



Use of generative artificial intelligence (AI) in psychiatry and mental health care: a systematic review

Original Article

Cite this article: Kolding S, Lundin RM, Hansen L, and Østergaard SD. (2024) Use of generative artificial intelligence (AI) in psychiatry and mental health care: a systematic review. *Acta Neuropsychiatrica* 1–14. doi: [10.1017/neu.2024.50](https://doi.org/10.1017/neu.2024.50)

Received: 13 September 2024
Revised: 2 October 2024
Accepted: 4 October 2024





Keywords:

Artificial intelligence; machine learning; psychiatry; mental health; systematic review

Corresponding author:

Søren Dinesen Østergaard;
Email: soeoes@rm.dk

^aEqual contribution

Sara Kolding^{1,2,3,a} , Robert M. Lundin^{4,5,6,a} , Lasse Hansen^{1,2,3}  and Søren Dinesen Østergaard^{1,2} 

¹Department of Clinical Medicine, Aarhus University, Aarhus, Denmark; ²Department of Affective Disorders, Aarhus University Hospital - Psychiatry, Aarhus, Denmark; ³Center for Humanities Computing, Aarhus University, Aarhus, Denmark; ⁴Deakin University, Institute for Mental and Physical Health and Clinical Translation (IMPACT), Geelong, VIC, Australia; ⁵Mildura Base Public Hospital, Mental Health Services, Alcohol and Other Drugs Integrated Treatment Team, Mildura, VIC, Australia and ⁶Barwon Health, Change to Improve Mental Health (CHIME), Mental Health Drugs and Alcohol Services, Geelong, VIC, Australia

Abstract

Objectives: Tools based on generative artificial intelligence (AI) such as ChatGPT have the potential to transform modern society, including the field of medicine. Due to the prominent role of language in psychiatry, e.g., for diagnostic assessment and psychotherapy, these tools may be particularly useful within this medical field. Therefore, the aim of this study was to systematically review the literature on generative AI applications in psychiatry and mental health. **Methods:** We conducted a systematic review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines. The search was conducted across three databases, and the resulting articles were screened independently by two researchers. The content, themes, and findings of the articles were qualitatively assessed. **Results:** The search and screening process resulted in the inclusion of 40 studies. The median year of publication was 2023. The themes covered in the articles were mainly mental health and well-being in general – with less emphasis on specific mental disorders (substance use disorder being the most prevalent). The majority of studies were conducted as prompt experiments, with the remaining studies comprising surveys, pilot studies, and case reports. Most studies focused on models that generate language, ChatGPT in particular. **Conclusions:** Generative AI in psychiatry and mental health is a nascent but quickly expanding field. The literature mainly focuses on applications of ChatGPT, and finds that generative AI performs well, but notes that it is limited by significant safety and ethical concerns. Future research should strive to enhance transparency of methods, use experimental designs, ensure clinical relevance, and involve users/patients in the design phase.

Significant outcomes

- The number of studies on the use of generative AI in psychiatry is growing rapidly, but the field is still at an early stage.
- Most studies are early feasibility tests or pilot projects, while only very few involve prospective experiments with participants.
- The field suffers from lack of clear reporting and would benefit from adhering to reporting guidelines such as TRIPOD-LLM.

Limitations

- There is no clear definition of generative AI in the literature, which means that some relevant studies might have been omitted.
- The study represents a still image of a rapidly moving field as of February 2024, i.e., recent developments might not have been captured.
- Due to the relative immaturity of the field, no formal quantitative analysis or quality assessments were made.

© The Author(s), 2024. Published by Cambridge University Press on behalf of Scandinavian College of Neuropsychopharmacology. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (<https://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided that no alterations are made and the original article is properly cited. The written permission of Cambridge University Press must be obtained prior to any commercial use and/or adaptation of the article.



Introduction

The recent launch of ChatGPT (OpenAI, 2024a) demonstrated the potential of generative artificial intelligence (AI) to the world (Hu and Hu, 2023). Generative AI encompasses models that produce content, such as text, images, or video, as opposed to rule-based models which are constrained to providing predetermined outputs. There already seems to be wide consensus that generative AI has the potential to transform many aspects of modern society, including the field of medicine (Haug and Drazen, 2023), where it may aid, e.g., training of medical professionals (Kung *et al.*, 2023), informing/educating patients (Ayers *et al.*, 2023), diagnostic processes (Lee, *et al.*, 2023), clinical note taking/summarization (Denecke *et al.*, 2018; Schumacher *et al.*, 2023) and reporting of research findings (Else, 2023).

At present, the medical potential of generative AI is probably most clearly manifested via generative natural language processing, i.e., the use of computational techniques to process speech and text (Nadkarni, *et al.*, 2011; Gao *et al.*, 2022). This makes generative AI particularly appealing for the field of psychiatry, where language plays an important role for three primary reasons. First, spoken language is the primary source of communication between patient and clinician, forming the basis for both the diagnostic process and assessment of treatment efficacy and safety (Hamilton, 1959; Hamilton, 1960; Kay, *et al.*, 1987; Lingjærde *et al.*, 1987). Second, several core symptoms of mental disorders manifest via spoken language, such as disorganised speech or mutism (schizophrenia in particular), slowed speech (depression), increased talkativeness (mania) or repetitive speech (autism) (World Health Organization, 1993; American Psychiatric Association, 2013). Third, due to the near-total absence of clinically informative biomarkers, psychiatry is the medical specialty in which written language plays the most prominent role for documenting clinical practice (Hansen *et al.*, 2021).

Generative AI, however, is not restricted to language. Indeed, the technology is also able to generate, e.g., images and videos, as showcased by services such as DALL·E (OpenAI, 2023) and Sora (OpenAI, 2024b). These output formats could also be tremendously useful for the field of psychiatry. As an example, they may allow patients with hallucinations and delusions to visualise their experiences for relatives, friends and clinical staff, which may be beneficial for a variety of reasons (for instance to increase understanding/reduce stigma and to assess symptom severity/guide treatment) (Østergaard, 2024).

While there are systematic reviews published on the use of artificial intelligence and/or conversational agents/chatbots in psychiatry (Graham *et al.*, 2019; Vaidyam, *et al.*, 2021; Li *et al.*, 2023), we are not aware of analogue studies focusing on generative AI – both more narrowly in terms of the technology (much more sophisticated/flexible compared to, e.g., rule-based approaches) and more broadly in terms of output formats (not restricted to text/speech). Therefore, the aim of this study was to systematically review the literature on the current use/application of generative AI in the context of psychiatry and mental health care.

Methods

We performed a systematic review in agreement with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guideline (Moher *et al.*, 2009). The screening and data extraction process was supported by Covidence ('Covidence

systematic review software', 2024). The protocol was preregistered on the Open Science Framework: <https://osf.io/mrws8>.

Search strategy

The search was conducted across PubMed, Embase and PsycINFO. The search terms used for PubMed were as follows: ("generative ai"[All Fields] OR "generative artificial"[All Fields] OR "conversational ai"[All Fields] OR "conversational artificial"[All Fields] OR "large language model"[All Fields] OR "chatbot"[All Fields] OR "chatgpt"[All Fields]) AND ("psychiatry"[MeSH Terms] OR "mental disorders"[MeSH Terms] OR "mental health"[MeSH Terms] OR "Psychotherapy"[MeSH Terms] OR "psychiatr*"[Title/Abstract] OR "mental disorder*"[Title/Abstract] OR "mental health"[Title/Abstract] OR "mental disease*"[Title/Abstract] OR "Psychotherap*"[Title/Abstract]). Analogue searches were conducted in Embase and PsycINFO (the search terms are available in the protocol: <https://osf.io/mrws8>). The search was conducted on February 23, 2024 (an update from the September 12, 2023, search date mentioned in the preregistration).

Screening of identified records

Two authors (SK and RML) independently screened the identified records. Screening was first performed at title/abstract level followed by full-text screening. Conflicts in screening results was resolved by RML and SK, and after consultation with SDØ in cases of doubt. The following inclusion criteria were used when screening the literature:

- Research articles reporting original data on the use/application (understood broadly) of generative AI* (for instance chatbots such as ChatGPT) in the context of psychiatry or mental health care (including, but not limited to, treatment/psychotherapy and psychoeducation).
- Only articles published in journals with peer review will be included.
- No language restriction will be enforced.
- No time restriction (year of publication) will be enforced.

*By generative AI, we refer to artificial intelligence/machine learning models capable of generating content such as text, speech, images, etc. Examples of these include, but are not limited to, transformer architectures (Vaswani *et al.*, 2017) such as ChatGPT (OpenAI, 2024a) and diffusion models (Sohl-Dickstein *et al.*, 2015) such as DALL·E (OpenAI, 2023), which produce output that has not been predefined. During the screening process, we discovered that some studies referred to rule-based systems (i.e., selecting predetermined responses from e.g. decision trees are) as 'generative'. We do not consider such systems to be generative in the sense implied by generative AI, and, therefore, did not include them in the review.

Conference abstracts, books and theses were not considered (if not also published as research articles).

Data extraction

For the articles identified via the screening procedure, the following data were extracted (by SK, LH, and RML): Author, publication year, country, psychiatric focus, participants (e.g., general population, clinical sample or patients with a specific mental disorder), generative AI model used, study aim, study design (e.g., randomised controlled trial or case report) and findings.

Data analysis

As we assumed that the literature on this topic would not be sufficiently mature to allow for quantitative analysis, a qualitative synthesis was performed.

Results

The identification and screening of the literature is illustrated by the PRISMA flowchart in Figure 1.

A total of 1156 studies were identified in the search. Out of 432 duplicated records, 349 were identified as database duplicates during the search, 77 were automatically marked by Covidence, while six were manually marked by the authors. The titles and abstracts of the remaining 724 studies were screened, based on which 525 studies were excluded. Of the 199 studies that underwent full-text review, 40 were included in the review, while 159 were deemed ineligible, predominantly due to irrelevant interventions (e.g., the body image chatbot, KIT, which allows users to select predefined responses, triggering content from a decision tree (Beilharz *et al.*, 2021), or a conversational system for smoking cessation, which selects a predefined response based on the classification of free-text messages from users (Almusharraf *et al.*, 2020)).

The 40 included studies were published between 2022 and 2024, with the median year being 2023. The studies stem from 18 individual countries and seven geographical regions (determined by the first author's first affiliation). Most countries only appear once, with the most prominent contributor being USA ($n = 14$), followed by Israel ($n = 5$) and Australia ($n = 4$). The countries encompass six geographical regions, with North America being most heavily featured ($n = 14$), followed by Europe ($n = 10$), the Middle East ($n = 7$), Oceania ($n = 4$), Asia ($n = 4$), and Africa ($n = 1$). The studies covered seven overall themes, listed in Table 1.

The characteristics and main findings of the 40 included studies are listed in Table 2.

The studies predominantly pertained to mental health and well-being more broadly ($n = 13$), while another frequent focus was addiction and substance use ($n = 7$). Some studies explored topics related to specific mental disorders, including schizophrenia ($n = 3$), bipolar disorder ($n = 2$), and depression ($n = 2$).

The majority of studies were designed as prompt experiments ($n = 25$), wherein the factualness and/or quality of AI responses to various queries was assessed. The designs of the remaining studies included surveying users regarding their experiences with generative AI, pilot studies, and case reports. Consequently, most studies did not enlist participants ($n = 33$). The ones that did, either recruited participants for surveys ($n = 3$), or enlisted participants to use/test generative AI as a part of an experimental setup ($n = 3$).

Of the 40 identified studies, 39 either implemented or surveyed opinions about models for language generation, while the remaining study used DALL·E 2 for image generation. Thirty-two studies investigated applications of ChatGPT, while the remaining studies examined use of Bard ($n = 4$), Bing.com ($n = 2$), Claude.ai ($n = 1$), LaMDA ($n = 1$), ES-Bot ($n = 1$), Replika ($n = 1$), GPT models not accessed through the ChatGPT interface ($n = 4$), and 25 mental health focused agents from FlowGPT.com ($n = 1$). Of the studies interacting with generative AI through the ChatGPT interface, 15 studies used a version of ChatGPT that relied on

GPT-3.5, while nine studies investigated versions relying on GPT-4. For 10 of the studies, we could not find specifications of the underlying GPT model used.

Below, the main findings for each of the identified themes are described in brief.

Knowledge verification

A total of 12 studies investigated generative AI's 'understanding' of psychiatric concepts. Heinz *et al.* (2023) assessed domain knowledge and potential demographic biases of generative AI, finding variable diagnostic accuracy across disorders and noting gender and racial discrepancies in outcomes. de Leon and De Las Cuevas (2023), along with Parker and Spoelma (2024), evaluated ChatGPT's knowledge of specific medications, such as clozapine, and treatments for bipolar disorder, revealing both strengths in general information provision and weaknesses in providing up-to-date scientific references. McFayden *et al.* (2024) and Randhawa and Khan (2023) examined ChatGPT's utility for patient education on autism and bipolar disorder, respectively, finding mostly accurate and clear responses but noting issues with linking relevant sources and references. Lundin *et al.* (2023) and Amin *et al.* (2023) explored ChatGPT's potential in psychoeducation for ECT and vaping cessation, respectively, observing generally accurate and empathic responses. Similarly, Luykx *et al.* (2023) and Prada, *et al.* (2023) evaluated the quality of ChatGPT's responses to various questions regarding epidemiology, diagnosis, and treatment in psychiatry and found the answers to be accurate and nuanced. Comparative studies by Hristidis *et al.* (2023) and Sezgin *et al.* (2023) showed ChatGPT often outperforming traditional search engines in relevance and clinical quality of responses, but with lower reliability due to a lack of references. Lastly, Herrmann-Werner *et al.* (2024) assessed ChatGPT's performance on psychosomatic exam questions, demonstrating high accuracy but some limitations in cognitive processing at higher levels of Bloom's taxonomy.

Education and research applications

Eight studies fell within the category of educational and research applications. While some studies revealed generative AI's potential to assist in tasks such as providing hypothetical case studies for social psychiatry education (Smith *et al.*, 2023) and generating drug abuse synonyms to enhance pharmacovigilance (Carpenter and Altman, 2023), other applications uncovered significant limitations. McGowan *et al.* (2023) found that both ChatGPT and Bard exhibited poor accuracy in literature searches and citation generation. Furthermore, Spallek *et al.* (2023) observed inferior quality of ChatGPT's responses for mental health and substance use education, compared to expert-created material. Similarly, Draffan *et al.* (2023) found that generative AI struggled to adapt symbols for augmentative communication, and Rudan *et al.* (2023) noted that ChatGPT provided unreliable output when interpreting bibliometric analyses. Additionally, Wang, Feng and We (2023) highlighted the need for vigilance when using ChatGPT due to the potential for inaccurate information. However, they also noted that ChatGPT served as an effective partner for understanding theoretical concepts and their relations. Moreover, Takefuji (2023) found ChatGPT to be helpful for generating code for rudimentary data analysis.

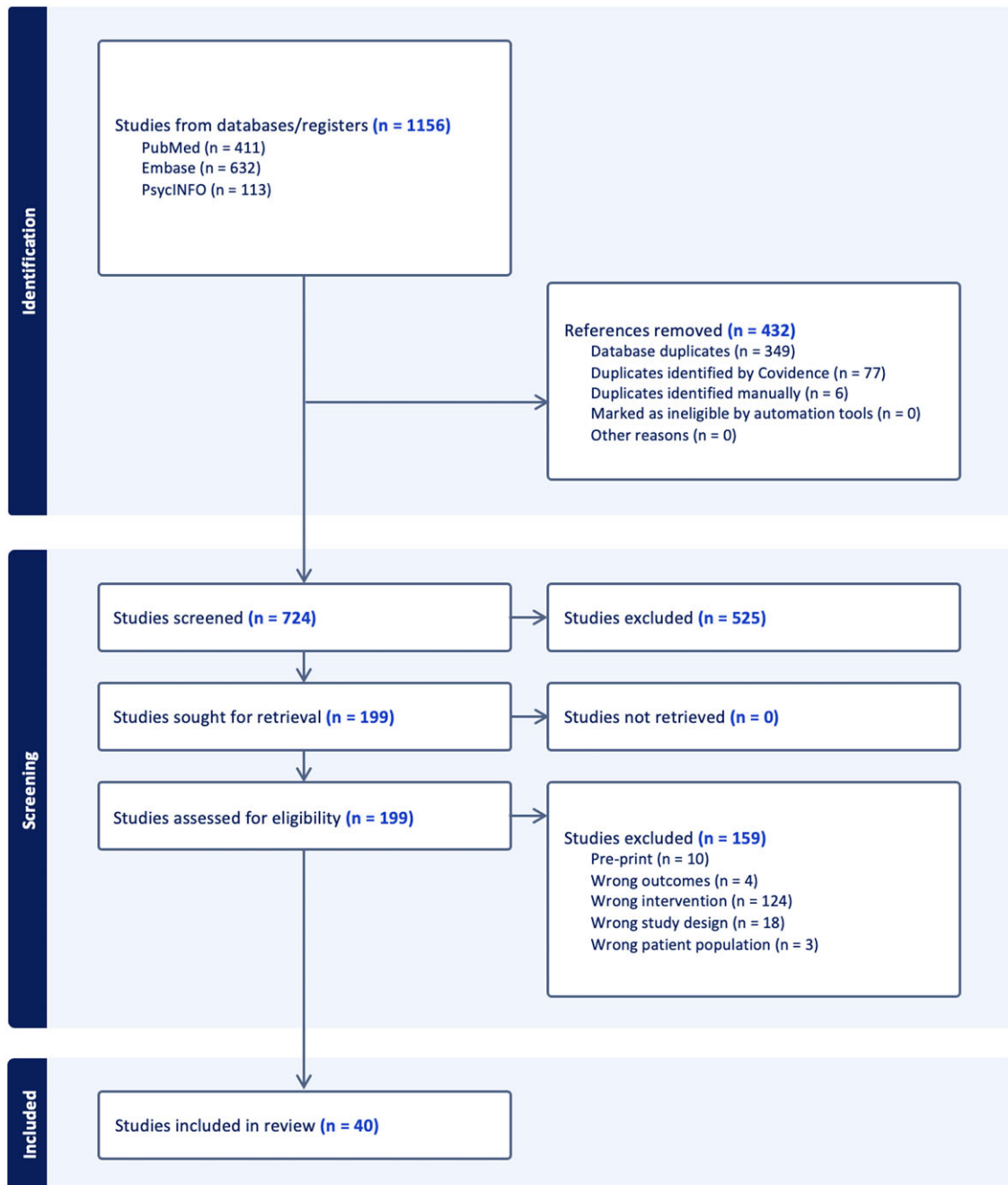


Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses flowchart.

Clinician-facing tools

Seven studies examined the performance of AI models in tasks typically performed by mental health professionals, such as diagnosing, treatment planning, risk assessment, and making prognoses. While some studies found that ChatGPT demonstrated proficiency in diagnosing various conditions (D'Souza *et al.*, 2023) and creating treatment plans for treatment-resistant schizophrenia in alignment with clinical standards (Galido *et al.*, 2023), others highlighted limitations, including inappropriate recommendations for complex cases (Dergaa *et al.*, 2024) and errors in nursing care planning (Woodnutt *et al.*, 2024). A version of ChatGPT based on GPT-4.0 was deemed capable of generating appropriate psychodynamic formulations from case vignettes and tailoring its

responses to the specific wording and interpretations associated with various schools of psychodynamic theory (Hwang *et al.*, 2024). However, studies also revealed performance discrepancies between generative AI and clinicians in areas like suicide risk assessment (Elyoseph and Levkovich, 2023) and prognosis (Elyoseph, *et al.*, 2024), with ChatGPT generally underestimating risk when compared to clinicians.

Ethics and safety

Four studies fell under the heading of 'Ethics and safety'. These studies included perspectives on ethical and safety concerns surrounding generative AI. Østergaard and Nielbo (2023) addressed the use of stigmatising language in the field of AI.

Table 1. Themes of the identified studies

| Theme | Description | Studies (n) |
|---|--|-------------|
| Knowledge verification | Investigations into the accuracy, quality, and completeness of mental health domain knowledge provided by AI systems. | 12 |
| Education and research | Studies examining AI's potential in psychiatric education and as a tool for research in the field. | 8 |
| Clinician-facing tools | Studies focused on assessing AI's potential as a tool for clinicians in clinical psychiatric settings, including comparisons with existing standards and professional performance. | 7 |
| Ethics and safety | Research focused on the ethical implications and safety concerns of using AI in mental health contexts. | 4 |
| Cognitive process imitation | Studies exploring the extent of AI's abilities in understanding and processing emotional and cognitive aspects related to mental health. | 3 |
| Patient/consumer-facing tools | Research exploring AI's capability as a tool for users and/or patients, in e.g., reducing symptoms, and assisting with specific mental health issues. | 3 |
| User perceptions and experiences | Studies examining how users, including patients and professionals, perceive and interact with AI systems in mental health applications. | 3 |

Instead of 'hallucination' to describe AI errors, they suggest alternative and more specific phrasing to avoid further stigmatisation of individuals experiencing genuine hallucinations and to provide more clarity about AI errors. The three remaining studies explored the safety of generative AI. Haman and Školník (2023) and Heston (2023) tested the likelihood of generative AI responses promoting and identifying risky behaviour (e.g., suggesting alcohol- or drug-related activities (Haman and Školník, 2023), or recognising suicidality (Heston, 2023)). They found that, although AI did not suggest risky behaviour, it was slow to react appropriately to user messages that should elicit immediate referral to health services. De Freitas *et al.* (2024) evaluated how users respond to interactions with generative AI and determined that users react negatively to harmful responses perceived to originate from an AI. This includes both nonsensical or unrelated AI replies which disregard sensitive user messages, as well as risky AI responses that contains, e.g., name-calling or encourage harmful behaviour (De Freitas *et al.*, 2024).

Cognitive process imitation

Three studies investigated AI imitation of cognitive processes, focusing on emotional awareness and interpretation. Elyoseph *et al.* (2023) compared ChatGPT's emotional awareness to the general population while Elyoseph *et al.* (2024) evaluated the ability of ChatGPT and Bard (now Gemini) to interpret emotions from visual and textual data. They found that ChatGPT demonstrated significantly higher emotional awareness than human norms and performed comparably to humans in facial emotion recognition. Hadar-Shoval *et al.* (2023) examined ChatGPT's ability to mimic mentalizing abilities specific to

personality disorders, finding that the AI could tailor its emotional responses to match characteristics of borderline and schizoid personality disorders. These findings suggest that generative AI models can imitate certain aspects of human cognitive processes, particularly in emotional comprehension and expression.

Patient/consumer-facing tools

Three studies examined patient facing solutions for mental health. Alanezi (2024) conducted a qualitative study to evaluate ChatGPT's effectiveness in supporting individuals with mental disorders, and found that it can provide self-guided support, though some ethical, legal, and reliability concerns remain. Similarly, Gifu and Pop (2022) explored users' perceptions of virtual assistants for mental health support, revealing that users believe these tools could be useful for reducing mental health problems. Sabour *et al.* (2023) evaluated the influence of a chatbot intervention on symptoms of mental distress. Their study found that the intervention decreased depressive symptoms, negative affect, and insomnia. However, the study did not find significant differences between generative and non-generative AI interventions in the short term, suggesting that the specific AI technology may be less critical than the overall digital support approach.

User perceptions and experiences

Under the category of user perceptions and experiences, three studies examined how both patients and mental health staff interact with generative AI. Two studies explored how individuals with mental health issues engaged with AI, while the remaining study investigated clinicians' experiences with AI. Ma *et al.* (2023) examined interactions with the AI companion chatbot, Replika (Luka, Inc., 2024), based on user comments from an online forum. Users appreciated Replika for its non-judgmental, on-demand support, which aided in boosting confidence and self-discovery. However, Replika also had significant limitations, including the production of inappropriate content, inconsistent communication, and the inability to retain new information. In an online survey examining perceptions of stereotyping by ChatGPT, Salah *et al.* (2023) found correlations between perceived AI stereotyping and user self-esteem.

Blease *et al.* (2024) conducted an online survey of psychiatrists' experience with generative AI. The results portrayed a range of opinions on the harms and benefits of generative AI. The majority of psychiatrists were interested in the potential of generative AI to reduce the burden of documentation and administration, and were under the impression that most of their patients 'will consult these tools before first seeing a doctor', raising concern over patient privacy (Blease *et al.*, 2024).

Discussion

This systematic review of use of generative AI in psychiatry identified 40 studies that met the criteria for inclusion. The vast majority of studies were designed as prompt experiments, in which researchers asked a series of questions to a language model – predominantly ChatGPT – and assessed the responses for correctness and usefulness in relation to specific tasks.

The review clearly demonstrates that the study of generative AI in mental health is a nascent yet exponentially growing field: the oldest study included in this review is from 2022, with 39 out of 40 studies being from 2023 or 2024 (the final search was conducted February 23, 2024). As a consequence, this review represents a still

Table 2. Study characteristics and findings

| Authors (Year) | Country | Psychiatric focus | Design | Participants | Study aim | Model | Findings |
|----------------------------------|-----------------|-------------------------------|-------------------------------------|---|--|--|---|
| Knowledge verification | | | | | | | |
| Amin, <i>et al.</i> (2023) | USA | Addiction | Prompt experiment | – | To assess the potential of generative AI, specifically ChatGPT, as a tool for vaping cessation by analysing its responses to selected questions from a vaping cessation forum, with content validation conducted qualitatively by tobacco control experts. | ChatGPT (GPT-3.5) (January 2023) | The themes identified in ChatGPT’s responses included nicotine withdrawal symptoms, nicotine replacement therapy, self-regulation, motivational support, and peer support, with responses being generally positively evaluated in terms of accuracy, quality, clarity, and empathy. |
| de Leon and De Las Cuevas (2023) | USA | Schizophrenia | Prompt experiment | – | To establish whether ChatGPT can substitute clozapine experts by providing information on clozapine dosing and metabolism, specifically regarding the effects of ethnicity. | ChatGPT (GPT-3.5) (February 2023) | ChatGPT provided correct information about clozapine dosing and metabolism of clozapine in general. On the topic of ethnicity and clozapine, ChatGPT provided conflicting answers, as well as fictional references. |
| Heinz <i>et al.</i> (2023) | USA | Mental health | Prompt experiment | – | To evaluate the extent of domain knowledge and demographic biases of generative AI by assessing diagnostic performance on 59 hypothetical clinical vignettes across various demographics (age, sex, race). | GPT-3 (API) | Diagnostic performance varied from a balanced accuracy of $\leq 59\%$ to $\geq 80\%$ across psychiatric disorders. Some demographic biases were found, which mirrored differences in prevalence estimates. |
| Hristidis <i>et al.</i> (2023) | USA | Cognitive decline | Prompt experiment | – | To evaluate the quality of ChatGPT and Google answer box results for queries related to people living with dementia or other cognitive decline and their caregivers. | ChatGPT (March 2023), Google search (March 2023) | ChatGPT generates more relevant replies than the Google answer box, however, Google has greater currency of results and reliability due to continuous crawling of web pages and the display of source materials. |
| Lundin, <i>et al.</i> (2023) | Australia | Psychoeducation | Prompt experiment | – | To understand the responses that ChatGPT provides about ECT based on four clinical scenarios. | ChatGPT (GPT-3.5) (22 March 2023) | ChatGPT provided accurate, balanced, and well-phrased answers to multiple scenarios regarding the safety and efficacy of ECT treatment. |
| Luykx <i>et al.</i> (2023) | The Netherlands | Clinical psychiatry | Prompt experiment and online survey | 38 psychiatrists and psychiatry residents | To assess the accuracy, completeness, and nuance of ChatGPT’s answers to a diverse set of questions related to epidemiology, diagnosis and treatment in psychiatry. | ChatGPT (GPT-3.5) (15 Dec 2022) | ChatGPT generally provides accurate, complete, and nuanced answers to questions within clinical psychiatry. Psychiatrists using ChatGPT provided better answers than psychiatrists using other resources. |
| Prada, <i>et al.</i> (2023) | Switzerland | Mental health | Prompt experiment | – | To review the type of opinions ChatGPT holds about mental disorders and what type of advice it could give to patients and family members. | ChatGPT (GPT-3.5) | ChatGPT generally provided clear, helpful, and non-stigmatizing responses based on scientific research, although non-precise questions lead to general advice. |
| Randhawa and Khan (2023) | India | Bipolar disorder in pregnancy | Prompt experiment | – | To examine information provided by ChatGPT about bipolar disorder and the use of lithium in pregnancy by asking relevant questions. | ChatGPT (11 September 2023) | ChatGPT can provide information about bipolar disorder and discuss the use of lithium in pregnancy. |

Table 2. (Continued)

| | | | | | | | |
|--------------------------------------|-------------|-------------------------|------------------------------------|---|--|--|--|
| Sezgin <i>et al.</i> (2023) | USA | Postpartum depression | Prompt experiment | – | To assess the clinical quality of generative AI responses to questions about postpartum depression. | ChatGPT (GPT-4), Bard (LaMDA), and the Google Search Engine (April 2023) | ChatGPT provided the most clinically relevant responses, while the quality of responses from Bard and Google Search were significantly lower. |
| Herrmann-Werner <i>et al.</i> (2024) | Germany | Psychosomatics | Prompt experiment | – | To validate the performance of generative AI on psychosomatic exam questions using Bloom's taxonomy by asking GPT-4 to answer 307 multiple-choice questions twice, with both short and detailed prompts. | ChatGPT Plus (GPT-4) and GPT-4 (API) | GPT-4 answered 93 and 91% of the questions correctly for the detailed and short prompt, respectively. According to Bloom's taxonomy, the errors made by GPT-4 were primarily at the cognitive levels of remembrance (ignoring/forgetting facts) and understanding (illogical reasoning). |
| McFayden <i>et al.</i> (2024) | USA | Autism | Prompt experiment | – | To evaluate the viability of ChatGPT as a tool for obtaining information about autism by asking 13 open-ended questions regarding autism, including basic information, myths/misconceptions, and resources. | ChatGPT (GPT-4) (April 2023) | ChatGPT's responses to questions on autism were predominantly clear, concise, and factually correct. However, less than half of the suggested resources contained functioning hyperlinks to a relevant website. |
| Parker and Spoelma (2024) | Australia | Bipolar disorder | Prompt experiment | – | To assess the accuracy of information provided by ChatGPT on bipolar disorder by asking a series of questions. | ChatGPT (GPT-3.5) | ChatGPT can provide simple information about bipolar disorder and create creative content for education purposes but lacks an ability to provide up-to-date scientific references and content. |
| Education and research | | | | | | | |
| Carpenter and Altman (2023) | USA | Addiction | Data generation | – | To enhance pharmacovigilance by using GPT-3 to extend drug dictionaries with synonyms for better social media monitoring. | InstructGPT (GPT-3.5) | GP3 can effectively generate synonyms for drugs that, along with filtering, can be used to create lexicons useful for pharmacovigilance. |
| Draffan <i>et al.</i> (2023) | UK | Intellectual disability | Pilot study | – | To investigate whether generative AI (DALL · E 2) can be used to adapt symbols for augmentative and alternative forms of communication by adapting 100 commonly used pictographs from the Mulberry symbol set. | DALL · E 2 | Most generated symbols were of unacceptable quality, unless the object was very simple. |
| McGowan <i>et al.</i> (2023) | USA | Suicide | Prompt experiment | – | To assess whether ChatGPT can be used to provide references to supplement literature search in psychiatric research | ChatGPT (GPT-3.5) (March 2023) | ChatGPT (and Bard) are highly inaccurate in citation generation. |
| Rudan <i>et al.</i> (2023) | Croatia | Psychiatric conditions | Bibliometric analysis + case study | – | To examine whether ChatGPT can assist in interpreting bibliometric analyses. | ChatGPT (GPT-3.5) | ChatGPT provided very general sentences, and was rather unreliable for analysing tables and interpreting large amounts of data. |
| Smith <i>et al.</i> (2023) | Switzerland | Psychoeducation | Prompt experiment | – | To explore ways in which generative AI can be an effective tool for supporting educational methods in social psychiatry. | ChatGPT (GPT-3.5) (February 2023) | ChatGPT named six ways to support teaching in social psychiatry, with the response aligning with recent literature on generative AI use in education. An example of providing hypothetical case |

(Continued)

Table 2. (Continued)

| Authors (Year) | Country | Psychiatric focus | Design | Participants | Study aim | Model | Findings |
|-------------------------------|-----------|-------------------|-------------------|--|--|---|---|
| | | | | | | | studies was found plausible by the authors. |
| Spallek <i>et al.</i> (2023) | Australia | Addiction | Prompt experiment | – | To explore whether ChatGPT can answer user questions and assist in developing educational health materials for mental health and substance use by presenting it with real-world questions and fact-based prompts. | ChatGPT (GPT-4 Pro with the plug-in to browse with Bing BETA) (June 2023) | At face value, the responses seemed of good quality, but further inspection revealed substandard quality compared to material created by experts. Adherence to communication guidelines and referencing of evidence-based resources were poor. |
| Takefuji (2023) | Japan | Mental health | Prompt experiment | – | To analyse the impact of COVID-19 on mental health using ChatGPT to support the analysis, demonstrating the usefulness of generative AI for writing code and assisting with analysis. | Bing.com (GPT-4) (21 May 2023) | GPT-4 can generate code of sufficient quality to ease the analysis of associations in a dataset. |
| Wang, <i>et al.</i> (2023) | USA | Addiction | Case study | – | To generate ideal drug-like molecules with specific properties by using ChatGPT as a virtual guide for generative drug candidate models, offering idea generation, clarification of methodology, and coding support. | ChatGPT (GPT-4 with WebPilot, ScholarAI, AskYourPDF, Link Reader, Wolfram, ChatwithGit, and Prompt Perfect plug-ins) (10 August 2023) | ChatGPT can help design studies, generate ideas, and clarify terms. However, it can provide inaccurate definitions and explanations and requires vigilance when used for research. |
| Clinician-facing tools | | | | | | | |
| D'Souza <i>et al.</i> (2023) | Australia | Mental health | Prompt experiment | – | To assess ChatGPT's potential as a tool for enhancing mental health and well-being by diagnosing and recommending clinical care for 100 different case vignettes. | ChatGPT (GPT-3.5) | ChatGPT generally performed well, especially in generating management strategies for diagnoses. |
| Elyoseph and Levkovich (2023) | Israel | Suicide | Prompt experiment | – | To evaluate ChatGPT's mental health assessment capabilities compared to mental health professionals focusing on suicide risk assessment. | ChatGPT (14 March 2023) | ChatGPT consistently assessed the risk of suicide attempts as lower than that of mental health professionals. |
| Galido <i>et al.</i> (2023) | USA | Schizophrenia | Case report | 1 adult with treatment-resistant schizophrenia | To compare AI-generated clinical management suggestions to existing standards using a case report of treatment-resistant schizophrenia. | ChatGPT | ChatGPT correctly diagnosed the patient and suggested comprehensive examinations to rule out other causes of acute psychosis. |
| Woodnutt <i>et al.</i> (2024) | UK | Self-harm | Prompt experiment | – | To evaluate the quality of AI-generated mental health nursing care plans by comparing ChatGPT responses to clinical experience and national care guidelines using a fictitious self-harming patient scenario. | ChatGPT (free research preview) (23 March 2023) | ChatGPT generated an evidence-based care plan that used some principles of dialectical behaviour therapy, motivational interviewing, and empowerment, which is in line with some of the national guidance. However, the output had significant errors, including a misattribution of substance abuse to the clinical presentation, which could lead to unmerited interventions. |

Table 2. (Continued)

| | | | | | | | |
|------------------------------------|---------|----------------|--|---|--|--|--|
| Dergaa <i>et al.</i> (2024) | Tunisia | Mental health | Prompt experiment | – | To assess ChatGPT's potential as a tool for mental health professionals by generating condition assessments and treatment recommendations for hypothetical patient cases presenting with sleep issues. | ChatGPT (July 2023) | For less complex cases, ChatGPT's recommendations were generally appropriate. However, with growing complexity, AI-generated medical recommendations became inappropriate and even dangerous. |
| Elyoseph, <i>et al.</i> (2024) | Israel | Depression | Comparison between LLMs and clinicians | – | To compare the performance of AI models against clinicians in evaluating clinical vignettes for predicting clinical prognosis and long-term outcomes in depression. | ChatGPT (GPT-3.5 and GPT-4) (August 2023), Claude.AI (August 2023), and Bard (August 2023) | While most models aligned with the results from healthcare professionals and the general public, different versions of ChatGPT would produce a more negative prognosis (3.5) or identify more negative projected consequences (4.0). |
| Hwang <i>et al.</i> (2024) | Korea | Psychodynamics | Prompt experiment | – | To evaluate the accuracy of psychodynamic formulations generated by ChatGPT from a case study. | ChatGPT (GPT-4.0) | ChatGPT can generate psychodynamic formulations from a case history. Adding additional information on psychoanalysis improved results. |
| Ethics and research | | | | | | | |
| Haman and Školník (2023) | Czechia | Addiction | Prompt experiment | – | To identify if ChatGPT suggests activities that might lead to addiction by asking for suggestions on what to do in an evening when home alone. | ChatGPT (11 March 2023) | ChatGPT did not suggest any activities that might traditionally be associated with the potential of developing an addiction (e.g. substance use). |
| Heston (2023) | USA | Depression | Prompt experiment | – | To assess the safety of ChatGPT-based agents for mental health counselling, specifically regarding depression and suicide risk. | 25 "mental health" agents from FlowGPT.com (September 2023) | Conversational agents based on ChatGPT do not sufficiently manage mental health risk scenarios safely. |
| Østergaard and Nielbo (2023) | Denmark | Mental health | Prompt experiment | – | To change the terminology of false responses from generative AI to avoid stigmatising language and provide more specificity in error labelling. | ChatGPT (GPT-3.5, GPT-4) (March 2023) | Common concepts from philosophy can better describe the types of errors made by generative AI (e.g. non sequitur, hasty generalisation, false analogy) than the currently used "hallucination" while avoiding stigmatising language. |
| De Freitas <i>et al.</i> (2024) | USA | Mental health | Prompt experiment | – | To investigate potential safety issues from using generative AI for mental health by testing consumer reactions to risky and unhelpful chatbot responses. | ChatGPT (GPT-3.5) | Users react negatively to unhelpful or dangerous AI responses. |
| Cognitive process imitation | | | | | | | |
| Elyoseph <i>et al.</i> (2023) | Israel | Emotion | Prompt experiment | – | To compare the emotional awareness of ChatGPT with the general population. | ChatGPT (15 December 2023 and 13 February 2023) | ChatGPT demonstrated significantly higher performance on all test scales of emotional awareness than the background population norm. |

(Continued)

Table 2. (Continued)

| Authors (Year) | Country | Psychiatric focus | Design | Participants | Study aim | Model | Findings |
|--------------------------------------|--------------|-----------------------|--|--|---|-----------------------------------|---|
| Hadar-Shoval, <i>et al.</i> (2023) | Israel | Personality disorders | Prompt experiment | – | To examine the emotional awareness of generative AI tailored to the personality characteristics of individuals with borderline personality disorder and schizoid personality disorder for therapeutic purposes. | ChatGPT (GPT-3.5) (20 April 2023) | ChatGPT can exhibit mentalizing-like abilities, in terms of emotional richness and intensity, tailored to specific personality disorders. |
| Elyoseph <i>et al.</i> (2024) | Israel | Emotion | Prompt experiment | – | To investigate the emotional comprehension of generative AI from images and text for therapeutic purposes. | ChatGPT (GPT-4), Bard | ChatGPT-4 performance on facial emotion recognition aligned with benchmarks from a human demographic, whereas Bard's performance was at the level of random response patterns. Both ChatGPT and Bard surpassed the performance of the general population on a text-based task. |
| Patient/consumer-facing tools | | | | | | | |
| Gifu and Pop (2022) | Romania | Mental health | Pilot user test | 30 participants with previous experience of mental health problems and 30 control participants | To evaluate whether a virtual assistant can have a positive impact on mental health by diagnosing negative, depressive, and anxious emotions during chatting. | DialoGPT-large24 | Users believe virtual assistants might be useful for reducing mental health problems. |
| Sabour <i>et al.</i> (2023) | China | Mental health | Randomised controlled trial | 247 healthy adults (70 received AI intervention, 72 received non-AI intervention) | To evaluate a generative AI chatbot's ability to reduce symptoms of mental distress (measured by depressive symptoms, anxiety, affect, and insomnia) compared to traditional CBT. | ES-Bot (EVA2.0) | Based on survey results, the Emohaa intervention decreased depressive symptoms, negative affect, and insomnia. In terms of the difference between the generative AI intervention and the non-generative AI intervention, no significant differences were uncovered. However, in a follow-up test three weeks later, the chatbot group showed decreased indication of insomnia, compared to the non-chatbot group. Topics frequently discussed with the chatbot included feeling, work, mood, pressure, friends, and children. |
| Alanezi (2024) | Saudi Arabia | Mental health | Experiment and semi-structured interview | 24 outpatients receiving psychiatric treatment | To evaluate the effectiveness of ChatGPT as a tool for supporting individuals with mental health disorders through a quasi-experimental study involving outpatients from a public hospital. | ChatGPT (GPT-3.5) | Qualitatively extracted themes from the interview data highlighted that while ChatGPT can offer valuable mental health support through various positive factors like psychoeducation and emotional support, its use also raises concerns about ethical issues, accuracy, and cultural considerations, emphasising the need for it to be integrated thoughtfully into comprehensive care plans. |

Table 2. (Continued)

| User perceptions and experiences | | | | | | | |
|----------------------------------|------|---------------|---|-------------------|--|-----------------------------|---|
| Ma, <i>et al.</i> (2023) | USA | Mental health | Healthcare consumer experience analysis | – | To explore the potential of generative AI in conversational agents for mental health support through a qualitative analysis of 2917 Reddit comments from 462 users about the mental health app, Replika. | Replika | The analysis of the Reddit posts revealed that users of Replika found the app to provide non-judgmental support on demand, which could aid in boosting confidence and self-discovery. However, Replika also produced both violent and sexual content, and had problems sustaining consistent communication and retaining new information. |
| Salah <i>et al.</i> (2023) | Oman | Mental health | Online survey | 732 participants | To investigate the associations of user perception of ChatGPT with psychological well-being and the impact of stereotypes in generative AI on user well-being. | ChatGPT | Associations were found between user perception of stereotyping by ChatGPT and self-esteem. |
| Blease <i>et al.</i> (2024) | USA | Mental health | Online survey | 138 psychiatrists | To understand the opinion of psychiatrists on the use of generative AI. | ChatGPT, Bard, and Bing.com | The majority of psychiatrists were interested in the potential of generative AI to reduce the burden of documentation and administration, and were under the impression that most of their patients ‘will consult these tools before first seeing a doctor’, raising concern over patient privacy. |

image of a field in rapid expansion. Indeed, most studies included in this review were pilot studies or feasibility studies exploring potential use cases, investigating user perceptions, or identifying potential ethical and safety concerns of prospective generative AI tools.

The relative immaturity of the field is evident in the absence of consensus on the definition of AI and generative AI in the studies screened as part of this review. The term 'AI' is used very loosely, often simply to describe a classification model. The majority of studies excluded based on the type of intervention were claiming to be 'powered by AI' which meant having a classification model tag, e.g., the sentiment of free-text input, which would then, in turn, trigger a pre-specified response. While this might fall under the broadest definition of generative AI, as the input does result in a textual output, we deemed it necessary to narrow our definition of generative AI to only include content generated in a less deterministic/preestablished manner (e.g., as seen in transformer and diffusion models such as those empowering ChatGPT, DALL-E, Sora and their equivalents).

Most of the identified studies focused on natural language implementations of generative AI, particularly ChatGPT, either by testing its psychiatric knowledge base or evaluating its capabilities as a mental health conversational companion. Though most of the included studies found that generative AI performed well at various tasks, some studies also highlighted potential safety issues. I.e., due to the inherent lack of predefinition in generative AI output, responses cannot be reliably predicted, and, thus, protection from ethical and safety breaches cannot be guaranteed. For these reasons, it is crucial for users, patients, practitioners, and their organisations to carefully consider and scrutinise the legal and ethical aspects of using generative AI.

While we did not conduct a formal quality assessment of the studies included in the review (a large proportion of studies were too preliminary/informal to allow for such assessment), it was our impression that many studies were of relatively low quality and had limited clinical relevance. Specifically, most studies were severely underspecified, both in terms of technology used (such as the type and version of models) and study design (e.g., specification of specific prompts), limiting reproducibility. Additionally, although many studies could be considered pilot studies, their results were often overgeneralised and overstated beyond what could reasonably be claimed from the results. Therefore, to advance the field of generative AI for mental health we propose the following guidelines for future research: First, to facilitate reproducibility and clarity of findings, we highly recommend studies to follow a set of reporting guidelines for generative AI, such as TRIPOD-LLM, to ensure that all relevant items are reported (Gallifant *et al.*, 2024). Second, we encourage the field to move beyond simple 'knowledge testing' and prompt experiments and towards rigorously planned clinical trials involving users/patients and tasks with greater clinical relevance. Indeed, it is noteworthy that only a handful of studies recruited participants to interact with the technology, while even fewer structured the interaction (intervention) in a systematic manner. Also, future studies should ideally take the user/patient perspective into account in the design phase (i.e., co-design).

While several studies deemed the responses from generative AI to be clear and in accordance with scientific knowledge, some studies found that generative AI underestimates the risk of e.g. suicide (Haman and Školník, 2023; Heston, 2023) and handles crisis scenarios in an less than ideal manner (Heston, 2023). Therefore, it is essential that chatbots developed for mental health/

patient support ensure adequate handling of all levels of illness/symptom severity – including suicidal ideation.

This study should be interpreted in light of its limitations. First, the field is in its nascence and tangible new developments may happen quickly. This review merely represents a snapshot of the state of the field as of February 23, 2024, and new developments are likely to have emerged since the data collection concluded. Second, we implemented a broad search strategy; however, we cannot rule out the possibility that some relevant studies may have been overlooked. Third, it was not feasible to do a quantitative analysis due to heterogeneity of the studies. Fourth, while the literature identified in this review predominantly emphasised the clinical/care potential of generative AI in the context of mental health/psychiatry (likely due to the databases used for the search), it is apparent that there are important legal/ethical challenges that need to be addressed. An exhaustive review of the literature on these challenges would require a broader search strategy than employed here.

In conclusion, the field of generative AI in psychiatry and mental health is in its infancy, though evolving and growing exponentially. Unfortunately, many of the identified studies investigating the potential of generative AI in the context of mental health/psychiatry were poorly specified (particularly with regard to the methods). Therefore, moving forward, we suggest that studies using generative AI in psychiatric settings should aim for more transparency of methods, experimental designs (including clinical trials), clinical relevance, and user/patient inclusion in the design phase.

Acknowledgements. The authors are grateful to librarian Helene Sognstrup (Royal Danish Library) for her assistance with the search strategy and to Arnault-Quentin Vermillet (Aarhus University) and Jean-Christophe Philippe Debost (Aarhus University Hospital – Psychiatry) for translation from French.

Author contribution. Conception and design: SDØ, RML, and SK. Provision of study data: SDØ. Screening of data: SK and RML. Data analysis: SK, LH, and RML. Interpretation: All authors. Manuscript writing: All authors. Final approval of the manuscript: All authors.

Financial support. There was no specific funding for this study. Outside this study, SDØ is supported by the Novo Nordisk Foundation (grant number: NNF20SA0062874), the Lundbeck Foundation (grant numbers: R358-2020-2341 and R344- 2020-1073), the Danish Cancer Society (grant number: R283-A16461), the Central Denmark Region Fund for Strengthening of Health Science (grant number: 1-36-72-4-20), The Danish Agency for Digitisation Investment Fund for New Technologies (grant number 2020-6720), and Independent Research Fund Denmark (grant number: 7016-00048B and 2096-00055A).

Competing interests. SDØ received the 2020 Lundbeck Foundation Young Investigator Prize. SDØ owns/has owned units of mutual funds with stock tickers DKIGI, IAIMWC, SPIC25 KL and WEKAFKI, and owns/has owned units of exchange traded funds with stock tickers BATE, TRET, QDV5, QDVH, QDVE, SADM, IQQH, USPY, EXH2, 2B76, IS4S, OM3X and EUNL. The remaining authors report no conflicts of interest.

References

- Alanezi F (2024) Assessing the effectiveness of chatGPT in delivering mental health support: a qualitative study. *Journal of Multidisciplinary Healthcare* 17, 461–471. DOI: [10.2147/JMDH.S447368](https://doi.org/10.2147/JMDH.S447368).
- Almusharraf F, Rose J and Selby P (2020) Engaging unmotivated smokers to move toward quitting: design of motivational interviewing-based chatbot through iterative interactions. *Journal of Medical Internet Research* 22(11), e20251. DOI: [10.2196/20251](https://doi.org/10.2196/20251).

- American Psychiatric Association** (2013) *Diagnostic and statistical manual of mental disorders*, 5th edn. Washington, DC: American Psychiatric Publishing.
- Amin S, Kawamoto CT and Pokhrel P** (2023) Exploring the chatGPT platform with scenario-specific prompts for vaping cessation. *Tobacco Control*. <https://pubmed.ncbi.nlm.nih.gov/37460216/>.
- Archambault D and Kouroupetroglou G** (2023) AI supporting AAC pictographic symbol adaptations. *Studies in Health Technology and Informatics* **306**, 215–221. DOI: [10.3233/shti230622](https://doi.org/10.3233/shti230622).
- Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, Faix DJ, Goodman AM, Longhurst CA, Hogarth M, Smith DM** (2023) Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine* **183**(6), 589–596. DOI: [10.1001/jamainternmed.2023.1838](https://doi.org/10.1001/jamainternmed.2023.1838).
- Beilharz F, Sukunesan S, Rossell SL, Kulkarni J, Sharp G** (2021) Development of a Positive Body Image Chatbot (KIT) With Young People and Parents/Carers: Qualitative Focus Group Study. *Journal of Medical Internet Research* **23**(6), e27807. DOI: [10.2196/27807](https://doi.org/10.2196/27807).
- Blease C, Worthen A and Torous J** (2024) Psychiatrists' experiences and opinions of generative artificial intelligence in mental healthcare: An online mixed methods survey. *Psychiatry Research* **333**, 115724. DOI: [10.1016/j.psychres.2024](https://doi.org/10.1016/j.psychres.2024).
- Carpenter KA and Altman RB** (2023) Using GPT-3 to build a lexicon of drugs of abuse synonyms for social media pharmacovigilance. *Biomolecules* **13**(2), 387. DOI: [10.3390/biom13020387](https://doi.org/10.3390/biom13020387).
- de Leon J and De Las Cuevas C** (2023) Will ChatGPT substitute for us as clozapine experts? *Journal of Clinical Psychopharmacology* **43**(5), 400. DOI: [10.1097/JCP.0000000000001734](https://doi.org/10.1097/JCP.0000000000001734).
- Denecke K, Hochreutener S, Pöpel A, May R** (2018) Self-anamnesis with a conversational user interface: concept and Usability study. *Methods of Information in Medicine* **57**(05/06), 243–252. DOI: [10.1055/s-0038-1675822](https://doi.org/10.1055/s-0038-1675822).
- Dergaa I, Fekih-Romdhane F, Hallit S, Loch AA, Glenn JM, Fessi MS, Ben Aissa M, Souissi N, Guelmami N, Swed S, El Omri A, Bragazzi NL and Ben Saad H** (2024) ChatGPT is not ready yet for use in providing mental health assessment and interventions. *Frontiers in Psychiatry* **14**, 1277756. DOI: [10.3389/fpsy.2023.1277756](https://doi.org/10.3389/fpsy.2023.1277756).
- De Freitas J, Uğuralp AK, Oğuz-Uğuralp Z, Puntoni S** (2024) Chatbots and mental health: insights into the safety of generative AI. *Journal of Consumer Psychology* **34**(3), 481–491. DOI: [10.1002/jcpsy.1393](https://doi.org/10.1002/jcpsy.1393).
- D'Souza RF, Amanullah S, Mathew M and Surapaneni KM** (2023) Appraising the performance of chatGPT in psychiatry using 100 clinical case vignettes. *Asian Journal of Psychiatry* **89**, 103770. DOI: [10.1016/j.ajp.2023.103770](https://doi.org/10.1016/j.ajp.2023.103770).
- Else H** (2023) Abstracts written by chatGPT fool scientists. *Nature* **613**(7944), 423–423. DOI: [10.1038/d41586-023-00056-7](https://doi.org/10.1038/d41586-023-00056-7).
- Elyoseph Z, Hadar-Shoval D, Asraf K and Lvovsky M** (2023) ChatGPT outperforms humans in emotional awareness evaluations. *Frontiers in Psychology* **14**, 1199058. DOI: [10.3389/fpsyg.2023.1199058](https://doi.org/10.3389/fpsyg.2023.1199058).
- Elyoseph Z and Levkovich I** (2023) Beyond human expertise: the promise and limitations of chatGPT in suicide risk assessment. *Frontiers in Psychiatry* **14**, 1213141. DOI: [10.3389/fpsy.2023.1213141](https://doi.org/10.3389/fpsy.2023.1213141).
- Elyoseph Z, Levkovich I and 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**(Suppl 1), e002583. DOI: [10.1136/fmch-2023-002583](https://doi.org/10.1136/fmch-2023-002583).
- Elyoseph Z, Refoua E, Asraf K, Lvovsky M, Shimoni Y and Hadar-Shoval D** (2024) Capacity of generative AI to interpret human emotions from visual and textual data: pilot evaluation study. *JMIR Mental Health* **11**(1), e54369. DOI: [10.2196/54369](https://doi.org/10.2196/54369).
- Galido PV, Butala S, Chakerian M and Agustines D** (2023) A case study demonstrating applications of chatGPT in the clinical management of treatment-resistant schizophrenia. *Cureus* **15**(4), e38166. DOI: [10.7759/cureus.38166](https://doi.org/10.7759/cureus.38166).
- Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, Demner-Fushman D, Dligach D, Daneshjou R, Fernandes C, Hansen LH, Landman A, Lehmann L, McCoy LG, Miller T, Moreno A, Munch N, Restrepo D, Savova G, Umeton R, Gichoya JW, Collins GS, Moons KGM, Celi LA and Bitterman DS, .** (2024) The TRIPOD-LLM statement: a targeted guideline for reporting large language models use, medRxiv. DOI: [10.1101/2024.07.24.24310930](https://doi.org/10.1101/2024.07.24.24310930).
- Gao Y, Dligach D, Christensen L, Tesch S, Laffin R, Xu D, Miller T, Uzuner O, Churpek MM and Afshar M** (2022) A scoping review of publicly available language tasks in clinical natural language processing. *Journal of the American Medical Informatics Association* **29**(10), 1797–1806. DOI: [10.1093/jamia/ocac127](https://doi.org/10.1093/jamia/ocac127).
- Gifu D and Pop E** (2022) Smart solutions to keep your mental balance. *Procedia Computer Science* **214**, 503–510. DOI: [10.1016/j.procs.2022.11.205](https://doi.org/10.1016/j.procs.2022.11.205).
- Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim H-C, Jeste DV** (2019) Artificial intelligence for mental health and mental illnesses: an overview. *Current Psychiatry Reports* **21**(11), 116. DOI: [10.1007/s11920-019-1094-0](https://doi.org/10.1007/s11920-019-1094-0).
- Hadar-Shoval D, Elyoseph Z and Lvovsky M** (2023) The plasticity of chatGPT's mentalizing abilities: personalization for personality structures. *Frontiers in Psychiatry* **14**, 1–234397. DOI: [10.3389/fpsy.2023](https://doi.org/10.3389/fpsy.2023).
- Haman M and Školnik M** (2023) Behind the chatGPT hype: are its suggestions contributing to addiction? *Annals of Biomedical Engineering* **51**(6), 1128–1129. DOI: [10.1007/s10439-023-03201-5](https://doi.org/10.1007/s10439-023-03201-5).
- Hamilton M** (1959) The assessment of anxiety states by rating. *British Journal of Medical Psychology* **32**(1), 50–55. DOI: [10.1111/j.2044-8341.1959.tb00467.x](https://doi.org/10.1111/j.2044-8341.1959.tb00467.x).
- Hamilton M** (1960) A rating scale for depression. *Journal of Neurology, Neurosurgery & Psychiatry* **23**(1), 56–62.
- Hansen L, Enevoldsen KC, Bernstorff M, Nielbo KL, Danielsen AA and Østergaard SD** (2021) The PSYchiatric clinical outcome prediction (PSYCOP) cohort: leveraging the potential of electronic health records in the treatment of mental disorders. *Acta Neuropsychiatrica* **33**(6), 323–330. DOI: [10.1017/neu.2021.22](https://doi.org/10.1017/neu.2021.22).
- Haug CJ and Drazen JM** (2023) Artificial intelligence and machine learning in clinical medicine, 2023. *New England Journal of Medicine* **388**(13), 1201–1208. DOI: [10.1056/NEJMra2302038](https://doi.org/10.1056/NEJMra2302038).
- Heinz MV, Bhattacharya S, Trudeau B, Quist R, Song SH, Lee CM and Jacobson NC** (2023) Testing domain knowledge and risk of bias of a large-scale general artificial intelligence model in mental health. *Digit Health* **9**, 20552076231170499. DOI: [10.1177/20552076231170499](https://doi.org/10.1177/20552076231170499).
- Herrmann-Werner A, Festl-Wietek T, Holderried F, Herschbach I, Griewatz J, Masters K, Zipfel S, Mahling M** (2024) Assessing chatGPT's mastery of bloom's taxonomy using psychosomatic medicine exam questions: mixed-methods study. *Journal of Medical Internet Research* **26**(1), e52113. DOI: [10.2196/52113](https://doi.org/10.2196/52113).
- Heston TF** (2023) Safety of large language models in addressing depression. *Cureus* **15**. <https://pubmed.ncbi.nlm.nih.gov/38111813/>.
- Hristidis V, Ruggiano N, Brown EL, Ganta SRR and Stewart S** (2023) ChatGPT vs google for queries related to dementia and other cognitive decline: comparison of results. *Journal of Medical Internet Research* **25**, e48966. DOI: [10.2196/48966](https://doi.org/10.2196/48966).
- Hu K and Hu K** (2023) ChatGPT sets record for fastest-growing user base - analyst note. Reuters, 2 February. Available at: <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/> (Accessed: 9 September 2024).
- Hwang G, Lee DY, Seol S, Jung J, Choi Y, Her ES, An MH and Park RW** (2024) Assessing the potential of chatGPT for psychodynamic formulations in psychiatry: An exploratory study. *Psychiatry Research* **331**, 115655. DOI: [10.1016/j.psychres.2023.115334](https://doi.org/10.1016/j.psychres.2023.115334).
- Kay SR, Fiszbein A and Opler LA** (1987) The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin* **13**(2), 261–276. DOI: [10.1093/schbul/13.2.261](https://doi.org/10.1093/schbul/13.2.261).
- Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V and Dagan A** (2023) Performance of chatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digital Health* **2**(2), e0000198. DOI: [10.1371/journal.pdig.0000198](https://doi.org/10.1371/journal.pdig.0000198).
- Lee P, Bubeck S and Petro J** (2023) Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine* **388**(13), 1233–1239. DOI: [10.1056/NEJMsr2214184](https://doi.org/10.1056/NEJMsr2214184).
- Li H, Zhang R, Lee Y-C, Kraut RE and Mohr DC** (2023) Systematic review and meta-analysis of AI-based conversational agents for promoting mental

- health and well-being. *npj Digital Medicine* 6(1), 1–14. DOI: [10.1038/s41746-023-00979-5](https://doi.org/10.1038/s41746-023-00979-5).
- Lingjærde O, Ahlfors UG, Bech P, Dencker SJ and Elgen K (1987) The UKU side effect rating scale: a new comprehensive rating scale for psychotropic drugs and a cross-sectional study of side effects in neuroleptic-treated patients. *Acta Psychiatrica Scandinavica* 76(Suppl 334), 100–100. DOI: [10.1111/j.1600-0447.1987.tb10566.x](https://doi.org/10.1111/j.1600-0447.1987.tb10566.x).
- Luka, Inc (2024) Replika, Available at: <https://replika.com> (Accessed: 13 September 2024).
- Lundin RM, Berk M and Østergaard SD (2023) ChatGPT on ECT: can large language models support psychoeducation? *The Journal of ECT* 39(3), 130–133. DOI: [10.1097/YCT.0000000000000941](https://doi.org/10.1097/YCT.0000000000000941).
- Luyck JJ, Gerritse F, Habets PC and Vinkers CH (2023) The performance of chatGPT in generating answers to clinical questions in psychiatry: a two-layer assessment. *World Psychiatry* 22(3), 479–480. DOI: [10.1002/wps.21145](https://doi.org/10.1002/wps.21145).
- Ma Z, Mei Y and Su Z (2023) Understanding the Benefits and Challenges of Using Large Language Model-based Conversational Agents for Mental Well-being Support. In: AMIA Annual Symposium Proceedings, 1105–1114.
- McFayden TC, Bristol S, Putnam O and Harrop C (2024) ChatGPT: artificial intelligence as a potential tool for parents seeking information about autism. *Cyberpsychology, Behavior, and Social Networking* 27(2), 135–148. DOI: [10.1089/cyber.2023.0202](https://doi.org/10.1089/cyber.2023.0202).
- McGowan A, Gui Y, Dobbs M, Shuster S, Cotter M, Selloni A, Goodman M, Srivastava A, Cecchi GA and Corcoran CM (2023) ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry Research* 326, 115334. <https://pubmed.ncbi.nlm.nih.gov/37499282/>.
- Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine* 6(7), e1000097. DOI: [10.1371/journal.pmed.1000097](https://doi.org/10.1371/journal.pmed.1000097).
- Nadkarni PM, Ohno-Machado L and Chapman WW (2011) Natural language processing: an introduction. *Journal of the American Medical Informatics Association* 18(5), 544–551. DOI: [10.1136/amiainl-2011-000464](https://doi.org/10.1136/amiainl-2011-000464).
- OpenAI (2023) DALL-E. 2. Available at: <https://openai.com/dall-e-2> (Accessed: 30 May 2023).
- OpenAI (2024a) ChatGPT. Available at: <https://chatgpt.com/>. (Accessed: 9 September 2024).
- OpenAI (2024b) Sora: creating video from text. Available at: <https://openai.com/sora/>, (Accessed: 10 September 2024).
- Østergaard SD (2024) Can generative artificial intelligence facilitate illustration of- and communication regarding hallucinations and delusions? *Acta Psychiatrica Scandinavica* 149(6), 441–444. DOI: [10.1111/acps.13680](https://doi.org/10.1111/acps.13680).
- Østergaard SD and Nielbo KL (2023) False responses from artificial intelligence models are not hallucinations. *Schizophrenia Bulletin* 49(5), 1105–1107. DOI: [10.1093/schbul/sbad068](https://doi.org/10.1093/schbul/sbad068).
- Parker G and Spoelma MJ (2024) A chat about bipolar disorder. *Bipolar Disorders* 26(3), 249–254. DOI: [10.1111/bdi.13379](https://doi.org/10.1111/bdi.13379).
- Prada P, Perroud N and Thorens G (2023) [Artificial intelligence and psychiatry: questions from psychiatrists to ChatGPT]. *Revue Médicale Suisse* 19(818), 532–536. DOI: [10.53738/revmed.2023.19.818.532](https://doi.org/10.53738/revmed.2023.19.818.532).
- Randhawa J and Khan A (2023) A conversation with chatGPT about the Usage of lithium in pregnancy for bipolar disorder. *Cureus* 15. <https://pubmed.ncbi.nlm.nih.gov/37933339/>.
- Rudan D, Marčinko D, Degmečić D and Jakšić N (2023) Scarcity of research on psychological or psychiatric states using validated questionnaires in low- and middle-income countries: a ChatGPT-assisted bibliometric analysis and national case study on some psychometric properties. *Journal of Global Health* 13, 04102. DOI: [10.7189/jogh.13.04102](https://doi.org/10.7189/jogh.13.04102).
- Sabour S, Zhang W, Xiao X, Zhang Y, Zheng Y, Wen J, Zhao J and Huang M (2023) A chatbot for mental health support: exploring the impact of Emohaa on reducing mental distress in China, *Front Digit Health* 20230504th, DOI: [10.3389/fgth.2023.1133987](https://doi.org/10.3389/fgth.2023.1133987).
- Salah M, Alhalbusi H, Ismail MM and Abdelfattah F (2023) Chatting with chatgpt: decoding the mind of chatbot users and unveiling the intricate connections between user perception, trust and stereotype perception on self-esteem and psychological well-being. *Current Psychology* 43(9), 7843–7858. DOI: [10.1007/s12144-023-04989-0](https://doi.org/10.1007/s12144-023-04989-0).
- Schumacher E, Rosenthal D, Nair V, Price L, Tso G and Kannan A (2023) Extrinsicly-Focused Evaluation of Omissions in Medical Summarization. Available at: <https://doi.org/10.48550/arXiv.2311.08303>.
- Sezgin E, Chekeni F, Lee J and Keim S (2023) Clinical accuracy of large language models and google search responses to postpartum depression questions: cross-sectional study. *Journal of Medical Internet Research* 25(1), e49240. DOI: [10.2196/49240](https://doi.org/10.2196/49240).
- Smith A, Hachen S, Schleifer R, Bhugra D, Buadze A and Liebrez M (2023) Old dog, new tricks? Exploring the potential functionalities of chatGPT in supporting educational methods in social psychiatry. *International Journal of Social Psychiatry* 69(8), 1882–1889. DOI: [10.1177/00207640231178451](https://doi.org/10.1177/00207640231178451).
- Sohl-Dickstein J, Weiss E, Maheswaranathan N and Ganguli S (2015) Deep Unsupervised Learning using Nonequilibrium Thermodynamics. International conference on machine learning, 2256–2265.
- Spallek S, Birrell L, Kershaw S, Devine EK, Thornton L (2023) Can we use chatGPT for mental health and substance use education? Examining its quality and potential harms. *JMIR Medical Education* 9(1), e51243. DOI: [10.2196/51243](https://doi.org/10.2196/51243).
- Takefuji Y (2023) Impact of COVID-19 on mental health in the US with generative AI. *Asian Journal of Psychiatry* 88, 103736 DOI: [10.1016/j.ajp.2023.103736](https://doi.org/10.1016/j.ajp.2023.103736).
- Vaidyam AN, Linggonogoro D and Torous J (2021) Changes to the psychiatric chatbot landscape: a systematic review of conversational agents in serious mental illness. *The Canadian Journal of Psychiatry / La Revue canadienne de psychiatrie* 66(4), 339–348. DOI: [10.1177/0706743720966429](https://doi.org/10.1177/0706743720966429).
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L and Polosukhin I (2017) Attention is all you need. *Advances in Neural Information Processing Systems*. 30. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
- Veritas Health Innovation (2024). Covidence systematic review software. Melbourne, Australia: Veritas Health Innovation. Available at: www.covidence.org.
- Wang R, Feng H and Wei G-W (2023) ChatGPT in drug discovery: a case study on anticocaine addiction drug development with chatbots. *Journal of Chemical Information and Modeling* 63(22), 7189–7209. DOI: [10.1021/acs.jcim.3c01429](https://doi.org/10.1021/acs.jcim.3c01429).
- Woodnutt S, Allen C, Snowden J, Flynn M, Hall S, Libberton P and Purvis F (2024) Could artificial intelligence write mental health nursing care plans? *Journal of Psychiatric and Mental Health Nursing*. (1), 79–86. DOI: [10.1111/jpm.12965](https://doi.org/10.1111/jpm.12965).
- World Health Organization (1993) *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*. Geneva: World Health Organization.