BJPsych Open (2026)
12, e246, 1–4. doi: 10.1192/bjo.2026.12101

Short report

Methodological transitions in mental health research driven by machine learning

Haojie Fu, Xudong Zhao and Fazhan Chen

Summary

Machine learning is increasingly reshaping psychological and mental health research by complementing traditional theory-driven approaches with data-driven predictive modelling. This encompasses three key transitions: from hypothesis-testing to pattern discovery; from controlled experimental settings to naturalistic and multidimensional data ecosystems; and from explanatory theoretical models to prediction-informed theory-building and clinical translation. Nevertheless, this methodological reorientation remains insufficiently synthesised. Our aim is to examine how machine learning is reshaping psychological research methodologies, focusing on its ability to process and analyse multidimensional data, its capacity to model complex associations and temporal patterns and its integration into clinical practice. We critically examine both the opportunities and boundaries of integrating predictive models with theoretical insights. We further clarify that predictive accuracy, interpretability and causal explanation are related but distinct aims, and that machine learning should complement rather than replace psychological theory, causal reasoning and clinical judgement.

Keywords

Mental health; machine learning; data analysis; methodological transition; artificial intelligence.

Copyright and usage

© The Author(s), 2026. Published by Cambridge University Press on behalf of Royal College of Psychiatrists. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Machine learning is transforming mental health research by enhancing diagnosis, prediction and treatment–response modelling.[1] Compared with many conventional statistical workflows, machine learning efficiently processes diverse data sources, including health records and wearable device inputs, enabling sophisticated predictive modelling of mental health risks.[2]

Traditional psychology has long used quantitative research methods alongside observational, psychometric and clinical approaches, utilising controlled experiments and statistical inference to study human behaviour. Controlled experiments manipulate independent variables while controlling extraneous factors to establish causal relationships under well-specified assumptions.[3] These methods emphasise objectivity and reproducibility, yet have limitations. Controlled experiments often suffer from low ecological validity, because their rigid conditions may not reflect real-world behaviour.

With the rise of machine learning, data-driven models are increasingly complementing and, in some cases, reshaping traditional research paradigms. By leveraging large-scale, high-dimensional data, these models can detect intricate patterns in psychological data that conventional statistical methods might overlook.

Different analytical methods provide researchers with specialised tools to address various research questions, each with unique strengths and applications. Various machine learning approaches offer different computational strategies for analysing psychological data, ranging from linear models and decision trees that provide interpretable results to more complex methods like neural networks and reinforcement learning that can capture intricate patterns but may sacrifice transparency.[4] Given the diverse nature of psychological research, no model is universally superior. Instead, selecting the appropriate model based on research objectives allows for more effective and targeted applications.

Although machine learning offers powerful predictive capabilities for mental health research, its implementation raises a fundamental challenge: balancing predictive power with scientific understanding. As these computational approaches increasingly complement traditional methods, they must provide not only accurate predictions but also insights into why and how these predictions are made. This brings us to a critical aspect of the evolving research paradigm in psychological science. However, this methodological reorientation in mental health research driven by machine learning is not succinctly summarised in the literature. Our aim is to examine how machine learning is transforming psychological research methodologies, particularly how data-driven approaches can complement theory-driven research and the implications for future research paradigms. This methodological reorientation is characterised by three key transitions: from hypothesis testing to pattern discovery; from controlled experimental settings to naturalistic and multidimensional data ecosystems; and from explanatory theoretical models to prediction-informed theory-building and clinical translation. Throughout this article we emphasise that predictive utility, interpretability, causal explanation and clinical actionability are related but distinct aims.

Recent reviews, reporting guidelines and representative studies indicate several observable methodological trends in machine learning-based mental health research, summarised in Box 1.

Box 1 Observable methodological trends in machine learning-based mental health research

(a) *Broader data ecosystems:* from questionnaire-based and experimental datasets to electronic health records, wearable sensors, speech, text and digital behavioural traces.
(b) *Prediction-oriented tasks:* increasing use of predictive modelling for risk stratification, early warning and treatment-response estimation.
(c) *Stronger validation expectations:* growing emphasis on external validation, calibration, fairness and clinical utility.

Together, these trends support the framing of machine learning as a bounded methodological reorientation in mental health research.

https://doi.org/10.1192/bjo.2026.12101 Published online by Cambridge University Press

From explainable machine learning to causal awareness

The inevitable need for interpretability

Psychological and mental health research is not just about building better predictive models, it is also about understanding causal relationships and identifying key influencing factors. In clinical practice, doctors and therapists need to interpret the reasoning behind predictions to develop effective treatment plans and communicate with patients. Complex machine learning models, often referred to as 'black-box' models due to their opaque internal workings, present significant challenges in clinical settings.[5] Although some argue that predictive accuracy may be sufficient for certain applications,[6] we contend that explainability is particularly valuable in mental health research, where understanding mechanisms is crucial for theory development and where transparency can help ensure fairness and prevent biases. For example, a study using the Secure Anonymised Information Linkage database developed machine learning models to predict childhood mental health problems and used Shapley Additive Explanations to identify model-contributing factors, thereby improving interpretability for clinical applications.[7] Such approaches are particularly useful when candidate predictors are numerous, non-linear or heterogeneous, although model comparisons and data type-specific choices remain necessary.[7,8] This example shows how integrating explainability into machine learning models can be essential for moving from accurate prediction towards clinically interpretable and theory-relevant hypotheses, rather than proving causal mechanisms directly.

From linear causality to dynamic systems thinking

Psychological phenomena are rarely the result of simple linear cause–effect relationships: rather, they emerge from complex, multifactorial interactions. Whereas many conventional designs and models emphasise linear or additive associations, numerous studies have demonstrated their limitations in capturing the dynamic and feedback-driven nature of mental processes.[3] To advance psychological research we need to shift from linear causality to a systems-thinking approach, where multiple factors interact through feedback loops and dynamic changes over time.[9] Machine learning is a useful tool that, through its powerful non-linear function approximation ability and adaptive feature-learning mechanism, can effectively capture complex dynamic structures.[10,11] For example, techniques like recurrent neural networks[12] and long short-term memory networks[13] are particularly well-suited for modelling temporal dependencies in longitudinal or sequential data, enabling researchers to better understand how mental health conditions evolve over time. However, temporal prediction is not equivalent to causal identification. Claims about feedback loops or cyclic causality require appropriate repeated-measure designs, temporal ordering and causal assumptions, rather than being established by the model architecture alone. By modelling non-linear and dynamic temporal patterns, machine learning may help researchers generate hypotheses about how mental health conditions develop and how interventions may influence outcomes over time.

A useful distinction is therefore needed among four related but different aims. Prediction asks whether a model can accurately estimate future or unobserved outcomes. Interpretability asks which features or patterns contributed to a model output. Causal inference asks what would happen under a hypothetical intervention, conditional on assumptions such as adequate confounding control and temporal ordering. Mechanistic explanation asks why and through which psychological or biological processes an effect occurs. Machine learning therefore serves as a computational bridge linking prediction, interpretation, causal reasoning and mechanistic hypotheses.

Large language models as research tools in mental health

The field of psychological research has been transformed by various machine learning methods, including random forests, support vector machines, neural networks and unsupervised learning approaches. The selection of a particular method typically depends on factors such as input data types, research objectives and the deployment context. In this section we focus on large language models (LLMs) as a particularly impactful example, due to their ability to process and understand vast amounts of unstructured textual data. This makes LLMs directly relevant to the transition from controlled or questionnaire-based data to naturalistic language data, including interviews, clinical notes and digital communication. LLMs have proven useful in areas like sentiment analysis, text-based mental health assessment and diagnostic interviews, where the interaction between language and psychological processes is central.[14,15] By providing this view, we aim to clarify the role of LLMs within the broader machine learning landscape as research tools rather than as substitutes for human participants or clinical judgement.

Machine learning as a tool, not a research object

The methodological status of LLMs in psychological research requires careful clarification. Some scholars argue that psychology also influences machine learning, particularly through behavioural experiments on LLMs. This has led to a new paradigm in artificial intelligence research – treating LLMs as computational models of human cognition, such as language comprehension, reasoning and decision-making. Researchers analyse LLMs' internal representations and computational processes to infer human cognitive mechanisms,[16] while others view LLMs as simulators of human behaviour, using task-based and situational studies to predict human responses in similar contexts.[17]

However, this approach faces substantial criticism. LLMs inherently contain biases, potentially leading to misleading conclusions when applied to human cognition.[18–20] Machine learning models, including LLMs, are fundamentally tools – they assist in studying psychological phenomena but cannot replace the research subjects themselves. They are useful for role-play and computational modelling, provided their outputs are validated against human data.[19] It is crucial to distinguish between a tool and the actual object of study, avoiding the fallacy of equating LLMs with human cognition or behaviour. Thus, LLMs may be appropriate for processing clinical text, generating hypotheses or modelling bounded task responses, but they should not be treated as direct evidence of human mental states without empirical validation.

Proper application: integrating LLMs as research tools through validation and controlled sampling

Sampling control is fundamental to generating explainable insights from machine learning models in psychological research. Sampling biases can affect the reliability of results[21] and, in language model research, it is essential to increase the diversity of the samples to avoid biased outcomes.[18] Additionally, the results must be validated to prevent mistakenly treating the model's biases as the actual results.[16,19] Beyond sampling control, LLMs can serve as valuable research tools when technically refined for psychological applications. These models demonstrate advanced text-processing

https://doi.org/10.1192/bjo.2026.12101 Published online by Cambridge University Press

Table 1 Distinguishing large language models (LLMs) as research tools from participant substitutes in mental health research

LLM role	Methodological implication
Research tool	Appropriate for processing clinical text, extracting features from interviews or notes, generating hypotheses and supporting bounded task modelling, provided that outputs are validated against human-coded or empirical data.
Participant proxy/human substitute	Should not be used for direct inference about real populations, mental states or clinical outcomes unless the task is narrowly defined and externally validated against human behavioural or clinical data.

LLMs may assist in psychological and mental health research, but their outputs should not be treated as direct evidence of human cognition, behaviour or clinical status, without task-specific validation.

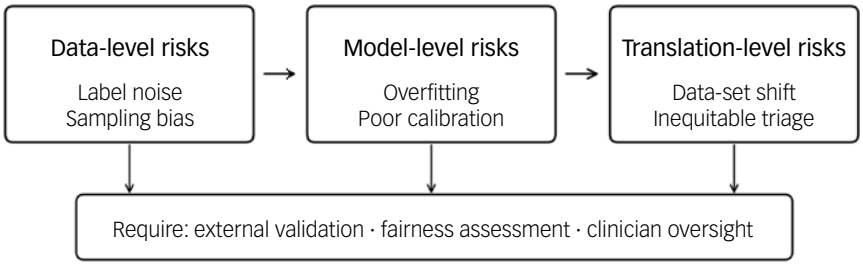


Fig. 1 Practical barriers in the clinical translation of machine learning models.

capabilities that can rival general practitioners in certain medical assessments. Research indicates that domain-specific fine-tuning significantly enhances LLMs' psychological applications, by adapting their parameters to specialised contexts. Combined with controlled sampling and human data validation, such technical adaptation allows researchers to use LLMs for more nuanced mental health predictive modelling. For example, one *Nature* study showed that a fine-tuned large prediction model could simulate aspects of human cognitive behaviour when validated against human behavioural data, thereby supporting theoretical interpretation, hypothesis generation and preliminary testing.[14] The practical value emerges when these refined models are deployed within clearly defined research parameters and evaluated against established psychological standards.

The discussion of LLMs highlights a shift in psychological research. On the one hand, they can be integrated into established research paradigms; on the other, LLMs, through fine-tuning and application in new contexts, provide new methods to challenge traditional approaches. This dual role demonstrates how LLMs are changing and strengthening psychological and mental health research in both innovative and practical ways, particularly by expanding access to naturalistic language data, supporting pattern discovery in unstructured text and generating hypotheses that require subsequent human-data validation.

Building on recent discussions about the opportunities and risks of using LLMs in psychological research, Table 1 distinguishes LLMs as research tools from their more controversial use as participant proxies or human substitutes.[16–20]

Methodological reorientation in psychological and mental health research

With advances in computing technology and interdisciplinary collaboration, psychological and mental health research is undergoing a methodological reorientation. This transformation reflects not only new analytic tools, but also a change in how researchers connect data, prediction, interpretation, causal reasoning and clinical application.

From theoretical to prediction-informed modelling

Psychological research is increasingly incorporating predictive models, which support the forecasting of future behaviours or mental health outcomes alongside theoretical interpretation.[4] Predictive modelling can complement hypothesis-driven approaches, and offers practical utility when models are validated across settings. With the integration of machine learning and big data analytics, predictive models are being widely used to assess mental health risks, detect early warning signs of disorders and inform intervention strategies. Although explanatory models remain essential for theory-building, their combination with predictive approaches enhances both the understanding and practical application in clinical and research settings. A study utilising machine learning causal inference to personalise therapy assignments for grief among veterans demonstrated its practical value in optimising treatment approaches.[22] This example shows how integration of explainability into machine learning models can be essential in moving from accurate predictive models towards clinically interpretable and theory-relevant hypotheses.[14] Interpretable machine learning can help researchers identify potential causal relationships, generate new research hypotheses and refine existing theories.[12]

From experimental control to causal inference under assumptions

Traditional psychology has relied on controlled experiments to establish causal relationships. However, this approach has significant limitations, including ethical constraints, ecological validity concerns and difficulties in manipulating real-world variables. The advancement of causal inference techniques has empowered researchers to estimate causal effects from observational data under explicit assumptions, thus partly addressing several of these methodological constraints.[3,23] Causal inference techniques offer the advantage of formalising causal questions in observational data, avoiding some ethical challenges. However, they also come with limitations, such as the challenge of distinguishing correlation from causation and the potential for confounding factors to skew results.[24] These advancements are particularly impactful in computational psychiatry, where understanding the causal mechanisms of mental disorders is critical.[25] This shift expands our ability to investigate mental illness aetiology and intervention efficacy beyond the constraints of controlled laboratory settings, but only when predictive modelling is combined with appropriate study design, domain theory and causal identification strategies.

Figure 1 summarises these barriers at the data, model and clinical translation levels. Key risks include data-set shift across

https://doi.org/10.1192/bjo.2026.12101 Published online by Cambridge University Press

settings, label noise in psychiatric outcomes, overfitting in high-dimensional data, weak external validation, poor calibration and inequitable algorithmic triage. These risks do not negate the value of machine learning, but they require transparent reporting, external validation, fairness assessment and clinician oversight before clinical use.

Implications and future directions

Machine learning is catalysing a methodological reorientation in psychological and mental health research, rather than merely updating methodologies. The essence of this transformation lies not in replacing theory-driven explanation with data-driven prediction, but in connecting theory, data, prediction, interpretation and clinical translation more systematically.

This methodological reorientation is characterised by three key transitions: from hypothesis-testing to pattern discovery; from controlled experiments to naturalistic and multidimensional data ecosystems; and from explanatory theoretical models to prediction-informed theory-building and clinical translation. These developments not only enhance our understanding of complex psychological phenomena but also establish a robust foundation for personalised precision interventions.

Future work should therefore prioritise theory-constrained, externally validated and clinically accountable machine learning. The value of machine learning lies not in replacing existing research traditions but in making them more scalable, integrative and responsive to real-world mental health needs.

Haojie Fu, Biomedical Big Data Centre, West China Hospital, Sichuan University, Chengdu, China; and Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, Shanghai, China; **Xudong Zhao**, Department of Psychosomatic Medicine, Shanghai East Hospital, School of Medicine, Tongji University, Shanghai, China; **Fazhan Chen**, Clinical Research Centre for Mental Disorders, Shanghai Pudong New Area Mental Health Centre, School of Medicine, Tongji University, Shanghai, China; School of Humanities, Tongji University, Shanghai, China; and Shanghai Key Laboratory of Data Science, Shanghai, China

Correspondence: Haojie Fu. Email: fuhj.alen@gmail.com

First received 8 Mar 2025 UTC, final revision 30 May 2026 UTC, accepted 5 Aug 2026 UTC

Data availability

This article does not involve the creation or analysis of any new data, and therefore data availability is not applicable.

Author contributions

H.F. prepared this article and edited subsequent revisions. X.Z. and F.C. provided critical feedback and assisted with the interpretation of results. All authors reviewed and approved the final version of the manuscript.

Funding

This work was funded by STI2030-Major Projects (no. 2021ZD0202000), the National Center for Mental Health (no. SXRH-RW-2025034), the Shanghai Health System Key Discipline Construction Project (no. 2024ZDXK0011), the Shanghai Pudong New Area Science and Technology Development Fund (no. PKJ-2023-Y79) and the Healthcare Talents Elite Programme of Shanghai Pudong New Area (no. 2025PDWSYCBJ-09).

Declaration of interest

None.

References

1 Vannucci A, Fields A, Heleniak C, Bloom PA, Harmon C, Nikolaidis A, et al. Machine learning for identifying caregiving adversities associated with greatest risk for mental health problems in children. *Nat Ment Health* 2025; **3**: 71–82.

2 Abd-alrazaq A, AlSaad R, Aziz S, Ahmed A, Denecke K, Househ M, et al. Wearable artificial intelligence for anxiety and depression: scoping review. *J Med Intern Res* 2023; **25**: e42672.

3 Sheetal A, Feng Z, Savani K. Using machine learning to generate novel hypotheses: increasing optimism about COVID-19 makes people less willing to justify unethical behaviors. *Psychol Sci* 2020; **31**: 1222–35.

4 Haghish EF, Nes RB, Obaidi M, Qin P, Stänicke LI, Bekkhus M, et al. Unveiling adolescent suicidality: holistic analysis of protective and risk factors using multiple machine learning algorithms. *J Youth Adolesc* 2024; **53**: 507–25.

5 Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, et al. Clinical applications of machine learning algorithms: beyond the black box. *BMJ* 2019; **364**: l886.

6 Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* 2021; **3**: e745–50

7 Crowley R, Parkin K, Rocheteau E, Massou E, Friedmann Y, John A, et al. Machine learning for prediction of childhood mental health problems in social care. *BJPsych Open* 2025; **11**: e86.

8 Chekroud AM, Bondar J, Delgadillo J, Doherty G, Wasil A, Fokkema M, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry* 2021; **20**: 154–70.

9 Zhao X, Ma X, Yao Y, Wan C, Ng E. Chinas little emperors show signs of success. *Science* 2013; **339**: 905–6.

10 Fu H, Tang S, Zhao X. Limitations of correlation coefficients in research on functional connectomes and psychological processes. *Human Brain Mapp* 2025; **46**: e70287.

11 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*: 6000–10. Curran Associates Inc., 2017.

12 Ji-An L, Benna MK, Mattar MG. Discovering cognitive strategies with tiny recurrent neural networks. *Nature* 2025; **644**: 993–1001.

13 Xie W, Wang C, Lin Z, Luo X, Chen W, Xu M, et al. Multimodal fusion diagnosis of depression and anxiety based on CNN-LSTM model. *Comput Med Imaging Graph* 2022; **102**: 102128.

14 Binz M, Akata E, Bethge M, Brändle F, Callaway F, Coda-Forno J, et al. A foundation model to predict and capture human cognition. *Nature* 2025; **644**: 1002–9.

15 Suri G, Slater LR, Ziaee A, Nguyen M. Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *J Exp Psychol Gen* 2024; **153**: 1066–75.

16 Bao H-W-S. The Fill-Mask Association Test (FMAT): measuring propositions in natural language. *J Person Soc Psychol* 2024; **127**: 537–61.

17 Zhao Y, Huang Z, Seligman M, Peng K. Risk and prosocial behavioural cues elicit human-like response patterns from AI chatbots. *Sci Rep* 2024; **14**: 7095.

18 Abdurahman S, Atari M, Karimi-Malekabadi F, Xue MJ, Trager J, Park PS, et al. Perils and opportunities in using large language models in psychological research. *PNAS Nexus* 2024; **3**: 245.

19 Lin Z. Six fallacies in substituting large language models for human participants. *Adv Methods Pract Psychol Sci* 2025; **8**: 25152459251357566.

20 Lin Z. Large language models as psychological simulators: a methodological guide. *Adv Methods Pract Psychol Sci* 2026; **9**: 25152459251410153.

21 Berghoff R, Bylund E. Diversity in research on the psychology of language: a large-scale examination of sampling bias. *Cognition* 2025; **256**: 106043.

22 Argyriou E, Gros DF, Hernandez Tejada MA, Muzzy WA, Acierno R. A machine learning personalized treatment rule to optimize assignment to psychotherapies for grief among veterans. *J Affect Disord* 2024; **358**: 466–73.

23 Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat Machine Intell* 2020; **2**: 369–75.

24 VanderWeele TJ. Can sophisticated study designs with regression analyses of observational data provide causal inferences? *JAMA Psychiatry* 2021; **78**: 244–6.

25 Lucasius C, Ali M, Patel T, Kundur D, Szatmari P, Strauss J, et al. A procedural overview of why, when and how to use machine learning for psychiatry. *Nat Ment Health* 2025; **3**: 8–18.

https://doi.org/10.1192/bjo.2026.12101 Published online by Cambridge University Press