Trends
in Psychiatry and Psychotherapy

JOURNAL ARTICLE PRE-PROOF
(as accepted)

Editorial

Artificial Intelligence in Mental Health Care: Less Hype, More Accountability

Felipe Dalvi-Garcia, Evellin Cristine Cardoso, Laiana Azevedo Quagliato, Antonio Egidio Nardi

http://doi.org/10.47626/2237-6089-2025-1203

Original submitted Date: 17-Sep-2025

Accepted Date: 16-Nov-2025

Artificial Intelligence in Mental Health care: Less Hype, More Accountability

Short Title: Accountability in AI for mental health care

Felipe Dalvi-Garcia[a]*, Evellin Cristine Cardoso[b], Laiana Azevedo Quagliato[a], Antonio Egidio Nardi[a]

[a]Institute of Psychiatry, Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil.
[b]Informatics Institute, Federal University of Goiás, Goiânia, GO, Brazil.

*__Corresponding Author:__
Felipe Dalvi-Garcia, MD, PhD
Institute of Psychiatry, Federal University of Rio de Janeiro
71, Venceslau Brás Avenue, Botafogo, Rio de Janeiro – RJ, Zip Code 20290-140, Brazil.
Email: fdalvigarcia@gmail.com
Phone: +55 21 98582 1982

The healthcare sector for mental health is undergoing rapid transformation through artificial intelligence (AI), which provides diagnostic tools and predictive capabilities and digital treatment options through mobile applications, virtual reality platforms and chatbots. The research shows that models face some issues as they operate with restricted training data, require external validation, and have unclear decision-making processes. The guidelines SPIRIT-AI, CONSORT-AI, and British Standard BS 30440 establish requirements for developers to establish auditable protocols, lifecycle monitoring systems, and equity safeguards. Psychiatry needs to achieve a proper balance between developing new approaches and preserving ethical standards in the

present time. The process requires psychiatrists to maintain central involvement through AI data integration with patient narratives and medical records while ensuring privacy protection and algorithmic transparency. AI systems can provide useful functions for risk assessment and monitoring; yet they need to operate under medical staff oversight and strict regulatory frameworks. The combination of digital literacy programs with data protection systems as well as validation protocols will ensure these technologies achieve their intended goals for mental health care.

Clinical Decision Support Systems (CDSS) enhanced by Artificial Intelligence technologies are becoming breakthroughs in Medicine and healthcare practices (Golden et al., 2024). The field of Psychiatry has also taken advantage of AI technologies in different ways (Graham et al., 2019; Lee et al., 2021). AI has already supported differential diagnoses based on brain image analysis, heterogeneous data (e.g., speech, wearable sensors, behavioral data) and has been integrated to provide insights about the neurophysiology of mental illnesses. In this regard, patterns from longitudinal data have been deciphered to detect evolving psychiatric symptoms, biological and digital signals have been united to perform psychiatric screening and risk assessment. Research on digital phenotyping has produced promising, though heterogeneous, results for suicide attempt prediction, depression subtype identification, and schizophrenia relapse prediction, just no name a few (Tornero-Costa et al., 2023; Torous et al., 2021). Digital psychiatry now supports clinical practice through mobile applications, virtual reality therapies, chatbots to deliver care to patients at every stage of their treatment journey for diagnosis and treatment and prognosis assessment (Torous et al., 2021). Medical technology developments can help doctors identify diseases during their initial stages and

develop personalized treatment approaches which bring medical services to people who live in underserved communities.

The scientific literature shows some methodological issues and quality problems with AI systems used in psychiatric settings despite the general enthusiasm of AI technology for mental health care (Tornero-Costa et al., 2023). As highlighted by a recent systematic review, many studies use insufficient data to allow model generalizability in machine learning models, lack external validation, independent verification and international cooperation.

Traditionally rooted in qualitative observation and therapeutic relationships, Psychiatry has been slower to adopt these technologies (Graham et al., 2019). Digital health data continues to grow at a fast pace while patients need more advanced solutions to mental health emergencies, thus making these tools essential for psychiatric research and patient care. Before doing so, some *concerns* need to be addressed. First, interventions involving AI need to meet the same strict *evaluation criteria* which apply to all therapeutic and diagnostic solutions. The guidelines SPIRIT-AI (Cruz Rivera et al., 2020) for protocols and CONSORT-AI for reporting (Liu et al., 2020) provide specific requirements for clinical trials of AI-based interventions by demanding full protocol transparency, complete algorithm documentation and clear documentation of system errors and failures. These guidelines strive to enforce both auditable and reproducible interventions, as they require the interventions to disclose their decision-making process and safety protocols.

Before scale-up, AI systems need to undergo a documented third-party–auditable validation process, which includes safety, clinical efficacy, equity, usability/human factors,

environmental impact, and post-deployment monitoring according to the recently published British Standard BS 30440 framework for AI in healthcare (Sujan et al., 2023). Over the past decade, however, regulatory oversight of medical AI has expanded well beyond the United Kingdom. The United States (US) Food and Drug Administration along with other regulatory agencies, such as Health Canada and Pharmaceuticals and Medical Devices Agency in Japan, have been transitioning their device approval systems to adaptive, lifecycle-based frameworks which include good machine learning practices and pre-market-to-post-market monitoring models (Pantanowitz et al., 2024). The European Union's upcoming AI Act and the US-EU Trade and Technology Council's voluntary AI Code of Conduct show that regulations now focus on establishing standardized risk-based oversight systems and algorithmic responsibility (Palaniappan et al., 2024). These programs demonstrate a worldwide shift from separate fragmented regulatory policies to active governance systems which unite technological progress with protective measures for patients and equitable treatment across different populations (Ardic & Dinc, 2025; Palaniappan et al., 2024).

Additionally, as AI algorithms taking autonomous decisions start to influence diagnostic and therapeutic decisions, *explainability* arises as one of the major concerns. Algorithmic bias, overfitting, and opaque “black-box” predictions could severely compromise patient safety. AI tools demonstrate their highest effectiveness and safety when used under clinician supervision as part of supervised care systems rather than functioning independently (Torous et al., 2021). In the AI era, psychiatrists should learn to assess algorithms, not to defer them, and to recognize when model outputs should be dismissed due to ethical or evidentiary concerns (Blease et al., 2020). In practice, in conformance

with the American Psychiatry Association position, we believe that the psychiatrist's non-delegable work will remain: weaving AI signals into a formulation that balances biography and biology; making trade-offs under uncertainty; safeguarding consent, dignity, equity, and privacy; and leading teams that deliver measurement-based, crisis-safe care (Potash et al., 2025).

Clinician-informed studies also suggest that, for now, AI adds the most value in monitoring and prediction (for trajectories of suicidality) rather than prescriptive recommendations, reinforcing the importance of human participation in this loop (Blease et al., 2020; Fischer et al., 2025). Adapting to this moment means building digital literacy, demanding clear documentation of intended use and failure modes, and incorporating tools within auditable standards, so clinical judgment keeps up the final common pathway to care.

Another concern is *patient privacy and data security*, as psychiatric information is uniquely confidential. The evaluation of digital psychiatry tools shows that apps have different privacy policies and their contingency systems are not trustworthy while their standalone use outside medical supervision produces uncertain results (Torous et al., 2021). The deployment process needs the following essential elements: (a) threat models that describe re-identification, model inversion and data leakage vulnerabilities along with their security countermeasures; (b) simulated tests of crisis escalation procedures, and (c) proactive bias assessment with detailed subgroup analysis for language, race/ethnicity, gender and age variables; and (d) clinician override and accountability for critical decisions.

AI technology brings changes to psychiatry, but its safe application for effective treatment needs rigorous validation protocols with full transparency and patient protection measures. Standard guidelines and auditable lifecycle approaches offer valuable strategies to improve reproducibility, clarity, and safety. Academic governance proves to be insufficient for handling these matters because digital psychiatry issues show that more action is needed. We encourage these standards to be considered as reference points not only for researchers, but also for developers, regulators, and policymakers. By fostering a culture of clarity, pre-market evidence, lifecycle validation, real-world monitoring, strong data governance, and transparent documentation practices allows the development of responsible AI mental health solutions, building up both public trust and clinical value.

The near future will bring advancements to this field through the development of dependable systems that support open innovation practices that ensure fairness and medical safety. AI systems use may unintentionally reinforce existing social inequalities when no protective measures exist. Thus, the successful operation of AI-based predictive tools for early detection and treatment coordination depends on continuous post-deployment monitoring, human oversight and multidisciplinary governance framework.

Funding: The authors acknowledge the financial support from the Brazilian National Council for Scientific and Technological Development (CNPq), through the call CNPq/MS/SCTIE/DECIT No. 45/2022 - Mental Health.

Author contributions: CRediT TaxonomyFelipe Dalvi-GarciaConceptualization-Equal, Investigation-Equal, Methodology-Equal, Writing - original draft-Equal, Writing - review

& editing-EqualEvellin CardosoConceptualization-Equal, Investigation-Equal, Methodology-Equal, Writing - original draft-Equal, Writing - review & editing-EqualLaiana QuagliatoFunding acquisition-Equal, Supervision-Equal, Writing - original draft-Equal, Writing - review & editing-EqualAntonio NardiFunding acquisition-Equal, Supervision-Equal, Writing - original draft-Equal, Writing - review & editing-Equal

Handling Editor: Dr. Ives Passos

References

Ardic, N., & Dinc, R. (2025). Artificial Intelligence in Healthcare: Current Regulatory Landscape and Future Directions. *British Journal of Hospital Medicine*, *86*(8), 1–21. https://doi.org/10.12968/hmed.2024.0972

Blease, C., Locher, C., Leon-Carlyle, M., & Doraiswamy, M. (2020). Artificial intelligence and the future of psychiatry: Qualitative findings from a global physician survey. *Digital Health*, *6*. https://doi.org/10.1177/2055207620968355

Cruz Rivera, S., Liu, X., Chan, A. W., Denniston, A. K., Calvert, M. J., Ashrafian, H., Beam, A. L., Collins, G. S., Darzi, A., Deeks, J. J., ElZarrad, M. K., Espinoza, C., Esteva, A., Faes, L., Ferrante di Ruffano, L., Fletcher, J., Golub, R., Harvey, H., Haug, C., … Yau, C. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. In *The Lancet Digital Health* (Vol. 2, Issue 10). https://doi.org/10.1016/S2589-7500(20)30219-3

Fischer, L., Mann, P. A., Nguyen, M.-H. H., Becker, S., Khodadadi, S., Schulz, A., Edwin Thanarajah, S., Repple, J., Hahn, T., Reif, A., Salamikhanshan, A., Kittel-Schneider, S., Rief, W., Mulert, C., Hofmann, S. G., Dannlowski, U., Kircher, T., Bernhard, F. P., & Jamalabadi, H. (2025). AI for mental health: clinician expectations and priorities in computational psychiatry. *BMC Psychiatry*, *25*(1), 584. https://doi.org/10.1186/s12888-025-06957-3

Golden, G., Popescu, C., Israel, S., Perlman, K., Armstrong, C., Fratila, R., Tanguay-Sela, M., & Benrimoh, D. (2024). Applying artificial intelligence to clinical decision

support in mental health: What have we learned? *Health Policy and Technology*, *13*(2). https://doi.org/10.1016/j.hlpt.2024.100844

Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H. C., & Jeste, D. V. (2019). Artificial Intelligence for Mental Health and Mental Illnesses: an Overview. In *Current Psychiatry Reports* (Vol. 21, Issue 11). https://doi.org/10.1007/s11920-019-1094-0

Lee, E. E., Torous, J., De Choudhury, M., Depp, C. A., Graham, S. A., Kim, H. C., Paulus, M. P., Krystal, J. H., & Jeste, D. V. (2021). Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom. In *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* (Vol. 6, Issue 9). https://doi.org/10.1016/j.bpsc.2021.02.001

Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., Denniston, A. K., Chan, A. W., Darzi, A., Holmes, C., Yau, C., Ashrafian, H., Deeks, J. J., Ferrante di Ruffano, L., Faes, L., Keane, P. A., Vollmer, S. J., Lee, A. Y., Jonas, A., Esteva, A., Beam, A. L., … Rowley, S. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine*, *26*(9). https://doi.org/10.1038/s41591-020-1034-x

Palaniappan, K., Lin, E. Y. T., & Vogel, S. (2024). Global Regulatory Frameworks for the Use of Artificial Intelligence (AI) in the Healthcare Services Sector. In *Healthcare (Switzerland)* (Vol. 12, Issue 5). https://doi.org/10.3390/healthcare12050562

Pantanowitz, L., Hanna, M., Pantanowitz, J., Lennerz, J., Henricks, W. H., Shen, P., Quinn, B., Bennet, S., & Rashidi, H. H. (2024). Regulatory Aspects of Artificial Intelligence and Machine Learning. *Modern Pathology*, *37*(12), 100609. https://doi.org/10.1016/j.modpat.2024.100609

Potash, J. B., McClanahan, A., Davidson, J., Butler, W., Carroll, N., Ruble, A., Yaden, M., King, D., Torous, J., Zandi, P. P., Kennedy, K. G., Smith, T. E., Waghray, A., Trestman, R., & Wills, M. (2025). The Future of the Psychiatrist. *Psychiatric Research and Clinical Practice*, *7*(2), 80–90. https://doi.org/10.1176/appi.prcp.20240130

Sujan, M., Smith-Frazer, C., Malamateniou, C., Connor, J., Gardner, A., Unsworth, H., & Husain, H. (2023). Validation framework for the use of AI in healthcare: Overview of

the new British standard BS30440. *BMJ Health and Care Informatics*, *30*(1). https://doi.org/10.1136/bmjhci-2023-100749

Tornero-Costa, R., Martinez-Millana, A., Azzopardi-Muscat, N., Lazeri, L., Traver, V., & Novillo-Ortiz, D. (2023). Methodological and Quality Flaws in the Use of Artificial Intelligence in Mental Health Research: Systematic Review. In *JMIR Mental Health* (Vol. 10). https://doi.org/10.2196/42045

Torous, J., Bucci, S., Bell, I. H., Kessing, L. V., Faurholt-Jepsen, M., Whelan, P., Carvalho, A. F., Keshavan, M., Linardon, J., & Firth, J. (2021). The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry*, *20*(3). https://doi.org/10.1002/wps.20883