BMJ Mental Health

ChatGPT and mental healthcare: balancing benefits with risks of harms

Charlotte Blease [1,2], John Torous [3]

[1]Participatory eHealth and Health Data Research Group, Department of Women's and Children's Health, Uppsala Universitet, Uppsala, Sweden
[2]Digital Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts, USA
[3]Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts, USA

Correspondence to
Dr Charlotte Blease, Participatory eHealth and Health Data Research Group, Department of Women's and Children's Health, Uppsala Universitet, Uppsala, Sweden; charlotte.blease@uu.se

Received 15 September 2023
Accepted 6 October 2023

To cite: Blease C, Torous J. *BMJ Ment Health* 2023;**26**:1–3.

ABSTRACT

Against the global need for increased access to mental services, health organisations are looking to technological advances to improve the delivery of care and lower costs. Since November 2022, with the public launch of OpenAI's ChatGPT, the field of generative artificial intelligence (AI) has received expanding attention. Although generative AI itself is not new, technical advances and the increased accessibility of large language models (LLMs) (eg, OpenAI's GPT-4 and Google's Bard) suggest use of these tools could be clinically significant. LLMs are an application of generative AI technology that can summarise and generate content based on training on vast data sets. Unlike search engines, which provide internet links in response to typed entries, chatbots that rely on generative language models can simulate dialogue that resembles human conversations. We examine the potential promise and the risks of using LLMs in mental healthcare today, focusing on their scope to impact mental healthcare, including global equity in the delivery of care. Although we caution that LLMs should not be used to disintermediate mental health clinicians, we signal how—if carefully implemented—in the long term these tools could reap benefits for patients and health professionals.

CHATGPT AND MENTAL HEALTHCARE: BALANCING BENEFITS WITH RISKS OF HARMS

The WHO estimates that worldwide one in eight people live with mental illness.[1] Stigmatisation and human rights violations, combined with lack of resources, including shortfalls in mental health professionals, pose significant barriers to psychiatric care.[2] Consequently, clinician time is one of the scarcest resources in healthcare, and psychiatrists report looking to advances in artificial intelligence (AI) to improve efficiencies and assist with administrative tasks.[3]

Considering these challenges, recent advances in the field of generative AI and its potential to impact healthcare delivery have received considerable attention. Unlike search engines, which provide internet links in response to typed entries, a new generation of chatbots, such as OpenAI's GPT-4, powered by large language models (LLMs) offer responses that resemble conversations. LLMs use massive amounts of past data to predict the next word in a sequence. This probabilistic process combined with other technical advances means these models have aptitude in recognising, summarising and generating content.

On the face of it, these tools offer considerable promise to clinicians. Already in June 2023, in a Medical Economics survey in the USA, more than 1 in 10 clinicians already reported adopting chatbots such as ChatGPT, and nearly 50% expressed future intent to use these technologies for data entry, medical scheduling or research.[4] From the patient side, given the widespread global use of internet-enabled devices and the ease of access of LLM-powered chatbots, people with stigmatised mental health conditions may be especially inclined to adopt these tools. In light of these rapid advances, we examine the potential promise and the risks in mental healthcare and offer suggestions to ensure the enormous scope of these innovations is effectively and ethically harnessed.

The potential benefits of generative AI

While AI continues to evolve, some features appear nearly ready for use today. These include decreasing administrative burdens and improving documentation and clinical hypothesis generation. Surveys show that psychiatrists desire and anticipate assistance from advances in AI in undertaking administrative tasks,[3] and the rapidity with which LLMs generate narrative summaries of complex data strongly suggests potential to reduce work burdens, including updating clinical records.

Preliminary studies also suggest that LLMs can assist with writing empathic documentation. For example, a study comparing written responses of physicians and ChatGPT with 195 real-world health questions submitted to Reddit's AskDocs reported that ChatGPT's responses were, on average, four times longer.[5] In addition, a panel of blinded physicians rated ChatGPT responses as 'good' or 'very good' nearly four times more often than those submitted by doctors and rated the chatbot's responses almost 10 times more empathic than responses by doctors.

Other studies suggest chatbots powered by LLMs could assist mental health peers or clinicians in offering consistently high levels of support in patient-facing interactions, including those struggling with compassion fatigue. For example, a randomised controlled trial of responses submitted to TalkLife, a social media platform that offers peer support to mental health patients, found responses that were written in collaboration with a chatbot called 'HAILEY', short for '*H*uman-*AI* co*L*laboration approach for *E*mpath*Y*', were more likely to be rated as empathic than human-only responses.[6] Peer supporters who self-identified as struggling to offer empathic support were significantly rated as

more likely to provide empathic responses in the AI-in-the-loop scenario.

Aside from clinician documentation and patient-facing interactions, a key emerging strength of generative AI is hypothesis generation. Encouragingly, preliminary studies show the promise of GPT-4 in generating accurate lists of differential diagnoses, including in complicated clinical cases,[7] suggesting their potential to facilitate brainstorming in diagnostic and treatment decision-making.

The potential harms of generative AI

Using generative AI in mental healthcare also risks harm. LLMs are autoregressive, meaning they use past data to predict future data in generating responses; this probabilistic process means outputs can be inconsistent, often changing depending on the wording of queries. Most LLMs are not exclusively trained on medical texts and lack the capacity to discriminate the quality of webtext from which they draw their responses, meaning inferior content is treated the same way as reliable material, leading to risks of harm. For example, transcripts show that a chatbot encouraged a Belgian man to end his life to help stop climate change.[8]

As noted, LLMs have gained a fast reputation for their capacity to follow requests, such as writing responses in a requested conversational style, tone or literacy level. However, for a variety of reasons, biases are baked in, leading to the potential for 'algorithmic discrimination' whereby outputs may perpetuate or exacerbate unfair treatment,[9] and research shows these models can embed gender, race and disability biases, threatening their equitable application.[10 11] The sources of bias are multiple, the training data including omissions in clinical populations in medical publications (eg, PubMed), as well as stereotyping arising within social media (eg, Twitter/X, Reddit, Facebook), books, news media and images.[12] Human and societal biases can also be introduced via supervised learning techniques whereby workers who are very often poorly paid risk entrenching unwanted stereotypes via data labelling and feedback.

An additional source of harm is the tendency for LLMs to make up patently false information, referred to as 'hallucinations'.[13] This risk, combined with the sheer speed and authoritative nature of the conversational responses offered by LLM-powered chatbots, might render clinicians and patients more vulnerable to disinformation, risking safety. Relatedly, if patients are unaware that it is a chatbot rather than a human answering their queries, this could compromise patient trust. For example, in January 2023, a company called Koko publicly apologised for using ChatGPT to write emotional responses while deceiving users that the responses were generated by humans.[14]

Yet, because of the conversational fluency associated with chatbots such as OpenAI's GPT-4, patients and clinicians may become too trusting and be tempted to input sensitive patient data to solicit seemingly 'neutral' advice or recommendations, risking patient privacy. Earlier this year, the American Medical Association issued an advisory cautioning that ChatGPT and other LLM tools are not regulated and that clinicians should avoid entering patient data into generative AI systems.[15] Combined with the potential for data triangulation, without additional safeguarding measures, patients may lose control of their confidential health information.

Suggestions to enhance the ethical and effective use of generative AI

Many mental health professionals may already be concerned about the potential harms of generative AI. In some cases, the benefits of employing these models could outweigh the risks, so long as they are implemented appropriately. Several suggestions could be considered.

First, health systems must ensure that LLMs uphold or improve current standards of patient safety, and relatedly that these tools do not perpetuate or compound current inequities in the delivery of care. To this end, robust experimental work with both prompt engineering and tuning of the underlying models is first needed to establish clinical accuracy and quality of LLM responses. For example, to investigate the potential for biases, LLMs could be tested to explore the range of differential diagnoses and the quality of clinical notes created for different patient populations.

Reducing the potential for algorithmic discrimination will require a multilevel approach.[10 12] Thorough attention must be given to the quality of data fed into LLMs, the potential for bias in human agents involved in labelling and training AI, and therefore the diversity of participants involved in shaping these technologies.[16] Participatory design approaches, involving marginalised voices, including but not limited to those from low-income countries and patients from mental health communities, should be fully integrated into activities relating to the development and impact of these tools.

Second, clinicians and patients could be supported with resources and guidance about the limitations and benefits of using LLMs. These tools should not be used to replace clinician judgement, be relied on to complete documentation or fully substitute for human interactions. Instead, via medical curricula and ongoing professional training, clinicians could be supported in how these tools could be used to augment human capacities that should be overseen by clinicians.

Third, civic health professional and regulatory engagement will also be needed to review privacy concerns related to sensitive patient data in developing and using LLM chatbots.[17] In the USA, there are already efforts to integrate generative AI services into electronic healthcare systems that comply with the privacy standards of the 1996 Health Insurance Portability and Accountability Act.[18] In the European Union, under the General Data Protection Regulation, strong reasons must be given to process patient data without informed consent, such as for public health justifications, and authorities are currently reviewing whether OpenAI complies with this regulation.[19] Patients should be fully engaged in debates about how their health data are managed; decisions about when thresholds of acceptable use might be met should be informed by patients.

CONCLUSIONS

Generative AI in mental healthcare has the potential to offer significant benefits in clinical documentation, patient communication and medical decision-making. However, to minimise the risks of harm, these tools need to be more thoroughly studied and monitored. To this end and for psychiatrists to be equipped to lead policy and practice advances on the role of LLM in mental healthcare, improvements in digital education will be imperative.

Twitter Charlotte Blease @crblease and John Torous @JohnTorousMD

Contributors CB wrote the first draft. JT and CB revised the paper until both signed off on it.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Charlotte Blease http://orcid.org/0000-0002-0205-1165
John Torous http://orcid.org/0000-0002-5362-7937

REFERENCES

1 World Health Organization. Mental disorders. 2022. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders [Accessed 14 Sep 2023].
2 Mental Health and Substance Use Team, World Health Organization. World mental health report: transforming mental health for all. 2022. Available: https://www.who.int/publications/i/item/9789240049338 [Accessed 14 Sep 2023].
3 Blease C, Locher C, Leon-Carlyle M, *et al*. Artificial intelligence and the future of psychiatry: qualitative findings from a global physician survey. *Digit Health* 2020;6.
4 Shryock T. AI special report: what patients and doctors really think about AI in health care. In: *Medical Economics*. 2023. Available: https://www.medicaleconomics.com/view/ai-special-report-what-patients-and-doctors-really-think-about-ai-in-health-care [accessed 22 Aug 2023].
5 Ayers JW, Poliak A, Dredze M, *et al*. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med* 2023;183:589–96.
6 Sharma A, Lin IW, Miner AS, *et al*. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nat Mach Intell* 2023;5:46–57.
7 Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA* 2023;330:78–80.
8 El Atillah I. Man ends his life after an AI Chatbot 'encouraged' him to sacrifice himself to stop climate change. EuroNewsNext; 2023. Available: https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate [Accessed 11 Sep 2023].
9 Teno JM. Garbage in, garbage out-words of caution on big data and machine learning in medical practice. *JAMA Health Forum* 2023;4:e230397.
10 Gross N. What chatGPT tells us about gender: a cautionary tale about performativity and gender biases in AI. *Social Sciences* 2023;12:435.
11 King M. Harmful biases in artificial intelligence. *Lancet Psychiatry* 2022;9.
12 Ferrara E. Should chatGPT be biased? Challenges and risks of bias in large language models. 2023.
13 Alkaissi H, McFarlane SI. Artificial hallucinations in chatGPT: implications in scientific writing. *Cureus* 2023;15:e35179.
14 Ingram D. A mental health Tech company ran an AI experiment on real users. Nothing's stopping apps from conducting more. NBC News; 2023. Available: https://www.nbcnews.com/tech/internet/chatgpt-ai-experiment-mental-health-tech-app-koko-rcna65110 [Accessed 13 Aug 2023].
15 AMA. ChatGPT and generative AI: what physicians should consider. American Medical Association; 2023. Available: https://www.ama-assn.org/system/files/chatgpt-what-physicians-should-consider.pdf [Accessed 11 Sep 2023].
16 Birhane A, ChatGPT RD. Galactica, and the progress trap; 2022. Wired
17 Marks M, Haupt CE. AI Chatbots, health privacy, and challenges to HIPAA compliance. *JAMA* 2023;330:309–10.
18 Adams K. Epic to integrate GPT-4 into its EHR through expanded Microsoft partnership. MedCity News; 2023. Available: https://medcitynews.com/2023/04/epic-to-integrate-gpt-4-into-its-ehr-through-expanded-microsoft-partnership/ [Accessed 31 Jul 2023].
19 Minssen T, Vayena E, Cohen IG. The challenges for regulating medical use of chatGPT and other large language models. *JAMA* 2023;330:315–6.