The Digital Mind: New Concepts in Mental Health 2

From promise to practice: towards the realisation of AI-informed mental health care

Nikolaos Koutsouleris*, Tobias U Hauser, Vasilisa Skvortsova, Munmun De Choudhury*

In this Series paper, we explore the promises and challenges of artificial intelligence (AI)-based precision medicine tools in mental health care from clinical, ethical, and regulatory perspectives. The real-world implementation of these tools is increasingly considered the prime solution for key issues in mental health, such as delayed, inaccurate, and inefficient care delivery. Similarly, machine-learning-based empirical strategies are becoming commonplace in psychiatric research because of their potential to adequately deconstruct the biopsychosocial complexity of mental health disorders, and hence to improve nosology of prognostic and preventive paradigms. However, the implementation steps needed to translate these promises into practice are currently hampered by multiple interacting challenges. These obstructions range from the current technology-distant state of clinical practice, over the lack of valid real-world databases required to feed data-intensive AI algorithms, to model development and validation considerations being disconnected from the core principles of clinical utility and ethical acceptability. In this Series paper, we provide recommendations on how these challenges could be addressed from an interdisciplinary perspective to pave the way towards a framework for mental health care, leveraging the combined strengths of human intelligence and AI.

Lancet Digit Health 2022

Published Online
October 10, 2022
https://doi.org/10.1016/S2589-7500(22)00153-4

*Contributed equally

This is the second in a Series of two papers about the digital mind: new concepts in mental health. All papers in the Series are available at www.thelancet.com/series/the-digital-mind

Section for Precision Psychiatry, Department of Psychiatry and Psychotherapy, Ludwig Maximilian University, Munich, Germany (N Koutsouleris MD); **Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK** (N Koutsouleris); **Max Planck Institute of Psychiatry, Munich, Germany** (N Koutsouleris); **Max Planck UCL Centre for Computational Psychiatry and Ageing Research, London, UK** (T U Hauser PhD, V Skvortsova PhD); **Wellcome Centre for Human Neuroimaging, University College London, London, UK** (T U Hauser, V Skvortsova); **School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, USA** (M De Choudhury PhD)

Correspondence to:
Dr Nikolaos Koutsouleris, Section for Precision Psychiatry, Department of Psychiatry and Psychotherapy, Ludwig Maximilian University, D-80336 Munich, Germany
nikolaos.koutsouleris@med.uni-muenchen.de

The vision: precision in mental health

Mental health remains the only domain in medicine that depends entirely on the patient's ability to report their cognitive and emotional states, the course of their symptoms, and their interactions with relatives, friends, and colleagues. Similarly, current mental health-care practices demand that clinicians accurately recognise and map these dynamic states to diagnostic, prognostic, and therapeutic decisions under the constraints of varying resources, skillsets, and temperaments. This variability might lead to a hasty and superficial appraisal of the patient, or conversely, to a lengthy and exuberant exploration of the patient's mental landscape. The ideal, however, assumes that diagnosis and prognosis are precise procedures that inform the selection of the optimal treatment regimen for the patient, following the principles of evidence-based medicine (EBM).

A host of arguments have been put forward to explain why our clinical reality deviates from principles of EBM. In many settings, therapeutic traditions, stigma, and negative attitudes towards mental illness[1] impede the implementation of EBM.[2] Even if EBM is delivered, it might fall short of resolving the variability of patients' responses due to its reliance on group-level statistical evidence.[3,4] Thus, clinicians and patients are still forced into cumbersome trial-and-error searches of the best therapeutic strategy if mean-based EBM recom-mendations are inadequate. This clinical reality undermines patients' trust in a medical remedy of their illness, which in turn may exacerbate and prolongate disease pathology and precipitate poor outcomes.[5,6]

As Carr stated, "AI is the field of computer science that includes machine learning [algorithms], natural language processing, speech processing, robotics and similar automated decision-making".[7] The rapid growth of machine learning techniques within the artificial intelligence (AI) field (panel 1) has stirred hope that algorithms might be capable of overcoming the trial-and-error-driven status quo in mental health care by supporting precise diagnoses, prognoses, and therapeutic choices.[8] Recent reviews[9,10] have highlighted (1) the methodological strengths and weaknesses of machine learning techniques;[11,12] (2) their potential for clinical translation as the methodological backbone of personalised service delivery in mental health[8,13] (eg, in psychotic[14–16] or depressive disorders);[17–19] (3) their utility for analysing the biobehavioural and environmental heterogeneity of the diagnostic system into more manageable factors, subgroups, and dimensions;[20–24] and (4) their ethical implications for psychiatric care and research.[25–27] However, a synoptical and critical discussion of these aspects in light of the expected transformative impact of AI-driven precision medicine tools in mental health care is currently missing.

In parallel to the development of these data-driven, mechanism-agnostic methods, deeper insights into microscopic to macroscopic brain circuitry have given rise to mechanism-driven models that explain and simulate the development and expression of pathological human behaviour[28,29] and its response to therapeutic interventions.[30] In parallel, deep neural learning techniques in computational psychiatry, like generative embedding, are helping to link disease mechanisms to machine learning-based predictions, whereby inferences about patient-specific physiology or cognition are being used as features for subsequent supervised and unsupervised learning.[31] Thus, mechanism-driven models could provide in-silico probes of the

Panel 1: Glossary

Anchoring (cognitive bias)
Relying on initial impressions too early in the diagnostic process and failing to adjust initial impressions when considering new information.

Artificial intelligence (AI)
Field of study that aims to enable machines to perceive their environment and rationally act on the processed information. Generally, AI is differentiated into weak and strong AI. Weak AI comprises mathematical models and robots with a narrow transactional scope (eg, building specific elements of a car, or predicting a set of diagnoses). Strong AI consists of methods that emulate natural human behaviour by being able to process diverse information sources and accordingly adapt their behaviour.

AI-ready real-world database
A database making structured qualitative and quantitative patient-level information available for analysis by means of AI methods (eg, enabling the goal of producing decision support tools for clinical practice). The database adequately represents the mixture of disease phenotypes in the given population in terms of disease prevalence, and cross-sectional and longitudinal heterogeneity across societal strata and diverse population groups. Representativeness guards against model bias caused by training machine learning algorithms on convenience samples and by applying them to patients not included in the original training cohort.

Data-body-machine entanglements
A structured information exchange framework encompassing humans and machine elements that aims to augment and widen human pattern recognition abilities or decision-making skills for the purpose of improved health-care service delivery.

Electronic medical records
Clinical database that collects, visualises, and disseminates procedural health information of patients in the given health-care system over time (eg, referral sources, results of diagnostic procedures, and drug prescriptions).

Generative embedding
An integrated analytical strategy that first explains how the data might have arisen from the underlying pathophysiological and physiological processes, and then establishes machine learning models within the explanatory variable space. Thus, the goal of generative embedding is to make machine learning models' predictions transparent with respect to the disease-generating mechanisms.

Machine learning
An instantiation of AI that comprises methods to optimise mathematical functions for solving specific tasks (eg, explaining the variation of data [unsupervised machine learning], or predicting outcomes [supervised machine learning]). Earlier machine learning algorithms involved optimisation functions that typically weighted the elements of the input data (features) to generate output (shallow learning), whereas newer deep learning algorithms incorporate the feature engineering process previously accomplished by human experts.

Measurement-based care
A mental-health-care paradigm that relies on electronic medical records and that allows patients, relatives, and health-care providers to generate, process, analyse, and decide upon individualised quantitative information of mental conditions. Measurement-based care is supported by AI algorithms that aid health-care professionals, patients, and relatives to parse and filter rich health-care information into actionable health-care indications.

Natural language processing
Machine learning algorithms capable of parsing free written or spoken language into numerical representations that can be further analysed by predictive systems to predict disease outcomes or infer diagnoses, for example.

Premature closure (cognitive bias)
Accepting a diagnosis before it has been fully verified and believing in a single explanation of a situation without investigating other possibilities.

Reliance on authority (cognitive bias)
Relying unduly on authority or technology.

pathophysiological processes at work in a given patient,[28,32] as demonstrated in a case study of patients with monogenic ion channelopathies,[33] and thus facilitate more precise and personalised treatments (see Hauser and colleagues,[34] the companion paper in this Series).

Finally, digital phenotyping tools, such as wearables, ecological momentary assessments, and electronic medical records (EMR) rapidly expand the evidence base for both mechanism-driven and agnostic modelling of human cognition, behaviour, and social interactions, providing unprecedented opportunities for the implementation of predictive data science in mental health care.[16,35–37] Similarly, the costs for in-depth genetic and molecular testing are dropping, broadening the access to omics technologies beyond research and academia. Analogous to the rise of computer vision and natural language understanding to everyday ubiquity, this transition to a big-data, measurement-based care (MBC) paradigm in mental health care, which is an approach that uses symptom assessments to track patient outcomes over time and inform clinical decision making,[38] could fuel the integration of AI in psychiatric research and service delivery (panel 2).[39]

However, while these innovations promise to revolutionise health care, little progress has been made toward real precision mental health applications.[40,41] Implementation of these applications is often an afterthought. Wiens and colleagues[42] described current strategies as "far from optimal", and current machine learning approaches rarely consider implementation or stakeholder-driven considerations in model design or evaluation. Despite some promising examples of AI in real-world mental-health-care settings,[43] modelling and implementation phases remain largely disconnected. Accordingly, treatment strategies that rely on such advanced techniques might neglect grounded psychiatric evaluations, and might not display the empathic concern and awareness of human physicians.[44] As a result, individuals who only rely on AI-based interventions are often discouraged to pursue treatment.[45] Therefore, we will reflect on the current practices and unmet needs in the development, validation, and implementation phases of computational models from the end users' and health systems' perspectives. Thus, we aim to complement the methodological perspectives on psychiatric data science provided by Hauser and colleagues[34] and hope to elucidate the cascade of translational challenges that need to be addressed for computational methods to become a reality in mental health.

Data and modelling considerations for successful implementation into health care

Opportunities and challenges of big data in mental health care

The development of robust predictive models starts with high-quality, reliable, and sufficiently representative data that capture both the variability, complexity, and specificity of the targeted phenomena (figure). Outside of health care, these principles have been exemplified by the research fields of computer vision and machine translation. First, ImageNet, an image-based ontological database comprising 14 million images, laid the groundwork for super-human object recognition that can now be delivered on demand by convolutional neural networks. Publicly released in 2007, this database both democratised and stimulated predictive data science, because it encouraged both collaboration and competition among AI researchers worldwide. Based on ImageNet, this community annually competed for the best computer vision algorithms and within only 7 years, improved algorithmic object recognition from an accuracy of 71·8% to 97·3%.[46] Similarly, large corpora of parallel language segments—extracted from the internet using web crawlers, processed by human editors, and postprocessed by machine learning—enabled the machine translation field to transit from conventional statistical approaches to deep neural network algorithms.[47] A prominent example is the open-source OPUS collection, which contains 57 corpora covering 700 languages and 70 000 aligned bitexts across all corpora.[48] The availability of these data accelerated the development of deep autoencoders based on self-attentional and attentional techniques, such as CUBBITT, which recently reported human-level translation performance.[49]

Panel 2: **Key recommendations**

- Transition to a measurement-based system of mental-health-care delivery that provides privacy-protecting broad-consent policies to collect large-scale representative datasets for the training and validation of generalisable AI tools
- Implement debiasing strategies during data acquisition, model training and validation to minimise the risk of erroneous model predictions at the point-of-care level, and increasing health-care disparities at the system level
- Enrich electronic medical record systems with procedural metadata to render predictive health-care processes, decisions, and results as transparent and actionable as possible during the model interpretation phase
- Strengthen research into human–AI interactions to better delineate personal and system-level biases on the human side, and design AI methods that optimally adapt to the specific user and health-care context
- Invest into extending medical curricula towards the concepts, opportunities, and challenges of AI-driven digital health care

For **CUBBITT** see https://lindat.mff.cuni.cz/services/translation/

These examples illustrate that carefully curated datasets encompassing the heterogeneity of patient journeys across biological, behavioural, and environmental scales could encourage the development of machine learning algorithms in mental health care that can support both better prediction and understanding of diagnoses and forthcoming outcomes through inference and explanations (figure).[50–52] In particular, the imple-mentation of MBC based on digital phenotyping tools has been considered as the prerequisite for generating AI-ready mental-health-care data.[53] However, MBC services are hampered by challenges ranging from data security and confidentiality concerns to system-level service bottlenecks like the complete absence of training resources, difficulties in implementing broad consent mechanisms, and fragmentation of data formats.[54–57] Also, the widespread belief that clinical judgement is superior to quantitative measures might have stalled the transition toward a digital health-care framework, although empirical data support the increased quality of care delivery once such systems are implemented.[39] Ultimately, these limitations hamper the adoption of MBC practices, and delay the generation of AI-ready real-world databases of mental health disorders.

Promises and limitations of EMR-based databases for predictive data science

The Clinical Record Interactive Search (CRIS) system, developed by The National Institute for Health and Care Research Maudsley Biomedical Research Centre, is one example of how these challenges could be overcome. CRIS that has collected over 400 000 anonymised mental health records since 2007. More recently, CRIS has been augmented with a portfolio of AI-powered natural language processing (NLP) algorithms that sift records for

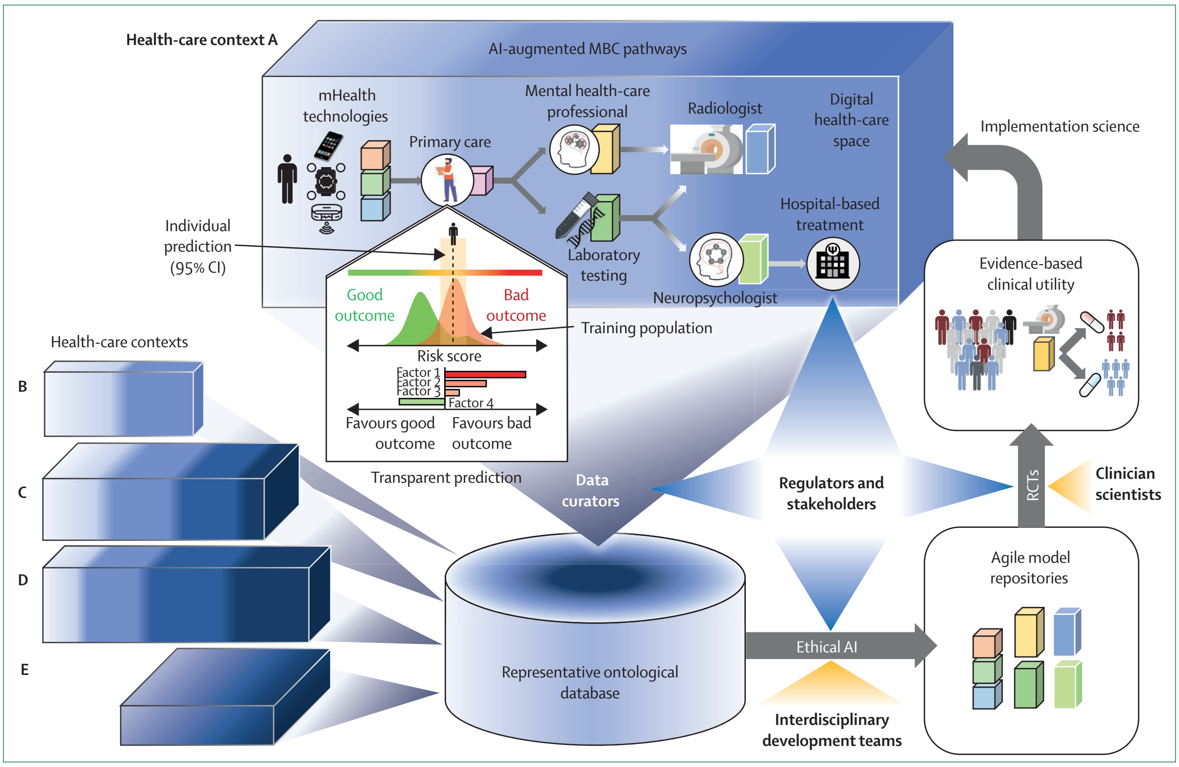

Figure: **An artificial intelligence (AI)-informed learning health-care system**

MBC pathways are embedded into a digital health-care space. The digital space covers different layers of granularity, including stakeholder-level information (eg, clinicians' notes, patient-reported symptoms and outcomes, examinations, and treatments) and system-level data (eg, disease prevalence and socioeconomic parameters in a given community). Pathways are supported by AI-based decision support tools, which provide transparent predictions for shared decision making. Data generated within the digital health-care space are curated and stored within a representative ontological database. Multiple health-care contexts (eg, representing different health-care systems) feed similarly into the database so that a given condition is represented through multiple data instances. At regular intervals, data are extracted by interdisciplinary model-development teams and analysed using, for example, ethical AI methods to develop new or recalibrate existing prediction tools. Tools are made available via model libraries to clinical scientists who test them for clinical utility in stratified RCTs. Tools that successfully augment or improve the AI-informed learning health-care system, are embedded into the digital health-care space. This process is organised and analysed by implementation scientists. To mitigate bias and safeguard predictive fairness across the value chain from bedside to bench and back to bedside, regulators and stakeholder organisations supervise the effects of AI-augmented health-care pathways on patients' outcomes, the data curation and model development processes, and the generation of model-based clinical evidence and implementation. Colour gradients for blocks B to E show the patients' journey from lightweight accessible examination methods towards specialised restricted assessment tools in a given health care context. mHealth=mobile health. MBC=measurement-based care. RCT=randomised controlled trial.

more than 80 different target phenotypes.[58] Furthermore, the predictive utility of EMR systems has been demonstrated using the IBM Explorys Platform, a database that collates longitudinal, anonymised EMR datasets from different health-care systems into a standardised format.[52] Based on a training sample of 102 030 individuals, a recurrent neural network learned to predict psychosis by analysing EMR-coded patient journeys and achieved an area-under-the-curve of 0·80 in the external validation data (n=4770). These findings indicate that EMR data composed of information tokens capturing prescriptions, procedures, encounters, admissions, observations, and laboratory results could outperform conventional measures of psychopathology, like observation-based ratings or patient-reported outcomes, both in terms of multivariate complexity and achievable sample size.

However, the strengths of EMR-based predictive feature spaces should be weighed against their obvious limitation: these data are collected to manage individual patient care, not to train predictive algorithms. As such, they reflect patients' footprints across highly variable and potentially idiosyncratic health-care contexts. Many of these event-based features are therefore susceptible to multiple sources of bias at various levels of the respective health-care system[59] (eg, the ordering and timing of laboratory tests is often dictated by the physiological process the test is measuring). Here, bias could stem from the fact that a great deal of information is contained in the context within which a laboratory test is taken[60]—the same numerical value of a creatinine test can have different interpretations for a chronic kidney disease patient and for a patient with acute kidney injury—but most EMR datasets ignore such context to include only numerical values.[61] EMR datasets might also be subject to change and inconsistency if service pathways and practices are updated, calling for ongoing recalibration of predictive algorithms and downstream stratified treatment decisions. Finally, it remains unclear how EMR-based AI algorithms could inform personalised treatments, because many of their predictive features are difficult to interpret or act on—either because they only indirectly relate to the biobehavioural processes underlying the targeted mental conditions, or because they depend on information not recorded in the EMR (panel 2). These unknowns could similarly concern other digital phenotypes that are collected passively by smartphones and other wearables. It is attractive to believe that low-cost, ubiquitous measures of social media use, physical activity, sleeping patterns, and heart rate variability could provide high-throughput substitutes for psychiatric or psychological ascertainment and patient-reported outcomes collected via questionnaires or ecological momentary assessments;[16,62] however, the available clinical and neurobiological evidence supporting such assumptions is still insufficient.[63,64]

Moving forward, the associations between new digital phenotypes and established psychometric, neuropsychological, and neurobiological data domains should be investigated to identify digital features that could be used as accurate proxies for a specific data domain. Thorough psychometric, clinical, and biological validation of digital phenotyping could facilitate big-data-based deep-learning approaches in psychiatric research and ultimately the implementation of diagnostic and prognostic models in clinical real world.

Tensions in selecting the right modelling approach

Besides considering the data requirements for model generation and validation, the focus should also rest on the type of models that might be best suited for implementation. Occam's razor and the robustness requirement for predictive models operating in the clinical real world might indicate a preference for simpler modelling approaches.[65] Conversely, given the promise of performance from deep learning,[66] this class of model might seem more attractive, although the trade-off between shallow and deep learning as a function of sufficiently large, well characterised samples needs to be systematically investigated going forward. In the first paper in this Series, Hauser and colleagues present an extended discussion surrounding this topic.[34]

Regarding implementation, as machine learning algorithms permeate contemporary information systems in domains including health care, studies have noted that machine learning is often presented as being universally applicable and that the application of machine learning without special expertise is actively encouraged.[67] A positive aspect is that universality can allow generalisability and ensure robustness beyond boutique models built for specific populations or applications. Machine learning can enable quick adaptation to novel, previously unseen contexts, by harnessing knowledge from historical data with similar patterns and characteristics (a feature offered by pretrained deep learning models), that has been harnessed to extract information from radiology reports[68] and from clinical narratives, as described.[69] Generalisability is also a particularly important consideration during implementation, because a prevalent criticism of using passive sensing data in mental health, has been that the models do not adapt to new contexts, given the high clinical heterogeneity in the experience of mental illness, and that models tend to be overfit or are biased to the particular sample or context, thus remaining underpowered to support translation and transferability.[70] Furthermore, generalisability is important because it can ensure modelling approaches are robust not only in terms of their performance on unseen data, but also against the messiness of real-world mental health data that often include inaccurate, incomplete, or missing entries, and the subjectivity of symptoms of mental ill health.[71]

Despite these strengths, researchers and practitioners will need to temper their enthusiasm, because although generalisable machine learning might deliver superior technical performance across a wide range of tasks, it could also compound societal inequities when they are naively adopted (eg, without taking into account the given application context and target groups).[72,73] Marginalised or low-resource communities often have unique mental health needs;[74] however, AI models of mental health are often built on majority identity populations or convenience samples of individuals who are likely to be willing to volunteer their personal and sensitive data for algorithmic inferences.[75] Consequently, these models are generalisable in many contexts for majority populations, but might not necessarily be sensitive to the needs, demands, and desires of individuals who are disproportionately disadvantaged.[76] As AI algorithms are implemented in the real world, they will need to factor in the situations of people who are marginalised in terms of access to mental health care and treatment. As mental illness has different effects on gender, racial, and ethnic groups,[77] researchers have also advocated the use of various debiasing approaches (eg, appropriate undersampling or oversampling of data to correct for demographic representation and readjusting model weights) in model development to calibrate performance against inequities.[78] A recent study showed that regional and racial biases present in machine learning models that were trained to rapidly detect COVID-19 on the basis of large-scale real-world patient data, could be successfully mitigated by using generative adversarial networks (panel 2).[79]

Building safeguards into models

It is important to consider the effects of AI in clinical decision making. When AI starts to permeate these decisions, incorporating safeguards into the models becomes even more important. Unscrupulous AI could exacerbate social inequities by aggravating bias or demonstrate prejudice by arriving at assessments for individuals based on race or gender without any grounded clinical rationale.[25] Non-diverse, unrepresentative training data can intensify already biased evaluations. Such biases could arise due to systemic racism, which leads to reduced access to treatment among minority ethnicities, resulting in racialised treatment and interventions (eg Black individuals being at an increased risk of compulsory detention).[80] Machine learning models trained with such data will over-represent specific human biases and propagate inequity and injustice in mental health care, inventing "new classes which do not correlate with protected characteristics".[7]

A safeguard to ensure effective implementation centres around ascertaining when, how, and for whom to prioritise predictive performance over providing transparent and interpretable predictions.[81] These considerations are complicated by the fact that many projects on mental health and AI include interdisciplinary teams, involving computer scientists and clinicians.[82] Different team members could have different goals for a project of this nature, stemming from their domain-specific training—computer scientists tend to prioritise predictive precision, whearas clinical researchers tend to be interested in the mechanisms and variable relationships that contribute to a mental health outcome (see the first paper in this Series by Hauser and colleagues[34] for a discussion of mechanism-driven and mechanism-agnostic models). We argue that, for well defined and circumscribed machine learning application scenarios at the point of care, predictive machine learning might be sufficient, but the problems commonly experienced in mental health care—where objectives are not always clear (eg, which specific treatments or diagnostic criteria best describe a patient's dynamic state), and where a multitude of environmental, clinical, and biological factors affect the outcome—mean an explanation is necessary for accountability.[83] In practice, both prediction and explanation should be balanced;[84] hence, future work needs to consider machine learning methods that are precise and interpretable to concurrently satisfy both interests in the implementation phase.

Regarding safeguards in implementation, researchers, developers, and clinical practitioners should come to a consensus about how much error is acceptable for a machine learning model and how to better explain error to relevant stakeholders, so that the derived information could be actionable (eg, to inform treatment decisions, including medication, hospitalisation, and crisis interventions). Clinical utility is a particularly notable safeguard for implementation considerations because previous research has shown that despite overall high accuracy, carefully chosen trade-offs between specificity and sensitivity of machine learning models will be needed for effective risk stratification.[85] For instance, smartphone sensing data are known to lack sufficient information on adverse events, such as experiences of paranoia. As a result, when such data are used to build algorithms, they might foster imprecise decision making that could even harm an individual because it does not incorporate relevant adverse effects, reactions, or events.[86] Therefore, researchers could consider model error-representation paradigms that adapt to the clinical context, the patients' unique needs, and respective ethical considerations.[13] Future research can explore the understanding and acceptability of error and uncertainty and how best to mitigate it in a principled way. Here, researchers can benefit from theory-driven computational models that incorporate the predictive capabilities of both model accuracy and performance, as well as acceptability and explainability in feature interpretation.

Safeguards extend beyond clinical utility. We argue that equity, in terms of gender, sex, race, ethnicity, socioeconomic status, and importantly mental-health-care access are equally important, especially given that most digital exhaust data from smartphones, wearables, and social media often exclude key contextual information needed to assess mental health at the interpersonal, cultural, social, economic, and environmental levels.[87] To safeguard against predictive biases and inequities in EMR systems like the CRIS system, we advocate a need to adopt strategies built for diverse populations spanning time, geography, and socioeconomic status, actively focusing heterogeneous subpopulations with contrasting individual differences, and leveraging signals from complementary sources that capture a wide range of patterns of behaviour and conditions. Analogous to the aforementioned computer vision and machine translation examples, an ontological database organisation (figure) that explicitly mitigates bias risks through multiple data representations of a given concept could foster the generation of more equitable predictive systems in mental health care.

Robust, adaptive, secure, and transparent tools for a new paradigm of mental health care

Thus far, we have described how implementation considerations need to be a part of the modelling process from the very start of the development of the underlying algorithms. Here we discuss outstanding issues that remain once AI models are incorporated at the point of care. With clinical translation in mind, Sendak and colleagues[88] proposed four steps necessary to facilitate model implementation: (1) design and development to support clinical decision making, (2) evaluation and validation, (3) diffusion and scaling across health-care settings, and (4) sustained engineering to remain current with clinical practice needs. Building upon these guidelines, we specifically ask if mental health care is ready for machine-learning-based diagnostic, prognostic and therapeutic decision making.

Despite smaller studies demonstrating efficacy and feasibility, many of the robotic therapists (panel 3) have encountered important obstacles. Concerns have been expressed about their clinical veracity and potential in supporting improved mental health outcomes for patients,[96] the upfront costs, utility, and potential hazards of specialised information technology infrastructure.[25] Proprietary algorithms, such as that from Facebook, have also been met with criticism due to their lack of transparency, threats due to big-data surveillance, and potentially increased involvement of law enforcement as an approach to crisis mitigation.[97] User consent and monetisation of sensitive data remain unclear issues, as was noted in the public backlash that followed Crisis Text Line after they shared their data with a for-profit organisation.[98] Finally, there are ethical dimensions to consider in deploying conversational AI agents for mental health.[99] Due to the lack of clinical grounding and context in these emerging technologies, responses to emergencies like the disclosure of immediate harm or suicidal ideation are often restricted and sometimes dangerously inappropriate, as Woebot and Wysa have reported.[100]

Readiness for real-world deployment

Important questions concerning model implementation typically relate to the optimal timing of model deployment, and the minimum levels of algorithmic precision required for informing decisions. Here, a gap exists between academically acceptable gains in predictive performance and practical demands (the latter might be higher); scholars have critiqued that the practical gains of many machine learning algorithms, ranging from predicting consumer behaviour to predicting stock market indices, do not necessarily improve much over state-of-the-art prediction models.[101] In the context of mental health, if model performance (eg, in terms of accuracy) matches clinicians' assessments, it could broaden the availability of diagnostic or prognostic services in the community in times when the psychiatric workforce is shrinking or when the demand is increasing, such as during or in the aftermath of the COVID-19 pandemic.[102] Nevertheless, many digital mental health projects aim at achieving improved precision, which is counteracted by the inherent uncertainty of machine learning. In this case, the question is: how do we support graceful failure, when our data or models cannot stand up to the potential use case? Or, how can algorithms be designed to do no harm, as per the Hippocratic oath? Researchers in digital mental health have long been concerned about the considerable harm caused by suboptimal models made publicly available in the interest of open science and reproducibility.[7] A good example comes from a recent study that built models to predict health-care-associated infections and found that attributes associated with risk at one site were protective in another.[103] If such a model were to be indiscriminately implemented at multiple sites and the variables under consideration were to be corrected for during implementation, it could result in harm at those sites where the variables provided protective advantage to patients. Thus, doing no harm would need to include strategies to operationalise and embed the contexts in which the model qualifies to be applied to certain patient groups. Similar to Wiens and colleagues,[42] we suggest that researchers go beyond predictive performance when comparing models, by including an application-specific "analysis of the trade-offs between simpler, faster, and more explainable models versus complex, slower but more accurate models".

Speaking further about real-world deployment, studies have repeatedly shown how individual technology-based mental health interventions do not exist in a vacuum;[104] rather, they form a part of socially situated and structurally influenced pathways to care. Unfortunately, most machine learning solutions in mental health are being developed in silos, without the inclusion of decision makers, experts, and end users.[42] Specifically, individual factors, such as gender identity, sexual orientation, or levels of distress, might influence the types of care that people seek, which in turn affects where and how they look for resources when in need.[105] Consequently, when attempting to access the desired care, barriers rooted in the design of mental health systems (eg, helplines) often limit people's ability to engage with the resource, and further shape their interactions with other forms of care. For instance, people were dissuaded from calling mental health helplines after having poor experiences in therapy and, conversely, they were hesitant to try therapy after poor experiences of helplines.[104] These intersections between individual needs, societal factors, and the design of the care system clearly demand that researchers consider how structural factors impede an individual's

Panel 3: Existing real-world implementations of artificial intelligence in mental health care

AI is increasingly incorporated into digital interventions, particularly web and smartphone apps, to enhance user experience and optimise personalised health care. Although many of these apps use simplistic data science rather than sophisticated AI, they have been recommended as adjuvant care and are being considered for reimbursement by insurers.[89] Second, the newsfeeds and forums of Facebook, Twitter, and Reddit provide rich material for natural language processing systems.[90] Consequently, Facebook has implemented machine learning tools to identify people at risk of self-harm.[91] Similarly, Crisis Text Line, a text-messaging-based crisis counselling hotline has been using machine learning to retrieve content that can signal a person at risk of suicide or self-harm.[92] The goal of such a program is to inform the person on hold to move to the front of the queue to be helped. Another example is the REACH VET program, which has leveraged its AI-ready electronic medical record database to introduce machine learning tools that identify individuals at high risk of suicide.[93] Chekroud and colleagues[94] similarly developed and deployed an AI-based decision support system to improve antidepressant treatment selection.[94] Third, with conversational AI steadily improving, AI agents incorporating sophisticated natural language processing can now simulate a modest conversation employing psychotherapeutic techniques, such as cognitive behavioural therapy.[95] Some examples include Woebot or Wysa, which provide mood tracking and cognitive behavioural therapy modules for the management of depression and anxiety.

For more on **Woebot** see https://woebot.io/

For more on **Wysa** see https://www.wysa.io/

care trajectory, and how the implementation of AI should recognise these complexities. Moreover, as Pendse and colleagues[106] argued, patients' needs might not necessarily be met by adding AI into existing pathways to care, and a Rawlsian notion of approaching justice in access to mental health-care solely on the basis of the existence of resources or institutions might not be enough.

Therefore, to support the translation of machine learning algorithms into technology-driven interventions, implementation research should identify ways to integrate them into both existing and newer pathways so that socioeconomic and political barriers in these pathways are overcome and not aggravated (panel 2).[107] Following Pendse and colleagues,[106] we stress the importance of analysing how interlocking and intersecting societal systems influence who can and cannot access adequate care, and introducing AI on the basis of those insights. AI implementations in mental health systems might consider intelligent matching of resources (eg, helpline volunteers or crisis counsellors) to an individual's needs. Auxiliary support systems that use conversational AI could provide preclinical emotional and informational support to people without a path or ability to access formal care, such as via online forums and communities. Also, these tools could accelerate people's access to more specialised mental health services that use machine-learning-based clinical and neurobiological workflows to guide health-care professionals' treatment decisions (figure).[108]

The future of AI-informed mental health work

Several questions about the future of mental health care should be considered in tandem with implementation approaches. In a future where machine learning algorithms are adopted to administer treatment to patients or support public health decision making, it will be important to ensure that they comply with existing work practices of human experts (psychiatrists and therapists). Specifically, when AI models are implemented, conflicts between human experts and models could arise. Such conflicts could result in situations in which algorithm-based prognostication is in disagreement with clinicians' judgements, patients' expectations, or patients' self-reports. Although clinical decision support based on deep learning methods could be helpful in many health settings (eg, radiology),[109] we argue that the mental health and psychiatry domains might benefit from weak AI, where such discrepancies and conflicts can be controlled more easily, and where there might be a need to develop multiple types of AI applications targeting specific tasks (eg, triggering an intervention *vs* recommending therapy) or symptom categories (eg, Research Domain Criteria informed treatment).[110] Furthermore, integrating these algorithms in existing work practices raises the question of skill acquisition. Recently, researchers have speculated that real-world deployment of AI in psychiatry would also require human experts to understand how machine learning and AI models work.[111] As a solution, researchers have advocated creating new roles within the treatment ecosystem called digital navigators, to enable better assimilation of AI technology in mental health care; essentially, such an individual can serve as an interface between the technology and the clinician, and the technology and hence the patient.[112] In future research, it would be worthwhile to investigate the minimum skill requirements for the sufficiently safe integration of AI algorithms in point-of-care services and thus empirically decide about the need additional health-care roles, such as digital navigators.

We acknowledge that this future of mental health care, where algorithms are a part of the decision making process, is likely to affect human–human relationships[113]—whether this is through the patient–clinician therapeutic alliance or coordination between public health personnel. To this end, the logistics of implementation need to be considered as some patients might perceive the clinicians' use of patients' sensitive data in AI models to be dehumanising[82] or even presenting a conflict of interest because of sensitive non-clinical information (eg, from social media or a GPS sensor) being revealed inadvertently. Eroding trust in AI due to risks of surveillance[114] and poor boundary regulation[115] might further damage the therapeutic relationship or present challenges to candid patient–clinician conversations. Therefore, as we consider how real-world machine learning systems would function over time, we should recognise the importance of sustaining social relationships and protecting them against complicated interactions and compromised interpersonal boundaries. To this end, future research would benefit from engaging with technofeminist scholarship and theories of care.[116] This literature provides a generative analytical framework for conceptualising new types of human–non-human interactions and data-body-machine entanglements. As AI-based technologies in mental health care are increasingly integrated into a wider range of medical settings and everyday activities,[117] these theories offer a starting point to rethink purely technical approaches to AI by rebalancing the focus of the implementation process toward stakeholder needs, interactions, and social concerns (panel 2).

Clinician-machine disconnects

One of the major concerns that have been noted in recent digital health research has been the threat AI algorithms may pose to clinical professions, particularly psychiatrists.[118] With its ability to produce precise AI models, scholars have noted that AI challenges orthodox boundaries of traditional medical expertise.[119] Some medical informaticians argue that the core functions of physicians—such as gathering and monitoring patient information, diagnostics, prognostics, and formulating personal treatment plans—are all susceptible to disintermediation.[120] However, other AI experts predict that health-care professionals will always play a role in medical care, with humans and machines working as team players.[121]

Although we think that the former dystopian scenario is less probable, the introduction of AI in mental health clinical practice might produce scenarios that have not yet been conceived of. For instance, AI could facilitate the work of psychiatrists by highlighting previously inaccessible or less understood symptoms, behavioural patterns, and collateral information that can inform evidence-based decisions, and helping to meet some of the growing demand, as discussed earlier. That said, because of the underlying uncertainty of machine learning models, clinicians will need to guard against well known, but now technology-amplified cognitive biases, such as anchoring, premature closure or reliance on authority (panel 1).[121] To counteract these biases, AI could take the form of mixed-initiative interfaces,[122] which is an interaction strategy whereby human and computer agents respond to each other's creative contributions to improve output and at the most appropriate time. In the context of mental health care, these interaction strategies could consist of models that strengthen the clinicians' and patients' shared decision making abilities by (1) automatically collating informational content from the literature on the patient's condition and prognosis at the point of care, and (2) adapting this information by integrating the users' profiles, preferences, contexts, and uncertainties via AI-powered dialogue techniques.

However, as humans and algorithms share initiatives in such a setup,[124] legal and ethical considerations surrounding the duty to rescue would also have to be navigated. Additional issues might include unclear boundaries of how much freedom there needs to be for the human and the machine, and how to navigate disconnects and disagreements. In a crisis scenario, for instance, whose final word should prevail? If the machine overrides the human in some situations, who will be held liable, given the many harms that can be caused when the chosen intervention is inadequate? To address these challenges, national and international regulatory policies will need to be defined in parallel to model implementation, thus avoiding the increase of health-care disparities through biased personal and institutional use of AI. Legal protection for the clinicians within the implementation framework will also have to be put in place, so that clinicians feel safe when using AI in their everyday work.

Conclusion

As digital technologies enrich mental health care, advancements in AI promise to build a new future in which individuals receive timely, accurate, and contextualised care, and clinicians are empowered to make better decisions (figure). This Series paper discussed the prospects of this future from the perspective of implementation. By integrating implementation considerations from the modelling to the deployment stages of AI-driven health care, we considered the challenges that lie ahead.

Moving forward, the partnership of computational and clinical researchers throughout the modelling-to-implementation pipeline will be needed to improve rigour and mitigate issues surrounding practical use (figure). Additionally, as the field moves towards generalising these findings to new types of data, new populations, or new opportunities for implementation, validity and clinical utility should be carefully maintained throughout the various modelling practices. Navigating these issues requires the need to find ways to balance between the disciplinary tensions that emphasise testing and evaluation of AI technology in different ways. Support from international collaborative and regulatory infrastructures can advance model creation towards proper validation and clinical implementation (panel 2). Furthermore, implementation considerations will also need to represent and augment heuristic decision making, such as how clinicians make diagnostic, prognostic, and therapeutic decisions, and respective mistakes.[125] Modelling these processes will be helpful in informing the next generation of predictive tools that optimally augment clinical decision making toward a cybernetic model of mental health-care.[108] Broadly, we hope our reflections will provide valuable perspectives and directions to the emergent field of digital mental health, especially when it comes to taking the potential of precision mental health tools from research to practical use.

Search selection and strategy criteria

The ideas and content of this work were developed through a series of discussions between the authors, which took place between Sep 21 and Dec 2, 2021. These discussions followed a multistage, iterative process, in which the authors drew upon their distinctive disciplinary trainings, experience, and expertise in computer science, (computational and precision) psychiatry, and digital mental health to identify candidate topics for consideration. This wide angle of perspectives on the development and integration of AI tools in future mental health-care led to the formulation of two primary foci (and respective papers) on modelling and implementation challenges of AI in mental health care. The two primary foci were discussed in-depth, respective content ideas were further grouped into sections, and keywords for literature search on PubMed and Google Scholar were formulated for each section. Text-based keyword search was performed using: ("mental disease" OR "mental illness" OR "mental disorder" OR "depression" OR "psychosis" OR "anxiety" OR "addiction" OR "suicide") AND ("precision medicine" OR "precision psychiatry" OR "computational psychiatry" OR "digital phenotyping" OR "machine learning" OR "pattern recognition" OR "pattern classification" OR "multivariate pattern analysis" OR "deep learning" OR "artificial intelligence") AND ("decision support" OR "diagnosis" OR "prognosis" OR "prediction" OR "mechanism" OR "simulation" OR "social media" OR "social network" OR "smartphones" OR "wearables"). In addition, relevant references from the obtained papers were also considered. Obtained papers (published after 2010) were critically (non-systematically) selected based on the representativeness of the work for the given section topics and the quality of research. The selection of references was discussed and supplemented as required to substantiate the authors' viewpoints from different perspectives. The focus was on papers published in English, but papers in German (ie, the language the authors also spoke) were also considered.

Contributors

NK, TUH, VS, and MDC conceptualised and wrote the original draft of the manuscript. NK and MDC conducted the literature search and reviewed the identified publications. NK, TUH, and MDC reviewed and edited the revised versions of the manuscript. VS supported the preparation and submission of the manuscript with administrative work. NK created the visual elements of the manuscript. MDC supervised the entire manuscript writing and revision process.

Declaration of interests

NK holds an honorary, unpaid advisory board position at the Spring Care and holds the patent US20160192889A1. The processing fees for this Review were provided by institutional funds from Ludwig Maximilian University (Munich, Germany). NK is supported by the National Institutes of Health (NIH), the Wellcome Trust, the German Innovation Fund (CARE project), the German Federal Ministry of Education and Research (COMMITMENT and BEST projects), and ERA PerMed (IMPLEMENT project). TUH collaborates on a grant funded by Koa Health; consults for Limbic; and has received funding from National Alliance for Research on Schizophrenia and Depression, Leverhulme Trust, and the European Research Council. TUH is also supported by a Sir Henry Dale Fellowship (number 211155/Z/18/Z) from the Wellcome Trust and the Royal Society, a grant from the Jacobs Foundation (number 2017-1261-04), the Medical Research Foundation, and a 2018 NARSAD Young Investigator grant (number 27023) from the Brain & Behavior Research Foundation. MDC has received funding from NIH (grant number R01MH117172); National Science Foundation (grant numbers 2027689 and 1816403), and the Injury Prevention Research Center, funded by the US Centers for Disease Control and Prevention, and Emory University. VS has received an individual fellowship from the EU's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant (agreement number 895213).

Acknowledgments

We thank the current Editor-in-Chief of *Lancet Digital Health*, Rupa Sarkar, and the former Editor-in-Chief of *Lancet Psychiatry*, Niall Boyce, for the initiating and providing valuable feedback through the conceptualisation and writing of this miniseries.

References

1 Svensson B, Hansson L. How mental health literacy and experience of mental illness relate to stigmatizing attitudes and social distance towards people with depression or psychosis: a cross-sectional study. *Nord J Psychiatry* 2016; **70:** 309–13.

2 Clarke T, Barwick M. Editorial perspective: A call to collective action–improving the implementation of evidence in children and young people's mental health. *Child Adolesc Ment Health* 2021; **26:** 73–75.

3 Blackstone EH. Precision medicine versus evidence-based medicine. *Circulation* 2019; **140:** 1236–38.

4 Beckmann JS, Lew D. Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities. *Genome Med* 2016; **8:** 134.

5 Schoeler T, Petros N, Forti MD, et al. Poor medication adherence and risk of relapse associated with continued cannabis use in patients with first-episode psychosis: a prospective analysis. *Lancet Psychiatry* 2017; **4:** 627–33.

6 Chien WT, Mui J, Gray R, Cheung E. Adherence therapy versus routine psychiatric care for people with schizophrenia spectrum disorders: a randomised controlled trial. *BMC Psychiatry* 2016; **16:** 42.

7 Carr S. 'AI gone mental': engagement and ethics in data-driven technology for mental health. *J Ment Health Abingdon Engl* 2020; **29:** 125–30.

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

9 Cho G, Yim J, Choi Y, Ko J, Lee S-H. Review of machine learning algorithms for diagnosing mental illness. *Psychiatry Investig* 2019; **16:** 262–69.

10 Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. *Psychol Med* 2019; **49:** 1426–48.

11 Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. *Annu Rev Clin Psychol* 2018; **14:** 91–118.

12 Fusar-Poli P, Hijazi Z, Stahl D, Steyerberg EW. The science of prognosis in psychiatry: a review. *JAMA Psychiatry* 2018; **75:** 1289–97.

13 Graham S, Depp C, Lee EE, et al. Artificial intelligence for mental health and mental illnesses: an overview. *Curr Psychiatry Rep* 2019; **21:** 116.

14 Kambeitz J, Kambeitz-Ilankovic L, Leucht S, et al. Detecting neuroimaging biomarkers for schizophrenia: a meta-analysis of multivariate pattern recognition studies. *Neuropsychopharmacology* 2015; **40:** 1742–51.

15 Korda AI, Andreou C, Borgwardt S. Pattern classification as decision support tool in antipsychotic treatment algorithms. *Exp Neurol* 2021; **339:** 113635.

16 Benoit J, Onyeaka H, Keshavan M, Torous J. Systematic review of digital phenotyping and machine learning in psychosis spectrum illnesses. *Harv Rev Psychiatry* 2020; **28:** 296–304.

17 Richter T, Fishbain B, Richter-Levin G, Okon-Singer H. Machine learning-based behavioral diagnostic tools for depression: advances, challenges, and future directions. *J Pers Med* 2021; **11:** 957.

18 Lin E, Lin C-H, Lane H-Y. Machine learning and deep learning for the pharmacogenomics of antidepressant treatments. *Clin Psychopharmacol Neurosci* 2021; **19:** 577–88.

19 Kambeitz J, Cabral C, Sacchet MD, et al. Detecting neuroimaging biomarkers for depression: a meta-analysis of multivariate pattern recognition studies. *Biol Psychiatry* 2017; **82:** 330–38.

20 Nigg JT, Karalunas SL, Feczko E, Fair DA. Toward a revised nosology for attention-deficit/hyperactivity disorder heterogeneity. *Biol Psychiatry Cogn Neurosci Neuroimaging* 2020; **5:** 726–37.

21 Goldstein-Piekarski AN, Holt-Gosselin B, O'Hora K, Williams LM. Integrating sleep, neuroimaging, and computational approaches for precision psychiatry. *Neuropsychopharmacol* 2020; **45:** 192–204.

22 Feczko E, Miranda-Dominguez O, Marr M, Graham AM, Nigg JT, Fair DA. The heterogeneity problem: approaches to identify psychiatric subtypes. *Trends Cogn Sci* 2019; **23:** 584–601.

23 Kim Y-K, Na K-S. Application of machine learning classification for structural brain MRI in mood disorders: critical review from a clinical perspective. *Prog Neuropsychopharmacol Biol Psychiatry* 2018; **80:** 71–80.

24 Bzdok D, Karrer TM, Habel U, Schneider F. Big data approaches in psychiatry: examples in depression research. *Nervenarzt* 2018; **89:** 869–74 (in German).

25 Lee EE, Torous J, De Choudhury M, et al. Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. *Biol Psychiatry Cogn Neurosci Neuroimaging* 2021; **6:** 856–64.

26 Ienca M, Ignatiadis K. Artificial intelligence in clinical neuroscience: methodological and ethical challenges. *AJOB Neurosci* 2020; **11:** 77–87.

27 Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell* 2019; **1:** 389–99.

28 Huys Q JM, Maia TV, Frank MJ. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat Neurosci* 2016; **19:** 404–13.

29 Hauser TU, Will G-J, Dubois M, Dolan RJ. Annual research review: developmental computational psychiatry. *J Child Psychol Psychiatry* 2019; **60:** 412–26.

30 Brown S-A. Building SuperModels: emerging patient avatars for use in precision and systems medicine. *Front Physiol* 2015; **6:** 318.

31 Stephan KE, Schlagenhauf F, Huys Q JM, et al. Computational neuroimaging strategies for single patient predictions. *NeuroImage* 2017; **145:** 180–99.

32 Hauser TU, Iannaccone R, Ball J, et al. Role of the medial prefrontal cortex in impaired decision making in juvenile attention-deficit/hyperactivity disorder. *JAMA Psychiatry* 2014; **71:** 1165–73.

33 Gilbert JR, Symmonds M, Hanna MG, Dolan RJ, Friston KJ, Moran RJ. Profiling neuronal ion channelopathies with non-invasive brain imaging and dynamic causal models: case studies of single gene mutations. *NeuroImage* 2016; **124:** 43–53.

34 Hauser TU, Skvortsova V, Choudhury MD, Koutsouleris N. The promise of a model-based psychiatry: building computational models of mental ill-health. *Lancet Digit Health* 2022; published online Oct 10. https://doi.org/10.1016/S2589-7500(22)00152-2.

35 Hänsel K, Lin IW, Sobolev M, et al. Utilizing Instagram data to identify usage patterns associated with schizophrenia spectrum disorders. *Front Psychiatry* 2021; **12:** 691327.

36 Ricard BJ, Marsch LA, Crosier B, Hassanpour S. Exploring the utility of community-generated social media content for detecting depression: an analytical study on Instagram. *J Med Internet Res* 2018; **20:** e11817.

37 Kim S, Lee K. Screening for depression in mobile devices using Patient Health Questionnaire-9 (PHQ-9) data: a diagnostic meta-analysis via machine learning methods. *Neuropsychiatr Dis Treat* 2021; **17:** 3415–30.

38 Scott K, Lewis CC. Using measurement-based care to enhance any treatment. *Cogn Behav Pract* 2015; **22:** 49–59.

39 Lewis CC, Boyd M, Puspitasari A, et al. Implementing measurement-based care in behavioral health: a review. *JAMA Psychiatry* 2019; **76:** 324–35.

40 Salazar de Pablo G, Studerus E, Vaquerizo-Serrano J, et al. Implementing precision psychiatry: a systematic review of individualized prediction models for clinical practice. *Schizophr Bull* 2021; **47:** 284–97.

41 Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging* 2018; **3:** 223–30.

42 Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019; **25:** 1337–40.

43 D'Alfonso S. AI in mental health. *Curr Opin Psychol* 2020; **36:** 112–17.

44 Fakhoury M. Artificial intelligence in psychiatry. *Adv Exp Med Biol* 2019; **1192:** 119–25.

45 Carroll KM, Rounsaville BJ. Computer-assisted therapy in psychiatry: be brave—it's a new world. *Curr Psychiatry Rep* 2010; **12:** 426–32.

46 Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis* 2015; **115:** 211–52.

47 Rivera-Trigueros I. Machine translation systems and quality assessment: a systematic review. *Lang Resour Eval* 2022; **56:** 593–619.

48 Aulamo M, Sulubacak U, Virpioja S, Tiedemann J. OpusTools and parallel corpus diagnostics. Proceedings of the 12th Language Resources and Evaluation Conference. France: European Language Resources Association, 2020. 3782–89.

49 Popel M, Tomkova M, Tomek J, et al. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nat Commun* 2020; **11:** 4381.

50 Kitanishi Y, Fujiwara M, Binkowitz B. Patient journey through cases of depression from claims database using machine learning algorithms. *PloS One* 2021; **16:** e0247059.

51 Hansen L, Enevoldsen KC, Bernstorff M, Nielbo KL, Danielsen AA, Østergaard SD. The PSYchiatric clinical outcome prediction (PSYCOP) cohort: leveraging the potential of electronic health records in the treatment of mental disorders. *Acta Neuropsychiatr* 2021; **33:** 323–30.

52 Raket LL, Jaskolowski J, Kinon BJ, et al. Dynamic ElecTronic hEalth reCord deTection (DETECT) of individuals at risk of a first episode of psychosis: a case-control development and validation study. *Lancet Digit Health* 2020; **2:** e229–39.

53 Fortney JC, Unützer J, Wrenn G, et al. A tipping point for measurement-based care. *Psychiatr Serv* 2017; **68:** 179–88.

54 Murphy JK, Michalak EE, Liu J, et al. Barriers and facilitators to implementing measurement-based care for depression in Shanghai, China: a situational analysis. *BMC Psychiatry* 2021; **21:** 430.

55 Cheah PY, Jatupornpimol N, Hanboonkunupakarn B, et al. Challenges arising when seeking broad consent for health research data sharing: a qualitative study of perspectives in Thailand. *BMC Med Ethics* 2018; **19:** 86.

56 Wei W-Q, Leibson CL, Ransom JE, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *J Am Med Inform Assoc* 2012; **19:** 219–24.

57 Williams BA. Constructing epidemiologic cohorts from electronic health record data. *Int J Environ Res Public Health* 2021; **18:** 13193.

58 Chilman N, Song X, Roberts A, et al. Text mining occupations from the mental health electronic health record: a natural language processing approach using records from the Clinical Record Interactive Search (CRIS) platform in south London, UK. *BMJ Open* 2021; **11:** e042274.

59 Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible sources of bias in primary care electronic health record data use and reuse. *J Med Internet Res* 2018; **20:** e9134.

60 Pivovarov R, Albers DJ, Sepulveda JL, Elhadad N. Identifying and mitigating biases in EHR laboratory tests. *J Biomed Inform* 2014; **51:** 24–34.

61 Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ* 2018; **361:** k1479.

62 Torous J, Staples P, Barnett I, Sandoval LR, Keshavan M, Onnela J-P. Characterizing the clinical relevance of digital phenotyping data quality with applications to a cohort with schizophrenia. *Npj Digit Med* 2018; **1:** 1–9.

63 Cohen AS, Schwartz E, Le T, et al. Validating digital phenotyping technologies for clinical use: the critical importance of "resolution". *World Psychiatry* 2020; **19:** 114–15.

64 Marsch LA. Digital health data-driven approaches to understand human behavior. *Neuropsychopharmacology* 2021; **46:** 191–96.

65 Babic B, Gerke S, Evgeniou T, Cohen IG. Beware explanations from AI in health care. *Science* 2021; **373:** 284–86.

66 Su C, Xu Z, Pathak J, Wang F. Deep learning in mental health outcome research: a scoping review. *Transl Psychiatry* 2020; **10:** 1–26.

67 Heuer H, Jarke J, Breiter A. Machine learning in tutorials—universal applicability, underinformed application, and other misconceptions. *Big Data Soc* 2021; **8:** 20539517211017590.

68 López-Úbeda P, Díaz-Galiano MC, Ureña-López LA, Martín-Valdivia MT. Pre-trained language models to extract information from radiological reports. Conference and Labs of the Evaluation Forum. Sept 21–24, 2021 (abstr 66).

69 Wei Q, Ji Z, Si Y, et al. Relation extraction from clinical narratives using pre-trained language models. AMIA Annual Symposium Proceedings. American Medical Informatics Association, 2019: 1236–45.

70 Trifan A, Oliveira M, Oliveira JL. Passive sensing of health outcomes through smartphones: systematic review of current solutions and possible limitations. *JMIR Mhealth Uhealth* 2019; **7:** e12649.

71 Mohr DC, Shilton K, Hotopf M. Digital phenotyping, behavioral sensing, or personal sensing: names and transparency in the digital age. *NPJ Digit Med* 2020; **3:** 1–2.

72 Smith MJ, Axler R, Bean S, Rudzicz F, Shaw J. Four equity considerations for the use of artificial intelligence in public health. *Bull World Health Organ* 2020; **98:** 290.

73 Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. *Annu Rev Biomed Data Sci* 2021; **4:** 123–44.

74 Schueller SM, Hunter JF, Figueroa C, Aguilera A. Use of digital mental health for marginalized and underserved populations. *Curr Treat Options Psychiatry* 2019; **6:** 243–55.

75 Wendler D, Kington R, Madans J, et al. Are racial and ethnic minorities less willing to participate in health research? *PLoS Med* 2006; **3:** e19.

76 Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. *J Glob Health* 2019; **9:** 020318.

77 Allen J, Balfour R, Bell R, Marmot M. Social determinants of mental health. *Int Rev Psychiatry Abingdon Engl* 2014; **26:** 392–407.

78 Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. *Nat Med* 2020; **26:** 16–17.

79 Yang J, Soltan AAS, Yang Y, Clifton DA. Algorithmic fairness and bias mitigation for clinical machine learning: insights from rapid COVID-19 diagnosis by adversarial learning. *medRxiv* 2022; published online Jan 14. https://www.medrxiv.org/content/10.1101/2022.01.13.22268948v1 (preprint).

80 Barnett P, Mackay E, Matthews H, et al. Ethnic variations in compulsory detention under the Mental Health Act: a systematic review and meta-analysis of international data. *Lancet Psychiatry* 2019; **6:** 305–17.

81 Adadi A, Berrada M. Explainable AI for healthcare: from black box to interpretable models. Embedded Systems and Artificial Intelligence. Singapore: Springer, 2020: 327–37.

82 Chancellor S, De Choudhury M. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ Digit Med* 2020; **3:** 1–11.

83 Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *N Engl J Med* 2017; **376:** 2507.

84 Pawar U, O'Shea D, Rea S, O'Reilly R. Explainable AI in healthcare. 2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA). Dublin, Ireland. IEEE 2020: 1–2.

85 Wilkinson J, Arnold KF, Murray EJ, et al. Time to reality check the promises of machine learning-powered precision medicine. *Lancet Digit Health* 2020; **2:** e677–80.

86 Bradstreet S, Allan S, Gumley A. Adverse event monitoring in mHealth for psychosis interventions provides an important opportunity for learning. *J Ment Health Abingdon Engl* 2019; **28:** 461–66.

87 Chancellor S, Birnbaum ML, Caine ED, Si-lenzio VM, De Choudhury M. A taxonomy of ethical tensions in inferring mental health states from social media. Proceedings of the conference on fairness, accountability, and transparency. Atlanta, GA, USA. Association for Computer Machinary 2019: 79–88.

88 Sendak M, D'Arcy J, Kashyap S, et al. A path for translation of machine learning products into healthcare delivery. *EMJ Innov* 2020; **10:** 19–00172.

89 Powell AC, Torous JB, Firth J, Kaufman KR. Generating value with mental health apps. *BJPsych Open* 2020; **6:** e16.

90 Birnbaum ML, Ernala SK, Rizvi AF, et al. Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from Facebook. *NPJ Schizophr* 2019; **5:** 17.

91 Card C. How Facebook AI helps suicide prevention. 2018. https://about.fb.com/news/2018/09/inside-feed-suicide-prevention-and-ai/ (accessed May 13, 2022).

92 Thompson LK, Sugg MM, Runkle JR. Adolescents in crisis: a geographic exploration of help-seeking behavior using data from Crisis Text Line. *Soc Sci Med* 2018; **215:** 69–79.

93 Reger GM, McClure ML, Ruskin D, Carter SP, Reger MA. Integrating predictive modeling into mental health care: an example in suicide prevention. *Psychiatr Serv* 2019; **70:** 71–74.

94 Chekroud AM, Zotti RJ, Shehzad Z, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry* 2016; **3:** 243–50.

95 D'Alfonso S, Santesteban-Echarri O, Rice S, et al. Artificial intelligence-assisted online social therapy for youth mental health. *Front Psychol* 2017; **8:** 796.

96 Neary M, Schueller SM. State of the field of mental health apps. *Cogn Behav Pract* 2018; **25:** 531–37.

97 Roemmich K, Andalibi N. Data subjects' conceptualizations of and attitudes toward automatic emotion recognition-enabled wellbeing interventions on social media. *Proc ACM Hum Comput Interact* 2021; **5:** 1–34.

98 Hick J, Lawler R. Crisis Text Line stops sharing conversation data with AI company. *The Verge*. Feb 1, 2022. https://www.theverge.com/2022/1/31/22906979/crisis-text-line-loris-ai-epic-privacy-mental-health (accessed Nov 5, 2022).

99 Kretzschmar K, Tyroll H, Pavarini G, Manzini A, Singh I, NeurOx Young People's Advisory Group. Can your phone be your therapist? Young people's ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support. *Biomed Inform Insights* 2019; **11:** 1178222619829083.

100 White G. Child advice chatbots fail to spot sexual abuse. British Broadcasting Corporation; published online Nov 12, 2018. https://www.bbc.com/news/technology-46507900 (accessed May 13, 2022).

101 Goel S, Hofman JM, Lahaie S, Pen-nock DM, Watts DJ. Predicting consumer behavior with Web search. *Proc Natl Acad Sci* 2010; **107:** 17486–90.

102 Satiani A, Niedermier J, Satiani B, Svendsen DP. Projected workforce of psychiatrists in the United States: a population analysis. *Psychiatr Serv* 2018; **69:** 710–13.

103 Oh J, Makar M, Fusco C, et al. A generalizable, data-driven approach to predict daily risk of *Clostridium difficile* infection at two large academic health centers. *Infect Control Hosp Epidemiol* 2018; **39:** 425–33.

104 Pendse SR, Niederhoffer K, Sharma A. Cross-cultural differences in the use of online mental health support forums. *Proc ACM Hum Comput Interact* 2019; **3:** 1–29.

105 Pendse SR, Nkemelu D, Bidwell NJ, et al. From treatment to healing: envisioning a decolonial digital mental health. CHI Conference on Human Factors in Computing Systems. New York, NY: Association for Computing Machinery, 2022: 1–23.

106 Pendse SR, Sharma A, Vashistha A, De Choudhury M, Kumar N. "Can I not be suicidal on a Sunday?": understanding technology-mediated pathways to mental health support. New York, NY, USA. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021: 1–16.

107 Torous J, Roberts LW. Needed innovation in digital health and smartphone applications for mental health: transparency and trust. *JAMA Psychiatry* 2017; **74:** 437–38.

108 Koutsouleris N, Dwyer DB, Degenhardt F, et al. Multimodal machine learning workflows for prediction of psychosis in patients with clinical high-risk syndromes and recent-onset depression. *JAMA Psychiatry* 2021; **78:** 195–209.

109 Strohm L, Hehakaya C, Ranschaert ER, Boon WP, Moors EH. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. *Eur Radiol* 2020; **30:** 5525–32.

110 Cuthbert BN, Insel TR. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Med* 2013; **11:** 126.

111 Rauseo-Ricupero N, Henson P, Agate-Mays M, Torous J. Case studies from the digital clinic: integrating digital phenotyping and clinical practice into today's world. *Int Rev Psychiatry* 2021; **33:** 394–403.

112 Wisniewski H, Gorrindo T, Rauseo-Ricupero N, Hilty D, Torous J. The role of digital navigators in promoting clinical care and technology integration into practice. *Digit Biomark* 2020; **4:** 119.

113 Fisher CE, Appelbaum PS. Beyond googling: the ethics of using patients' electronic footprints in psychiatric practice. *Harv Rev Psychiatry* 2017; **25:** 170–79.

114 Lee MK, Rich K. Who is included in human perceptions of AI?: Trust and perceived fairness around healthcare AI and cultural mistrust. New York, NY, USA. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems 2021: 1–14.

115 Lopez A, Schwenk S, Schneck CD, Grif-fin RJ, Mishkind MC. Technology-based mental health treatment and the impact on the therapeutic alliance. *Curr Psychiatry Rep* 2019; **21:** 1–7.

116 Lupton D. Toward a more-than-human analysis of digital health: inspirations from feminist new materialism. *Qual Health Res* 2019; **29:** 1998–2009.

117 Bartoletti I. AI in healthcare: ethical and privacy challenges. Conference on Artificial Intelligence in Medicine in Europe. Singapore: Springer, 2019: 7–10.

118 Brown C, Story GW, Mourão-Miranda J, Baker JT. Will artificial intelligence eventually replace psychiatrists? *Br J Psychiatry* 2021; **218:** 131–34.

119 Asan O, Bayrak AE, Choudhury A, et al. Artificial intelligence and human trust in healthcare: focus on clinicians. *J Med Internet Res* 2020; **22:** e15154.

120 Rodriguez F, Scheinker D, Harrington RA. Promise and perils of big data and artificial intelligence in clinical medicine and biomedical research. *Circ Res* 2018; **123:** 1282–84.

121 Fiske A, Henningsen P, Buyx A, et al. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *J Med Internet Res* 2019; **21:** e13216.

122 Committee on the Learning Health Care System in America, Institute of Medicine. Best Care at Lower Cost: The Path to Continuously Learning Health Care in America. Washington: National Academies Press, 2013. http://www.ncbi.nlm.nih.gov/books/NBK207225/ (accessed May 14, 2022).

123 Horvitz E. Principles of mixed-initiative user interfaces. New York, NY, USA. Proceedings of the SIGCHI conference on Human Factors in Computing Systems. 1999: 159–66.

124 Horvitz EJ. Reflections on challenges and promises of mixed-initiative interaction. *AI Mag* 2007; **28:** 19–22.

125 Todd PM, Gigerenzer G. What is ecological rationality? Ecological Rationality. Oxford: Oxford University Press, 2012.