Article

Not peer-reviewed version

Harnessing Large Language Models for Identification and Treatment of Obsessive-Compulsive Disorder

Inbar Levkovich *

Posted Date: 13 June 2024

doi: 10.20944/preprints202406.0857.v1

Keywords: Artificial Intelligence; Chat GPT; large language models; obsessive-compulsive disorder; vignette study

 doi:10.20944/preprints202406.0857.v1

Article

Harnessing Large Language Models for Identification and Treatment of Obsessive-Compulsive Disorder

Inbar Levkovich

Tel-Hai Academic College, Israel; levkovinb@telhai.ac.il

Abstract: Obsessive-Compulsive Disorder (OCD) is a mental health condition marked by recurrent intrusive thoughts or sensations that compel individuals to perform repetitive behaviors or mental acts. Obsessions and compulsions significantly disrupt daily life and cause considerable distress. Early identification and intervention improve long-term outcomes. This study aimed to evaluate the ability of four advanced artificial intelligence models (ChatGPT-3.5, ChatGPT-4, Claude, and Bard) to accurately recognize OCD compared to human professionals and to assess recommended therapies and stigma attributions. This study was conducted during March 2024 utilizing 12 vignettes. Each vignette depicted a client, either a young adult or a middle-aged male or female, attending an initial therapy session. Each vignette was evaluated ten times, resulting in 480 evaluations. The results were compared with those of a human sample of 514 psychotherapists, as reported by Canavan. Significant differences were found. AI models demonstrated higher OCD recognition rates and confidence levels than human professionals and showed 100% confidence in recognition, compared to 87% among psychotherapists. AI models also recommended evidence-based interventions more frequently, with ChatGPT-3.5 and Claude at 100%, ChatGPT-4 at 90%, and Bard at 60%, compared to 61.9% among psychotherapists. Additionally, AI models exhibited significantly lower stigma and danger estimations, though both AI and psychotherapists demonstrated high willingness to treat the described cases. The findings suggest that AI models surpass human professionals in recognizing OCD and recommending evidence-based treatments while also demonstrating lower stigma. These results highlight the potential of AI tools to enhance OCD diagnosis and treatment in clinical settings.

Keywords: Artificial Intelligence; Chat GPT; large language models; obsessive-compulsive disorder; vignette study

1. Introduction

Large-scale language models (LLMs), a subset of Natural Language Processing (NLP), are trained with extensive textual data to generate advanced language predictions [1]. The AI models considered in this study included ChatGPT-3.5, ChatGPT-4, Claude, and Bard, representing varying levels of linguistic and cognitive capability [2,3]. The use of artificial intelligence has been tested for many aspects of health [4], yet some mental disorders, and particularly obsessive-compulsive disorder (OCD), have received little research attention. Studies in the field of mental health indicate that AI models can optimize assistance and diagnostic processes, shorten the time required for administrative tasks, improve patient availability, and reduce stigma [5–8]. Despite these advantages, issues concerning ethics, privacy, and cultural bias still present difficulties [9–11]. Moreover, the results of these studies are inconsistent, often because of differing methodologies [12].

OCD is a prevalent mental health condition that significantly affects individual well-being [13–15]. OCD is characterized by repetitive intrusive thoughts (obsessions) and/or repetitive behaviors or by mental acts (compulsions) that individuals feel compelled to perform to alleviate their distress [16,17]. It has been widely recognized as a chronic condition [18,19]. Early identification and intervention can lead to better long-term outcomes [20]. Yet there is often a significant delay between symptom onset and problem recognition, partly because individuals with OCD are often reluctant to disclose their thoughts and behaviors [21] and their OCD frequently remains unrecognized even when disclosed [15]. Patients with OCD often experience substantial delays in receiving help due to

 doi:10.20944/preprints202406.0857.v1

stigma and a lack of knowledge about appropriate treatments [22–24]. Additionally, the scarcity of services for early detection and intervention exacerbates these delays, hindering timely access to effective treatment [21,25].

Recognizing OCD can be challenging due to the diversity of symptom presentations [26]. OCD generally manifests in four key subtypes: contamination, symmetry/incompleteness, responsibility for harm, and taboo intrusive thoughts [27,28]. Of these, taboo intrusive thoughts such as aggressive/violent, religious, or sexual obsessions are more difficult to recognize and have fewer treatment outcomes, often leading to misdiagnosis [29]. Thoughts of this type tend to conflict with an individual's self-concept and are reported as images, thoughts, doubts, and impulses [29].

Several studies have used AI to diagnose and track OCD using ordinary voice and image data. In a study involving adolescents with OCD and healthy controls, researchers analyzed speech patterns to investigate the relationship between OCD severity and specific vocal features [30]. Additionally, a study involving children and adolescents with OCD explored the feasibility and acceptability of using wearable biosensors, specifically wristbands, for monitoring OCD symptoms [31]. Neuroimaging, which is known for its standardization and quantifiability, has attracted increasing attention from researchers and clinicians due to its role in diagnosing and treating mental disorders [32]. Furthermore, machine-learning algorithms have demonstrated partial ability to predict the long-term course of OCD using clinical and cognitive information, thereby optimizing treatment options [33].

The Current Study

Despite the high prevalence of obsessive-compulsive behaviors, with rates ranging from 1.9 to 3.3 cases per 100 individuals in the United States [14], many factors prevent timely detection and treatment. Research indicates that primary care and mental health professionals are less likely to identify taboo intrusive thoughts as OCD than to recognize other OCD subtypes [34–37]. Misinterpretation of these thoughts can increase an individual's fear, anxiety, and depression [29]. Moreover, the stigmas held by mental health professionals and the treatment they provide are not based on evidence [38].

Considering recent advancements in artificial intelligence and its widespread availability, we sought to examine the ways in which different AI models can address the complexity of obsessive-compulsive behaviors compared to assessments by a human sample and across different languages. The aims of this study were twofold: to evaluate the ability of AI tools versus human professionals in accurately recognizing OCD, and to assess recommended therapies and stigma attributions.

Research Questions

This study examined the following research questions:

RQ1: How do various AI tools (ChatGPT-3.5, ChatGPT-4, Claude.AI, and Gemini) compare to human professionals (psychotherapists) in accurately recognizing OCD?

RQ2: What are the recommended therapies for individuals diagnosed with OCD according to the various AI tools (ChatGPT-3.5, ChatGPT-4, Claude.AI, and Gemini) compared to the recommendations of human professionals (psychotherapists)?

RQ3: How do various AI tools (ChatGPT-3.5, ChatGPT-4, Claude.AI, and Gemini) compare to human professionals (psychotherapists) in their assessments of stigma attributions and treatment willingness for OCD?

2. Materials and Methods

LLMs procedure

During March 2024, we evaluated the ability of LLMs—specifically ChatGPT-3.5 and 4 (by OpenAI; 3), Claude.AI (by Anthropic), and Gemini (by Google)—to identify OCD, make treatment recommendations, and assess stigma. We then compared these LLM evaluations with results from 514 psychotherapists in Ireland, as reported by Canavan. [27]. The same vignettes were also used by Glazier et al. [35,36] in their articles and underwent extensive validation. Five of the vignettes were translated into Spanish and distributed among healthcare providers in Latin America [37].

 doi:10.20944/preprints202406.0857.v1

Input source: Vignettes

To investigate the extent to which LLMs are sensitive to OCD, we used a series of vignettes developed by Canavan. [27]. The series included six different types of vignettes: four vignettes covering intrusive thoughts that were deemed taboo (pedophilia, homosexual, aggressive, and religious), as well as vignettes that covered contamination and symmetry. To assess the influence of gender on recognition rates, we created separate versions of each vignette for each gender, yielding a total of 12 vignettes (six vignette types × male/female). The demographic information of each client, excluding gender, was kept identical across all vignettes to minimize potential content bias [36]. Each of the vignettes was evaluated ten times in a new tab in ChatGPT-3.5 and 4, Claude.AI, and Gemini, yielding a total of 480 evaluations. Each vignette depicted a client, either a young adult or a middle-aged man or woman, at an initial therapy session. Unless otherwise stated, the clients have been experiencing symptoms for five years. These symptoms are time-consuming and cause significant anxiety and distress. The vignettes were adjusted to achieve a consistent length of 125-155 words each. The participating human professionals were given the following instructions: "You will be shown a scenario (vignette) describing a client with distressing symptoms. You will then be asked to answer questions related to your client. The vignettes are described in the following sections. The client (John/Lorraine) comes to an initial therapy session at your practice and describes the following symptoms. Unless otherwise stated, the symptoms have been present for five years, are time consuming and are causing the client significant anxiety and distress".

Vignette A+B: Sexual obsessions about children

John/Lorraine, a middle-aged man/woman, loved spending time with his/her nieces and nephews. But then he/she began seeing images that involved touching the children in a sexual manner. He/she had no desire to touch the children and did not experience any sexual arousal while seeing the images, but the worry of "what if" remained. He/she now tries to avoid being with the children and refuses to spend time alone with them. He/she knows that the thoughts come from within his/her own mind and are excessive in nature. Yet despite this knowledge, he/she remains upset by these thoughts and is not able to stop them.

Vignette C+D: Aggressive obsessions

John/Lorraine, a middle-aged man/woman, thought about pushing a woman standing next to him/her onto the railway tracks. He/she was afraid of this thought and feared he/she might act on it, so he/she immediately left the train station and caught a taxi home. Nevertheless, John/Lorraine remained worried and found himself/herself frequently visualizing the situation to make sure he/she did not actually harm the woman. John/Lorraine frequently finds himself/herself worrying that he/she may want to, or will, harm others and these thoughts greatly upset him/her. He/she knows that his/her thoughts come from within his/her own mind and are excessive in nature. Yet despite knowing this, he/she remains upset by the thoughts and is unable to stop them.

Vignette E+F: Religious obsessions

John/Lorraine, a middle-aged, highly religious man/woman, believes that one should not have any negative thoughts about religion. He/she becomes very upset upon noticing himself/herself having such negative religious thoughts (e.g., why does God allow bad things to happen to good people?). When these "bad" thoughts occur, as they frequently do, he/she becomes distressed and fears God will punish him/her. John/Lorraine then prays repeatedly to himself/herself until he/she feels safe from harm. This can go on for hours. He/she knows that these thoughts come from within his/her own mind and are excessive in nature. Yet despite knowing this, he/she remains upset by the thoughts and is not able to stop them.

Vignette G+H: Contamination obsessions

John/Lorraine, a middle-aged man/woman, constantly worries about dirt and germs. He/she is unable to complete many of his/her daily activities because he/she tries at all costs to avoid touching things that he/she thinks may be dirty. If John/Lorraine does touch a "dirty" object, he/she will immediately wash his/her hands to avoid contracting a disease. He/she knows that these thoughts are excessive in nature and come from within his/her own mind. Yet despite knowing this, he/she remains upset by the thoughts and is not able to stop them.

 doi:10.20944/preprints202406.0857.v1

Vignette I+ J: Symmetry obsessions

John/Lorraine, a middle-aged man/woman, worries when things are not in order or systematic. He/she becomes anxious when individuals move his/her belongings and feels he/she must immediately return the objects to their proper place. He/she also rearranges things that are not in order by placing them how they "should be". When things are not in proper order, John/Lorraine is unable to focus until the objects are returned to their correct place. He/she knows that these thoughts come from within his/her own mind and are excessive in nature. Yet despite knowing this, he/she remains upset by the thoughts and is not able to stop them.

Vignette K+ L: Homosexual obsessions

John, a young adult, has been in a committed relationship with his girlfriend for over five years. He loves her very much and is attracted to her. Although he is not sexually attracted to men, John is preoccupied by thoughts that he may be gay and worries that he is not living an honest life. Upon seeing men, John immediately assesses his body for any signs of sexual arousal and when he does not find any such signs, he experiences temporary relief. He knows that his thoughts come from within his own mind and are excessive in nature. Yet despite knowing this, he remains upset by the thoughts and is not able to stop them.

Measures

The large language models (LLMs) were asked to evaluate whether any of 21 specified mental health problems applied to the client in their vignette. The order of the mental health problems was randomized to avoid response-order biases. The LLMs were also asked to rank their confidence in answering this question on a 5-point Likert scale, ranging from "not at all confident" to "very confident."

The participating LLMs were then given a list of 20 options and asked to select the type of therapy they felt would be most beneficial as the first-line therapy for the client. They were allowed to select up to three therapy types. Again, the order of the therapies was randomized. The LLMs ranked their confidence in this selection on a 5-point Likert scale, ranging from "not at all confident" to "very confident."

Next, the participating LLMs were asked whether they believed the client in their vignette could harm another person. They answered on a 7-point Likert scale ranging from "very likely" to "highly unlikely." Finally, respondents indicated their willingness to work with the client on a 7-point Likert scale ranging from "very unlikely" to "highly likely." These questions were designed to assess two common forms of stigma related to mental health problems, as outlined by Brown (2008): (1) fear and dangerousness and (2) willingness to help and interact.

The questions are specified below.

1. In your opinion, do any of the following apply to the client? You can choose up to three responses.

a. Depression; b. Anger management issues; c. Narcissistic personality disorder; d. Bipolar disorder; e. Obsessive compulsive personality disorder; f. Sex addiction; g. Panic attacks; h. Perfectionism; i. Paranoid personality disorder; j. Obsessive compulsive disorder; k. Post-traumatic stress disorder; l. Pedophilia; m. Borderline personality disorder; n. Psychosis; o. None of these apply; p. Generalized anxiety disorder; q. Social anxiety disorder; r. Specific phobia; s. Impulse control disorder; t. Scrupulosity (strong moral or ethical standards); u. Other (please specify).

2. How confident are you in your choice(s) on the previous question?

a. Not confident at all; b. Slightly confident; c. Moderately confident; d. Confident; e. Very confident.

3. Based on the vignette, which form of therapy do you feel would be most beneficial as the first line of therapy for the client?

a. evidence-based intervention; b. Not evidence-based intervention.

4. How confident are you in your response(s) to the previous question?

a. Not at all confident; b. Slightly confident; c. Moderately confident; d. Confident; e. Very confident.

 doi:10.20944/preprints202406.0857.v1

5. Based on the evidence given in the vignette, how likely do you think it is that the client could harm another person?

a. Very likely; b. Likely; c. Somewhat likely; d. Neutral; e. Somewhat unlikely; f. Unlikely; g. Very unlikely.

6. Please state whether you would be willing to work with the client described in the vignette. (For the purposes of this question, please assume that you have the skills and experience necessary to do so.)

a. Very unwilling; b. Unwilling; c. Somewhat unwilling; d. Neutral; e. Somewhat willing; f. Willing; g. Very willing.

Statistical analysis

All analyses were conducted using R software version 4.4.0 [39] and R-Studio version 2023.06.1 [40]. All analyses were two-tailed, and statistical significance was determined at a significance level of $\alpha = 0.05$. A chi-square test of independence ($\chi 2$) was used to compare the rates of OCD identification against client gender. Additionally, $\chi 2$ analyses were used to compare the total rates of OCD identification across the six OCD presentations. The statistical methodology employed in this study was designed to provide a rigorous assessment of the effectiveness of various AI tools and human professionals in recognizing OCD. All research questions were examined using chi-square testing to assess differences. Due to the multiple comparisons involved in comparing each pair of entities, the Bonferroni correction method was used. This approach quantifies the differences in patterns, highlighting the potential of integrating AI into clinical decision-making. With respect to client gender, there were no significant differences in recognition rates across any of the vignette subtypes. Consequently, the data for the male and female versions of each vignette type were combined for the remainder of the analysis, as reported by Canavan [27].

3. Results

RQ1: Comparison of effectiveness of AI tools (ChatGPT-3.5, ChatGPT-4, Claude.AI, and Gemini) and of psychotherapists in recognizing OCD

To evaluate the differences in performance across these entities, chi-square tests were conducted to assess the statistical significance of the observed frequencies in each category. Given the multiple comparisons across entities, post-hoc pairwise chi-square tests with Bonferroni correction were employed to identify which specific pairs of entities exhibited significantly different recognition rates.

Chi-square tests revealed significant differences in total recognition rates across entities ($\chi^2(4) = 1906.6$, $p < .001$). Post hoc pairwise comparisons with Bonferroni correction pointed to significant differences between psychotherapists and all AI entities. Specifically, ChatGPT 3.5 ($\chi^2(1) = 421.61$, $p < .001$), ChatGPT-4 ($\chi^2(1) = 712.83$, $p < .001$), Claude.AI ($\chi^2(1) = 712.83$, $p < .001$), and Gemini ($\chi^2(1) = 712.83$, $p < .001$) all exhibited significantly higher OCD recognition rates than the psychotherapists. These findings suggest that AI tools outperform human professionals in recognizing OCD. Table 1 and Figure 1 illustrate the OCD recognition rates for the different entities.

Table 1. Comparison of OCD Recognition Rates across Different Entities.

Vignette	Psychotherapists	ChatGPT-3.5	ChatGPT-4	Claude.AI	Gemini
Total Recognition	47.3%	90.0%	100.0%	100.0%	100.0%
Pedophilia	36.1%	100.0%	100.0%	100.0%	100.0%
Aggressive	26.7%	100.0%	100.0%	100.0%	100.0%
Religious	36.7%	90.0%	100.0%	100.0%	100.0%
Homosexuality	31.7%	100.0%	100.0%	90.0%	100.0%
Contamination/ Symmetry	77.3%	30.0%	100.0%	100.0%	100.0%

 doi:10.20944/preprints202406.0857.v1

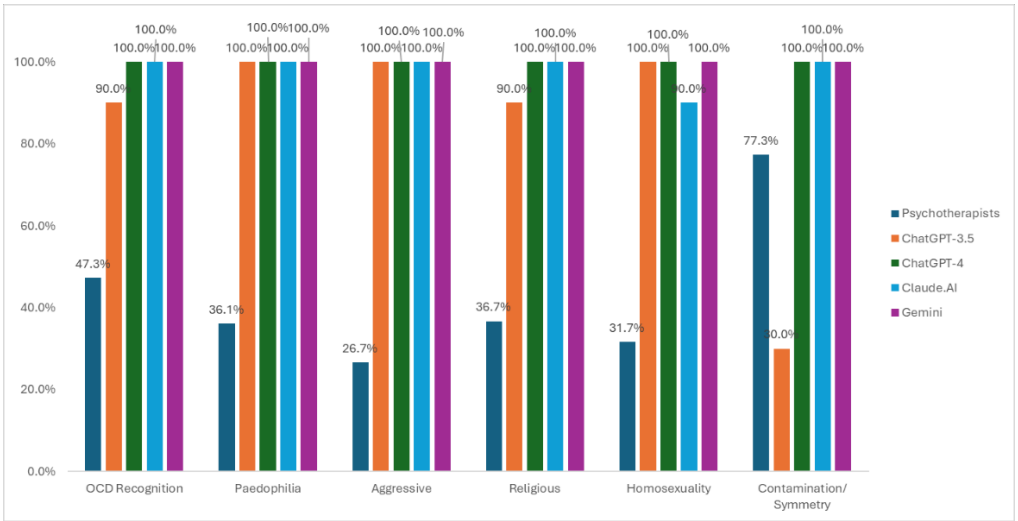

Figure 1. Comparison of OCD Recognition Rates across Different Entities.

The analyses were conducted independently for each vignette and revealed a similar pattern of results.

Pedophilia Vignette: Chi-square tests showed significant differences in recognition rates across entities ($\chi^2(4) = 2930.5$, $p < .001$). Post hoc pairwise comparisons with Bonferroni correction revealed significant differences between psychotherapists and all AI entities. Specifically, ChatGPT-3.5 ($\chi^2(1) = 621.21$, $p < .001$), ChatGPT-4, Claude.AI, and Gemini (all $\chi^2(1) = 936.08$, $p < .001$) demonstrated significantly higher OCD recognition rates compared to psychotherapists.

Aggressive Vignette: Chi-square tests showed significant differences in recognition rates across entities ($\chi^2(4) = 3435.7$, $p < .001$). Post-hoc pairwise comparisons with Bonferroni correction revealed significant differences between psychotherapists and all AI entities. ChatGPT-3.5, ChatGPT-4, Claude.AI, and Gemini (all $\chi^2(1) = 1153.9$, $p < .001$) exhibited significantly better OCD recognition rates than psychotherapists.

Religious Vignette: Chi-square tests showed significant differences in recognition rates across entities ($\chi^2(4) = 2423.7$, $p < .001$). Post-hoc pairwise comparisons with Bonferroni correction indicated significant differences between psychotherapists and all AI entities. ChatGPT-3.5 ($\chi^2(1) = 609.5$, $p < .001$), ChatGPT-4, Claude.AI, and Gemini ($\chi^2(1) = 923.19$, $p < .001$) demonstrated significantly higher OCD recognition rates compared to psychotherapists.

Homosexual Vignette: Chi-square tests revealed significant differences in recognition rates across entities ($\chi^2(4) = 2679.3$, $p < .001$). Post hoc pairwise comparisons with Bonferroni correction showed significant differences between psychotherapists and all AI entities. Claude.AI ($\chi^2(1) = 710.92$, $p < .001$), ChatGPT-3.5, ChatGPT-4, and Gemini (all $\chi^2(1) = 1034.2$, $p < .001$) demonstrated significantly higher OCD recognition rates than the psychotherapists.

Contamination/Symmetry Vignette: Chi-square tests indicated significant differences in recognition rates across entities ($\chi^2(4) = 2447.7$, $p < .001$). Post-hoc pairwise comparisons with Bonferroni correction revealed significant differences between psychotherapists and all AI entities. ChatGPT-4, Claude.AI, and Gemini (all $\chi^2(1) = 253.81$, $p < .001$) demonstrated significantly higher OCD recognition rates than psychotherapists. ChatGPT-3.5 ($\chi^2(1) = 447.96$, $p < .001$) was the only AI that performed worse than the psychotherapists.

With the minor exception of ChatGPT-3.5's performance in the contamination/symmetry vignette, in all cases AI exhibited significantly better OCD recognition rates than the psychotherapists, indicating that AI tools outperformed human professionals in recognizing OCD across various vignettes.

 doi:10.20944/preprints202406.0857.v1

Moreover, when examining the differences between psychotherapists and AI entities regarding their confidence in identifying the correct psychological diagnosis based on the provided vignettes, chi-square tests indicated significant differences across the entities ($\chi^2(4) = 533.88$, $p < .001$). Post-hoc pairwise comparisons with Bonferroni correction revealed significant differences between psychotherapists and all AI entities. ChatGPT-3.5, ChatGPT-4, Claude.AI, and Gemini (all $\chi^2(1) = 136.91$, $p < .001$), with the AI entities displaying significantly higher confidence rates than the psychotherapists (Figure 2).

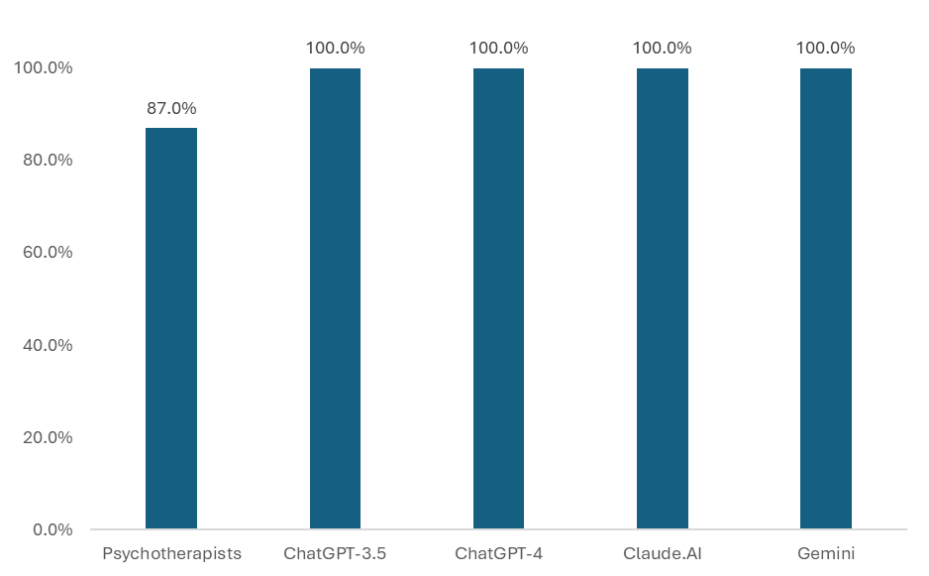

Figure 2. Comparison of confidence Recognition Rates Across Different Entities.

RQ2: Comparison between psychotherapists and AIs in their treatment decisions (whether or not they decided upon evidence-based intervention)

Based on the provided vignettes, chi-square tests indicated significant differences across entities ($\chi^2(4) = 1568.4$, $p < .001$). Post-hoc pairwise comparisons with Bonferroni correction revealed significant differences between psychotherapists and all AI entities. ChatGPT-3.5 ($\chi^2(1) = 786.61$, $p < .001$), ChatGPT-4 ($\chi^2(1) = 486.75$, $p < .001$), Claude.AI ($\chi^2(1) = 786.61$, $p < .001$), and Gemini ($\chi^2(1) = 54.513$, $p < .001$) demonstrated significantly higher evidence-based training (EBT) treatment recommendation rates than psychotherapists (Figure 3).

 doi:10.20944/preprints202406.0857.v1

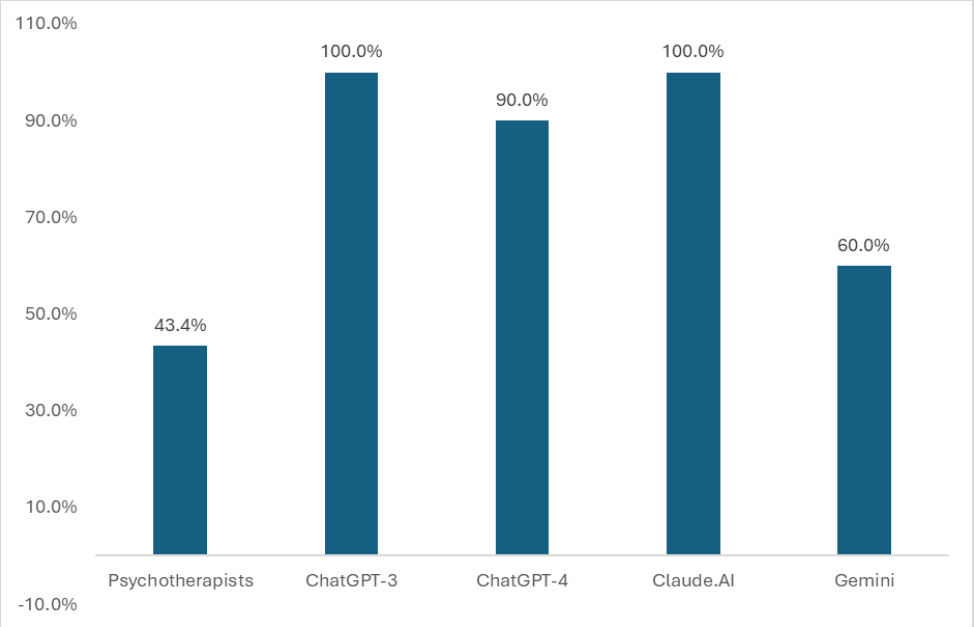

Figure 3. Comparison of EBT recommendation rates across different entities.

Moreover, chi-square tests that compared the EBT treatment recommendation rates between psychotherapists and AI entities only for those that recognized OCD pointed to significant differences across the entities ($\chi^2(4) = 1101.8$, $p < .001$). Post hoc pairwise comparisons with Bonferroni correction also revealed significant differences between psychotherapists and most AI entities. ChatGPT-3.5 ($\chi^2(1) = 468.19$, $p < .001$), ChatGPT-4 ($\chi^2(1) = 214.61$, $p < .001$), and Claude.AI ($\chi^2(1) = 468.19$, $p < .001$) demonstrated significantly higher EBT treatment recommendation rates compared to psychotherapists, whereas Gemini ($\chi^2(1) = 0.68$, $p = .41$) showed rates similar to those of psychotherapists (Figure 4).

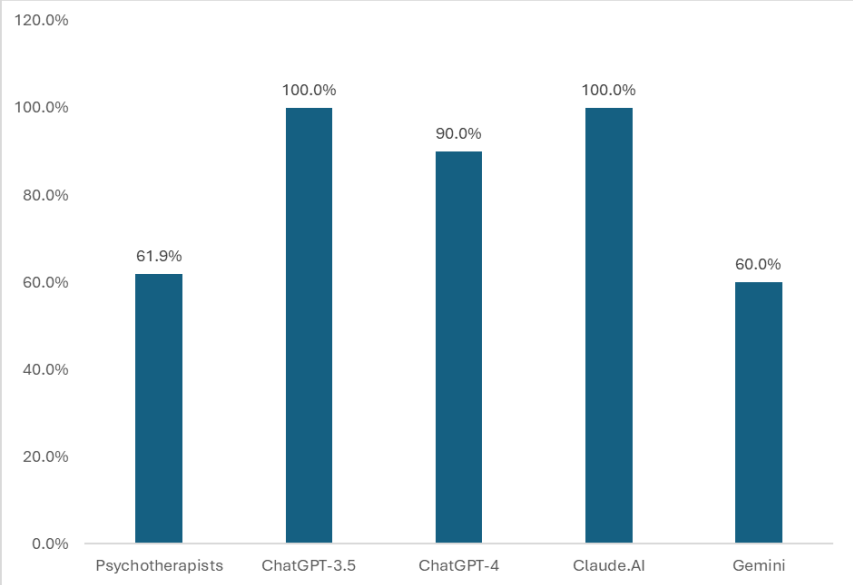

Figure 4. Comparison of EBT treatment recommendation rates across different entities only for those with recognized OCD.

Chi-square tests that examined the differences between psychotherapists and AIs regarding their confidence in their EBT treatment recommendation pointed to significant differences across the entities ($\chi^2(4) = 807.15$, $p < .001$). Post-hoc pairwise comparisons with Bonferroni correction also revealed significant differences between psychotherapists and all AI entities. ChatGPT-3.5 ($\chi^2(1) =$

 doi:10.20944/preprints202406.0857.v1

189.39, $p < .001$), ChatGPT-4 ($\chi^2(1) = 436.59$, $p < .001$), Claude.AI ($\chi^2(1) = 62.701$, $p < .001$), and Gemini ($\chi^2(1) = 436.59$, $p < .001$) all demonstrated significantly higher confidence rates than psychotherapists (Figure 5).

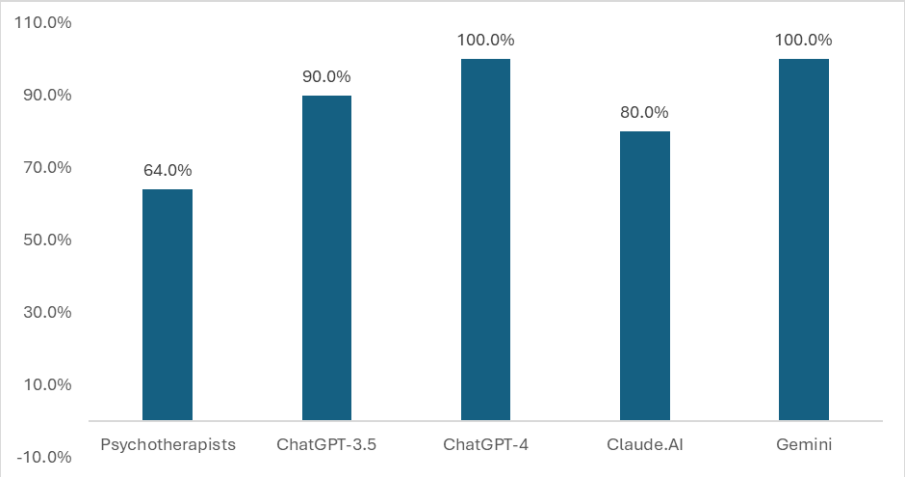

Figure 5. Comparison of confidence in EBT treatment recommendation across different entities.

RQ3: Comparison between AIs and psychotherapists in their stigma and danger estimations

Chi-square tests indicated significant differences across entities ($\chi^2(4) = 673.97$, $p < .001$) in stigma and danger estimations. Post-hoc pairwise comparisons with Bonferroni correction also revealed significant differences between psychotherapists and all AI entities. ChatGPT-3.5, ChatGPT-4, Claude.AI, and Gemini (all $\chi^2(1) = 175.29$, $p < .001$) all exhibited significantly lower estimations (all 0%) of danger than psychotherapists (16.3%).

Finally, in answer to the question of whether they would be willing to treat the person described in the vignettes, chi-square tests pointed to significant differences across the entities ($\chi^2(4) = 251.11$, $p < .001$). Post-hoc pairwise comparisons with Bonferroni correction also revealed significant differences between psychotherapists and all AI entities. ChatGPT-3.5, ChatGPT-4, Claude.AI, and Gemini (all $\chi^2(1) = 61.936$, $p < .001$) all demonstrated significantly higher willingness to treat the person than did the psychotherapists (Figure 6).

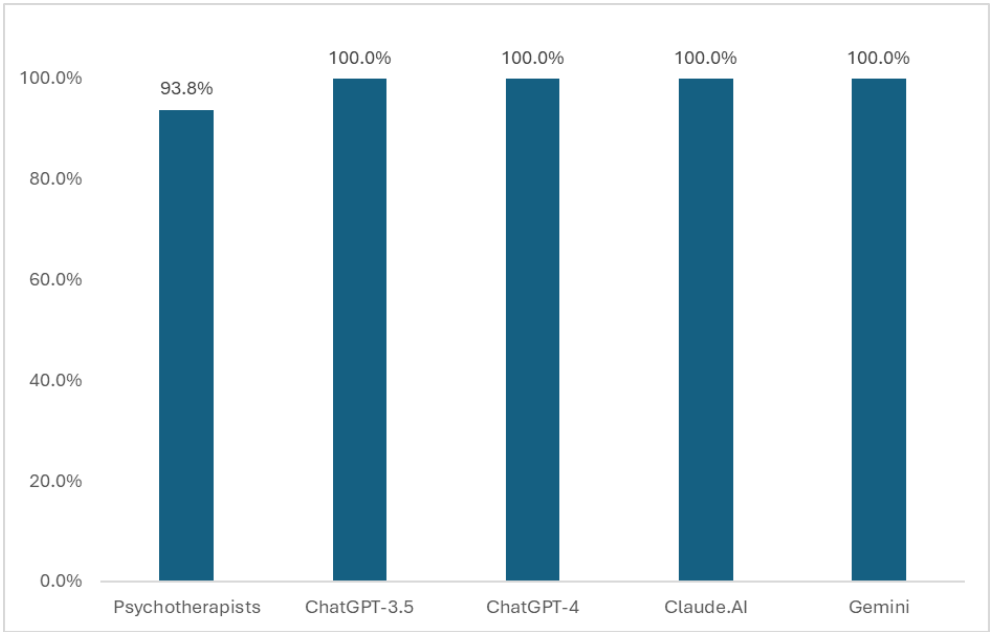

Figure 6. Comparison of treatment willingness across different entities.

 doi:10.20944/preprints202406.0857.v1

4. Discussion

The aim of this study was to evaluate the ability of four advanced artificial intelligence models (ChatGPT-3.5, ChatGPT-4, Claude, and Bard) to accurately recognize obsessive-compulsive disorder, compared to human professionals. Additionally, the study assessed the recommended therapies and stigma attributions provided by these AI models.

The study examined the effectiveness of AI tools (ChatGPT-3.5, ChatGPT-4, Claude.AI, and Gemini) in recognizing OCD. Significant differences were found between psychotherapists and all the AI models. All the artificial intelligence tools demonstrated significantly higher OCD recognition rates than did the psychotherapists. These findings suggest that AI tools outperform human professionals in recognizing OCD. All the AI models also demonstrated significantly higher confidence in their identification than the psychotherapists. Whereas the psychotherapists reported rates of 87% confidence in their recognition, ChatGPT-3.5, Claude.AI, and Gemini all reported a confidence rate of 100%. Identification data from artificial intelligence are much more significant than human identification data. Glazier et al. [36] evaluated the diagnostic accuracy of 208 primary care physicians in identifying various OCD presentations using vignettes depicting common OCD subtypes, including symmetry-related concerns. The study revealed that OCD symptoms were misdiagnosed in approximately 50.5% of cases. These misidentification rates varied by vignette type, with sexual obsessions (70.8–84.6%), aggression (80.0%), somatic concerns (40.0%), religious obsessions (37.5%), contamination (32.3%), and symmetry-related concerns (3.7%) revealing differing levels of diagnostic challenge. In a subsequent study, Glazier et al. [34] reported an overall misidentification rate of 38.9% across the OCD case vignettes. Glazier's study highlighted those physicians had substantially more difficulty recognizing uncommon or taboo-themed OCD symptoms, with misidentification rates as high as 75%, compared with more common OCD presentations such as contamination obsessions, which had a misidentification rate of 15.8%. Moreover, in a study utilizing these vignettes among mental health care providers, the rates of incorrect (non-OCD) diagnoses were significantly higher for vignettes depicting taboo thoughts (34.7%-52.7%) than for those depicting contamination (11%) and symmetry obsessions (6.9%) [37].

In the current study, AI models demonstrated significantly higher evidence-based intervention recommendations than psychotherapists. Whereas psychotherapists reported confidence rates of 61.9%, ChatGPT-3.5, and Claude.AI reported 100% confidence, ChatGPT-4 reported approximately 90%, and Gemini reported 60%. In comparison, confidence in evidence-based treatment recommendations among the AI models revealed levels of cultural intelligence ranging from 80-100%, whereas psychotherapists reported 64% confidence in evidence-based treatment. This finding supports previous research on artificial intelligence languages in mental health care in which AI exhibited better results than human samples [6–8]. Part of the delay between symptom onset and receipt of evidence-based interventions may be attributed to a lack of healthcare provider awareness of the diverse symptom presentations of OCD, particularly those related to taboo thoughts [22–24]. A study examining the diagnostic impressions and treatment recommendations of primary care physicians revealed that those who misidentified OCD vignettes were less likely to recommend empirically supported treatment [36].

In this study, we compared AI models and psychotherapists. Our findings reveal significant differences in stigma and estimations of the danger of the person described in the vignette. All LLM models demonstrated significantly lower stigma and danger estimations than did the psychotherapists. When asked if they would be willing to treat the person described in the vignettes, both the AI models and psychotherapists reported high willingness (AI models 100% and psychotherapists 93.8%). Stigma among the general population can discourage people from seeking help, and stigma among psychotherapists has the potential to be even more detrimental [41]. The behavior and attitude of mental health providers towards their clients can significantly affect treatment outcomes [42,43]. Nevertheless, the present study and the study by Canavan [27] provide encouraging data that show low levels of stigma among mental health professionals. This finding contrasts with several studies that have reported higher levels of stigma among mental health professionals. Despite the expectation that clinicians are trained to be free of bias, research indicates

 doi:10.20944/preprints202406.0857.v1

that they may hold negative views of individuals with mental illnesses, mirroring those of the general public [42,44]. Other studies have examined how the attitudes and prejudices of clinicians affect their professional work. Previous research found that stigma towards sexual obsessions is significantly greater than towards contamination obsessions among adults [45]. Additionally, Steinberg [38] pointed out that clinicians maintain more stigmas regarding clients with contamination, harm, and sexual obsessions than regarding those with scrupulous obsessions, leading to a decreased likelihood of clients revealing these thoughts.

To the best of our knowledge, this study is the first to compare the use of different AI models for identifying and treating OCD. The results for the AI models were very good, especially compared to a human sample. Yet it is important to note that the accuracy of LLM predictions is closely related to the quality and inclusiveness of the data used for training. Biases in the data or a lack of diversity in the demographic sample can result in incorrect predictions. In addition, LLM algorithms often operate as opaque systems, making it difficult to understand how they arrive at specific conclusions. These variations have significant implications for how LLMs report OCD. Recognizing these differences and their potential influence on patient outcomes is crucial for ongoing advancement and incorporation of LLMs into mental health care. Furthermore, the results highlight the importance of using various LLM versions in a complementary manner rather than relying exclusively on one. Intelligent language models can be used for training, support, and expert advice but should not replace skilled professionals who are trained to see the full picture.

Limitations

The current study has certain limitations that warrant acknowledgment. First, the study was confined to specific iterations of ChatGPT-3.5, ChatGPT-4, Claude, and Bard at a particular point in time, without considering subsequent versions. Future investigations should address this limitation by examining updates and improvements in forthcoming versions of these AI models. Second, the data were compared to a representative sample of human professionals (psychotherapists). While the sample was intended to be representative, it could not encompass the full spectrum of global psychotherapeutic practices. Therefore, the findings may not be universally generalizable. Third, this study utilized case vignettes which, although a useful methodological tool, simplify the complexities often encountered in real-life clinical scenarios. Although these vignettes were validated and used in several articles, they do not encompass the full range of symptoms and background details present in actual patients. Hence this methodology has inherent limitations, especially considering the multifaceted nature of OCD. Additionally, this study did not directly evaluate the clinical accuracy of large language models (LLMs) compared to human professionals. While the comparison provided valuable insights into the capabilities of AI models, future research should incorporate validation studies to assess the clinical utility and reliability of LLM predictions in real-world settings. Furthermore, the study did not explore potential cultural or demographic biases within LLMs.

5. Conclusions

This study evaluated the ability of four AI models (ChatGPT-3.5, ChatGPT-4, Claude, and Bard) to recognize OCD and recommend evidence-based treatments and compared their recognition and recommendations to human professionals. Utilizing 12 vignettes depicting clients with OCD, the study found that AI models significantly outperformed psychotherapists in terms of recognition rates, confidence levels, and recommendation of appropriate interventions. Additionally, AI models exhibited lower stigma and danger estimates in OCD cases.

These results underscore the potential of AI tools to enhance OCD diagnosis and treatment in clinical settings. AI models demonstrated higher confidence and more consistent recognition of OCD across various subtypes, including those related to taboo-intrusive thoughts, which often pose a challenge for human professionals to identify. These findings suggest that AI tools can be valuable in clinical practice and have the potential to improve the accuracy and timeliness of OCD diagnosis and treatment.

 doi:10.20944/preprints202406.0857.v1

Abbreviations

OCD: Obsessive-Compulsive Disorder
LLM: Large Language Models
NLP: Natural Language Processing
AI: Artificial Intelligence

Funding: This research received no external funding.

Institutional Review Board Statement: This study involves no human subjects and thus is exempt from Institutional Review Board (IRB) review.

Informed Consent Statement: Not applicable.

Data Availability Statement: The author has the research data, which is available upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, H.; Zhang, C.; Wang, Y. Revealing the technology development of natural language processing: A Scientific entity-centric perspective. *Information Processing & Management* **2024**, *61*, 103574. https://doi.org/10.1016/j.ipm.2023.103574
2. Motlagh, N.Y.; Khajavi, M.; Sharifi, A.; Ahmadi, M. The impact of artificial intelligence on the evolution of digital education: A comparative study of openAI text generation tools including ChatGPT, Bing Chat, Bard, and Ernie. *arXiv preprint arXiv:2309.02029* **2023**. https://doi.org/10.48550/arXiv.2309.02029
3. Rudolph, J.; Tan, S.; Tan, S. War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *Journal of Applied Learning and Teaching* **2023**, *6*. https://doi.org/10.37074/jalt.2023.6.1.23
4. Lee, H. The rise of ChatGPT: Exploring its potential in medical education. *Anatomical sciences education* **2023**. https://doi.org/10.1002/ase.2270
5. Borna, S.; Barry, B.A.; Makarova, S.; Parte, Y.; Haider, C.R.; Sehgal, A.; Leibovich, B.C.; Forte, A.J. Artificial Intelligence Algorithms for Expert Identification in Medical Domains: A Scoping Review. *European Journal of Investigation in Health, Psychology and Education* **2024**, *14*, 1182-1196. https://doi.org/10.3390/ejihpe14050078
6. Levkovich, I.; Elyoseph, Z. Suicide risk assessments through the eyes of Chatgpt-3.5 versus ChatGPT-4: vignette study. *JMIR mental health* **2023**, *10*, e51232. doi: 10.2196/51232
7. Elyoseph, Z.; Levkovich, I. Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment. *Frontiers in psychiatry* **2023**, *14*, 1213141. https://doi.org/10.3389/fpsyt.2023.1213141
8. Elyoseph, Z.; Levkovich, I. Comparing the perspectives of generative AI, mental health experts, and the general public on schizophrenia recovery: case vignette study. *JMIR Mental Health* **2024**, *11*, e53043. doi: 10.2196/53043
9. Gomez-Cabello, C.A.; Borna, S.; Pressman, S.; Haider, S.A.; Haider, C.R.; Forte, A.J. Artificial-Intelligence-Based Clinical Decision Support Systems in Primary Care: A Scoping Review of Current Clinical Implementations. *European Journal of Investigation in Health, Psychology and Education* **2024**, *14*, 685-698. https://doi.org/10.3390/ejihpe14030045
10. Tal, A.; Elyoseph, Z.; Haber, Y.; Angert, T.; Gur, T.; Simon, T.; Asman, O. The artificial third: utilizing ChatGPT in mental health. *The American Journal of Bioethics* **2023**, *23*, 74-77. https://doi.org/10.1080/15265161.2023.2250297
11. Haber, Y.; Levkovich, I.; Hadar-Shoval, D.; Elyoseph, Z. The artificial third: a broad view of the effects of introducing generative artificial intelligence on psychotherapy. *JMIR Mental Health* **2024**, *11*, e54781. doi: 10.2196/54781
12. Tornero-Costa, R.; Martinez-Millana, A.; Azzopardi-Muscat, N.; Lazeri, L.; Traver, V.; Novillo-Ortiz, D. Methodological and quality flaws in the use of artificial intelligence in mental health research: systematic review. *JMIR Mental Health* **2023**, *10*, e42045. doi: 10.2196/42045
13. Cervin, M. Obsessive-compulsive disorder: Diagnosis, clinical features, nosology, and epidemiology. *Psychiatric Clinics* **2023**, *46*, 1-16. https://doi.org/10.1016/j.psc.2022.10.006
14. Horwath, E.; Weissman, M.M. The epidemiology and cross-national presentation of obsessive-compulsive disorder. *Obsessive-Compulsive Disorder and Tourette's Syndrome* **2022**, 35-49.
15. Fineberg, N.A.; Dell'Osso, B.; Albert, U.; Maina, G.; Geller, D.; Carmi, L.; Sireau, N.; Walitza, S.; Grassi, G.; Pallanti, S. Early intervention for obsessive compulsive disorder: An expert consensus statement. *European Neuropsychopharmacology* **2019**, *29*, 549-565. https://doi.org/10.1016/j.euroneuro.2019.02.002

16. Barnhill, J. In *Obsessive-Compulsive and Related Disorders;* Textbook of psychiatry for intellectual disability and autism spectrum disorder; Springer: 2022; pp 625-654.
17. Wairauch, Y.; Siev, J.; Hasdai, U.; Dar, R. Compulsive rituals in Obsessive-Compulsive Disorder–A qualitative exploration of thoughts, feelings and behavioral patterns. *J. Behav. Ther. Exp. Psychiatry* **2024**, *84*, 101960. https://doi.org/10.1016/j.jbtep.2024.101960
18. Sharma, E.; Sharma, L.P.; Balachander, S.; Lin, B.; Manohar, H.; Khanna, P.; Lu, C.; Garg, K.; Thomas, T.L.; Au, A.C.L. Comorbidities in obsessive-compulsive disorder across the lifespan: a systematic review and meta-analysis. *Frontiers in psychiatry* **2021**, *12*, 703701. https://doi.org/10.3389/fpsyt.2021.703701
19. Melkonian, M.; McDonald, S.; Scott, A.; Karin, E.; Dear, B.F.; Wootton, B.M. Symptom improvement and remission in untreated adults seeking treatment for obsessive-compulsive disorder: A systematic review and meta-analysis. *J. Affect. Disord.* **2022**, *318*, 175-184. https://doi.org/10.1016/j.jad.2022.08.037
20. Liu, J.; Cui, Y.; Yu, L.; Wen, F.; Wang, F.; Yan, J.; Yan, C.; Li, Y. Long-term outcome of pediatric obsessive-compulsive disorder: A meta-analysis. *J. Child Adolesc. Psychopharmacol.* **2021**, *31*, 95-101. https://doi.org/10.1089/cap.2020.0051
21. Perris, F.; Sampogna, G.; Giallonardo, V.; Agnese, S.; Palummo, C.; Luciano, M.; Fabrazzo, M.; Fiorillo, A.; Catapano, F. Duration of untreated illness predicts 3-year outcome in patients with obsessive-compulsive disorder: A real-world, naturalistic, follow-up study. *Psychiatry Res.* **2021**, *299*, 113872. https://doi.org/10.1016/j.psychres.2021.113872
22. Thornicroft, G.; Bakolis, I.; Evans-Lacko, S.; Gronholm, P.C.; Henderson, C.; Kohrt, B.A.; Koschorke, M.; Milenova, M.; Semrau, M.; Votruba, N. Key lessons learned from the INDIGO global network on mental health related stigma and discrimination. *World Psychiatry* **2019**, *18*, 229-230. https://doi.org/10.1002/wps.20628
23. Leichsenring, F.; Sarrar, L.; Steinert, C. Drop-outs in psychotherapy: a change of perspective. *World Psychiatry* **2019**, *18*, 32. doi: 10.1002/wps.20588
24. Dell'Osso, B.; Benatti, B.; Oldani, L.; Spagnolin, G.; Altamura, A.C. Differences in duration of untreated illness, duration, and severity of illness among clinical phenotypes of obsessive-compulsive disorder. *CNS spectrums* **2015**, *20*, 474-478. doi:10.1017/S1092852914000339
25. Eisen, J.L.; Sibrava, N.J.; Boisseau, C.L.; Mancebo, M.C.; Stout, R.L.; Pinto, A.; Rasmussen, S.A. Five-year course of obsessive-compulsive disorder: predictors of remission and relapse. *J. Clin. Psychiatry* **2013**, *74*, 7286. https://doi.org/10.4088/JCP.12m07657
26. Shavitt, R.G.; van den Heuvel, O.A.; Lochner, C.; Reddy, Y.J.; Miguel, E.C.; Simpson, H.B. Obsessive-compulsive disorder (OCD) across the lifespan: Current diagnostic challenges and the search for personalized treatment. *Frontiers in psychiatry* **2022**, *13*, 927184. https://doi.org/10.3389/fpsyt.2022.927184
27. Canavan, R. Recognition rates, treatment recommendations and stigma attributions for clients presenting with taboo intrusive thoughts: A vignette-based survey of psychotherapists. *Counselling and psychotherapy research* **2024**, *24*, 51-62. https://doi.org/10.1002/capr.12557
28. Jacoby, R.J.; Blakey, S.M.; Reuman, L.; Abramowitz, J.S. Mental contamination obsessions: An examination across the obsessive-compulsive symptom dimensions. *Journal of obsessive-compulsive and related disorders* **2018**, *17*, 9-15. https://doi.org/10.1016/j.jocrd.2017.08.005
29. Bruce, S.L.; Ching, T.H.; Williams, M.T. Pedophilia-themed obsessive–compulsive disorder: Assessment, differential diagnosis, and treatment with exposure and response prevention. *Arch. Sex. Behav.* **2018**, *47*, 389-402. https://doi.org/10.1007/s10508-017-1031-4
30. Clemmensen, L.K.H.; Lønfeldt, N.N.; Das, S.; Lund, N.L.; Uhre, V.F.; Mora-Jensen, A.C.; Pretzmann, L.; Uhre, C.F.; Ritter, M.; Korsbjerg, N.L.J. Associations between the severity of obsessive-compulsive disorder and vocal features in children and adolescents: protocol for a statistical and machine learning analysis. *JMIR research protocols* **2022**, *11*, e39613. doi: 10.2196/39613
31. Lønfeldt, N.N.; Clemmensen, L.K.H.; Pagsberg, A.K. A wearable artificial intelligence feedback tool (wrist angel) for treatment and research of obsessive compulsive disorder: protocol for a nonrandomized pilot study. *JMIR research protocols* **2023**, *12*, e45123. doi: 10.2196/45123
32. Abd-Alrazaq, A.; Alhuwail, D.; Schneider, J.; Toro, C.T.; Ahmed, A.; Alzubaidi, M.; Alajlani, M.; Househ, M. The performance of artificial intelligence-driven technologies in diagnosing mental disorders: an umbrella review. *Npj Digital Medicine* **2022**, *5*, 87. https://doi.org/10.1038/s41746-022-00631-8
33. Segalàs, C.; Cernadas, E.; Puialto, M.; Fernández-Delgado, M.; Arrojo, M.; Bertolin, S.; Real, E.; Menchón, J.M.; Carracedo, A.; Tubío-Fungueiriño, M. Cognitive and clinical predictors of a long-term course in obsessive compulsive disorder: A machine learning approach in a prospective cohort study. *J. Affect. Disord.* **2024**, *350*, 648-655. https://doi.org/10.1016/j.jad.2024.01.157
34. Glazier, K.; Calixte, R.M.; Rothschild, R.; Pinto, A. High rates of OCD symptom misidentification by mental health professionals. *, 25, 3* **2013**, *25*, 201-209.
35. Glazier, K.; McGinn, L.K. Non-contamination and non-symmetry OCD obsessions are commonly not recognized by clinical, counseling and school psychology doctoral students. *Journal of Depression and Anxiety* **2015**, *4*, 10.4172. DOI: 10.4190/2167-1044.1000190

 doi:10.20944/preprints202406.0857.v1

36. Glazier, K.; Swing, M.; McGinn, L.K. Half of obsessive-compulsive disorder cases misdiagnosed: vignette-based survey of primary care physicians. *J. Clin. Psychiatry* **2015**, *76*, 7995.
37. Perez, M.I.; Limon, D.L.; Candelari, A.E.; Cepeda, S.L.; Ramirez, A.C.; Guzick, A.G.; Kook, M.; Ariza, V.L.B.; Schneider, S.C.; Goodman, W.K. Obsessive-compulsive disorder misdiagnosis among mental healthcare providers in Latin America. *Journal of obsessive-compulsive and related disorders* **2022**, *32*, 100693. https://doi.org/10.1016/j.jocrd.2021.100693
38. Steinberg, D.S.; Wetterneck, C.T. OCD taboo thoughts and stigmatizing attitudes in clinicians. *Community Ment. Health J.* **2017**, *53*, 275-280. https://doi.org/10.1007/s10597-016-0055-x
39. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2023. Vienna, Austria.
40. Team, RStudio. RStudio: Integrated development environment for R. RStudio, PBC. 2020. Boston, MA. URL: http://www.rstudio.com/
41. Chaves, A.; Arnáez, S.; Castilla, D.; Roncero, M.; García-Soriano, G. Enhancing mental health literacy in obsessive-compulsive disorder and reducing stigma via smartphone: A randomized controlled trial protocol. *Internet Interventions* **2022**, *29*, 100560. https://doi.org/10.1016/j.invent.2022.100560
42. Lauber, C.; Nordt, C.; Braunschweig, C.; Rössler, W. Do mental health professionals stigmatize their patients? *Acta Psychiatr. Scand.* **2006**, *113*, 51-59. https://doi.org/10.1111/j.1600-0447.2005.00718.x
43. Ponzini, G.T.; Steinman, S.A. A systematic review of public stigma attributes and obsessive–compulsive disorder symptom subtypes. *Stigma and Health* **2022**, *7*, 14. https://doi.org/10.1037/sah0000310
44. Kassam, A.; Glozier, N.; Leese, M.; Henderson, C.; Thornicroft, G. Development and responsiveness of a scale to measure clinicians' attitudes to people with mental illness (medical student version). *Acta Psychiatr. Scand.* **2010**, *122*, 153-161. https://doi.org/10.1111/j.1600-0447.2010.01562.x
45. Cathey, A.J.; Wetterneck, C.T. Stigma and disclosure of intrusive thoughts about sexual themes. *Journal of Obsessive-Compulsive and Related Disorders* **2013**, *2*, 439-443. https://doi.org/10.1016/j.jocrd.2013.09.001