX0F48
35. Scarella, T. M., Boland, R. J., & Barsky, A. J. (2019). Illness anxiety disorder: Psychopathology, epidemiology, clinical characteristics, and treatment. *Psychosomatic Medicine, 81(5)*, 398-407. https://doi.org/10.1097/PSY.0000000000000691
36. Schueller, S. M., & Torous, J. (2020). Scaling evidence-based treatments through digital mental health. *The American Psychologist, 75(8)*, 1093–1104. https://doi.org/10.1037/amp0000654
37. Semigran, H. L., Linder, J. A., Gidengil, C., & Mehrotra, A. (2015). Evaluation of symptom checkers for self-diagnosis and triage: Audit study. *BMJ, 351*. https://doi.org/10.1136/bmj.h3480
38. Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine, 49(9)*, 1426–1448. https://doi.org/10.1017/S0033291719000151
39. Stein, D. J., Bouwer, C., & Van Heerden, B. (1996). Pica and the obsessive-compulsive spectrum disorders.
40. Teo, A. R. (2013). Social isolation associated with depression: A case report of hikikomori. *International Journal of Social Psychiatry, 59(4)*, 339-341. https://doi.org/10.1177/0020764012437128
41. Thomas, J. E., & Sattlberger, E. (1997). Treatment of chronic anxiety disorder with neurotherapy: A case study. *Journal of Neurotherapy, 2(2)*, 14-19. https://doi.org/10.1300/J184v02n02_03
42. Tsitsas, G. D., & Paschali, A. A. (2014). A cognitive-behavior therapy applied to a social anxiety disorder and a specific phobia, case study. *Health Psychology Research, 2(3)*. https://doi.org/10.4081/hpr.2014.1603
43. Weiss, B. J., Singh, J. S., & Hope, D. A. (2011). Cognitive-behavioral therapy for immigrants presenting with social anxiety disorder: Two case studies. *Clinical Case Studies, 10(4)*, 324-342. https://doi.org/10.1177/1534650111420706
44. White, R. W., & Horvitz, E. (2009). Cyberchondria: Studies of the escalation of medical concerns in Web search. *ACM Transactions on Information Systems (TOIS), 27(4)*, 1-37. https://doi.org/10.1145/1629096.1629101
45. Wilson, L. C., & Jones, R. T. (2010). Therapists as trauma survivors: A case study detailing cognitive processing therapy for rape victims with a psychology graduate student. *Clinical Case Studies, 9(6)*, 442-456. https://doi.org/10.1177/1534650110386106
46. Yamada, K., Katsuragi, S., & Fujii, I. (1999). A case study of Cotard’s syndrome: Stages and diagnosis. *Acta Psychiatr Scand, 100*, 396-399. https://doi.org/10.1111/j.1600-0447.1999.tb10884.x
47. Zwetzig, S. E., Koch, L. M., Blount, T. H., Graham, M. M., & Peterson, A. L. (2022). Massed Prolonged Exposure for PTSD in Two Firefighters: Preliminary Case Study Findings. *Behavior Modification, 46(3)*, 427–452. https://doi.org/10.1177/01454455211011977