Current Psychiatry Reports
https://doi.org/10.1007/s11920-022-01378-5

PSYCHIATRY IN THE DIGITAL AGE (J SHORE, SECTION EDITOR)

Expectations for Artificial Intelligence (AI) in Psychiatry

Scott Monteith[1] · Tasha Glenn[2] · John Geddes[3] · Peter C. Whybrow[4] · Eric Achtyes[5,6] · Michael Bauer[7]

Accepted: 15 September 2022

Abstract

Purpose of Review Artificial intelligence (AI) is often presented as a transformative technology for clinical medicine even though the current technology maturity of AI is low. The purpose of this narrative review is to describe the complex reasons for the low technology maturity and set realistic expectations for the safe, routine use of AI in clinical medicine.

Recent Findings For AI to be productive in clinical medicine, many diverse factors that contribute to the low maturity level need to be addressed. These include technical problems such as data quality, dataset shift, black-box opacity, validation and regulatory challenges, and human factors such as a lack of education in AI, workflow changes, automation bias, and deskilling. There will also be new and unanticipated safety risks with the introduction of AI.

Summary The solutions to these issues are complex and will take time to discover, develop, validate, and implement. However, addressing the many problems in a methodical manner will expedite the safe and beneficial use of AI to augment medical decision making in psychiatry.

Keywords Artificial intelligence · Machine learning · Psychiatry · Technology maturity

Introduction

Artificial intelligence (AI) is presented today as a transformative technology in the delivery of healthcare that will improve the quality of care and increase physician efficiency. Amid the promise and excitement, there is little focus on realistic expectations for the time and effort required for the successful adaptation of new technology. In 1987, Nobel prize winning economist Robert Solow observed, "You can see the computer age everywhere but in the productivity statistics" [1]. This productivity paradox, a delay of years or even decades between the adoption of a new technology and productivity increases, has been found across all economic sectors [2]. The productivity paradox after investment in computer technology was repeatedly detected in the healthcare sector [3–5]. Successful adaption of any major technology requires complementary investment in process engineering, organizational changes, and widespread training on new skills and techniques, before there is an increase in productivity [2, 6].

In contrast to the high expectations for AI in medicine, the widespread adoption of AI products is associated with a similar productivity paradox [6, 7]. For example, the USA productivity growth from 2005 to 2019 was half of that from 1995 to 2004 [6, 7]. No amount of enthusiasm can overcome the difficulties in deploying new technology in any industry, especially medicine. The many unique and difficult issues involved in the delivery of healthcare will delay productivity increases with the implementation of AI technology.

The purpose of this narrative review is to discuss major challenges to successfully implementing AI in clinical

This article is part of the Topical Collection on *Psychiatry in the Digital Age*

✉ Scott Monteith
monteit2@msu.edu

1 Michigan State University College of Human Medicine, Traverse City Campus, Traverse City, MI 49684, USA

2 ChronoRecord Association, Fullerton, CA, USA

3 Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford, UK

4 Department of Psychiatry and Biobehavioral Sciences, Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles (UCLA), Los Angeles, CA, USA

5 Michigan State University College of Human Medicine, Grand Rapids, MI 49684, USA

6 Network180, Grand Rapids, MI, USA

7 Department of Psychiatry and Psychotherapy, University Hospital Carl Gustav Carus Medical Faculty, Technische Universität Dresden, Dresden, Germany

Published online: 10 October 2022

practice, with a focus on psychiatry. The issues relate to the maturity of AI technology, physician attitudes toward and knowledge of technology, workflow impacts, ongoing organizational support, patient safety, and problems of treating mental illness.

What AI is and is Not

AI is often promoted as the solution for many problems. Despite this assertion, the current AI technology does not possess human general intelligence, high-level reasoning, common sense, or the superhuman intelligence of science fiction [8–10]. Current AI technology is made possible by the massive databases available from the continuous creation and collection of data, including numbers, text, video, and audio, from diverse, interconnected computers and smart, everyday devices embedded with computing technology. Commercial AI products do not involve high-level reasoning or thought, but typically provide services based on large datasets that may augment human intelligence and decision making [8, 11•]. For example, after human evaluation, results from a search engine result may augment knowledge, or a spell checker may improve a document. In the commercial world, business models using AI tie decisions to large-scale datasets and focus on profits. The product recommender system used by Amazon is one example [12]. A profitable AI model with known errors may be acceptable to the corporate decision makers, regardless of inconvenience or costs to some customers [13, 14].

Most AI is based on machine learning (ML), including in medicine. ML blends concepts from many disciplines including computer science, statistics, and linguistics and includes many subsets such as deep learning [11•, 15–17]. ML algorithms use large training datasets to determine the best model for predicting an outcome, but the model itself remains an opaque "black box" [18–20]. ML has had the greatest success in situations with a very large signal-to-noise ratio (few data errors), such as visual or voice pattern recognition, or games with concrete rules [10, 20]. In contrast to ML, traditional statistical methods can be successfully estimated using both large and small datasets, but the model variables must be specified in advance. The focus of traditional statistical methods such as logistic regression is on understanding the relationships between independent data variables and the outcome or dependent variable. One example is using hypothesis testing to evaluate outcomes of controlled, clinical trials [21]. However, vast amounts of diverse data are increasingly available in medicine, including provider data from EHR, imaging, and genomics; patient data from Internet, smartphone and wearable activities, and data from non-medical sources such as government agencies [22, 23]. ML offers new approaches to the practice of psychiatry to analyze the massive datasets for prediction of the diagnosis, treatment selection, and illness course [24–26].

Maturity of AI Technology

The productivity paradox is related to the maturity of AI technology. When considering the introduction of AI in a safety–critical setting such as patient care, it is important to appreciate the current state of maturity of AI technology [27]. The technology readiness level (TRL) scale was developed by NASA in the 1970s to evaluate technical maturity, and has evolved to contain 9 levels [28]. The TRL scale produces a consistent measure to monitor progress in the development of new technology, promotes testing and verification to assess maturity, and provides assurance that the technology will function as intended [28, 29]. In the TRL scale, levels 1–4 refer to basic research in the laboratory, levels 5–6 to demonstrating the technology in a representative environment, and levels 7–9 to testing, validation, and successful deployment in an operational environment [28, 29]. The TRL scale is widely used internationally by governments and diverse industries [30].

The TRL scale was recently customized for rating the maturity of ML projects in a clinical setting, with TRL levels 3–4 referring to model prototyping and development and level 5 referring to model validation on other than the training population [31]. In an evaluation of 172 studies in intensive care medicine using ML, 160 (93%) scored a TRL level of 4 or below, with none successfully integrated into routine clinical care at level 9 [31]. In another evaluation of 494 studies in intensive care medicine using AI, 441 studies (89.3%) scored level 4 or below, 35 (7.1%) at level 5, with none successfully integrated into routine clinical care at level 9 [32•]. Technical maturity is very difficult to achieve for a new technology. The distance between academic discovery and successful commercialization is referred to as the "valley of death" [33–35]. The failure to advance the technology typically occurs between TRL levels 4–7, a stage often viewed as too applied for academia and too risky for commercial funding.

Recent Clinical Experiences Suggest AI Technology is Not Mature

Although there have been successes using AI in medicine, it is not widely deployed in routine clinical practice. There have also been unexpected results, errors, and failures using AI algorithms in many fields including medicine, as fundamental technological properties are being learned. AI image detection algorithms are routinely described as fragile and

brittle since very small changes to the data may result in incorrect labels [27, 36], as demonstrated by a change to just 2% of pixels in an image [37], and even after a one pixel attack [38]. AI image detection algorithms may incorrectly learn to include confounding variables, such as the label "PORTABLE" from a chest X-ray machine when diagnosing pneumonia [39], a ruler present in an image when diagnosing malignant skin cancer [40, 41], the scanner model and brand, and orders marked "urgent" when diagnosing a hip fracture [42], and the chest tube used for treatment when identifying a pneumothorax [43]. The inclusion of confounding variables, often not clinical, may limit generalization, lead to incorrect findings, and emphasize the need to further understand and validate AI algorithms.

Automated speech recognition is impacted by individual accents, by historical and cultural stereotypes, and a lack of diversity in ML training data that results in disparities and biases by race and for non-native English speakers [44–47]. Racial bias was found in automated measurements of speech coherence designed to identify thought disorders [48, 49]. The amount of environmental noise including indoor background conversations decreases the reliability of speech recognition systems [50, 51]. A review of speech recognition for clinical documentation across specialties found the word error rate ranged from 7.4 to 38.7%, and the percentage of documents with errors ranged from 4.8 to 71% [52]. In an analysis of notes generated by speech recognition dictated by emergency department physicians, 71% contained errors and 15% of errors were potentially critical [53]. Conversational clinical speech recognition is even more complex [54], with estimates of word error rates using 7 commercial products ranging between 34 and 65% [55]. In a psychotherapy setting, the word error rate in psychiatrist identified harm-related sentences using a commercial product was 34% [56]. Many challenges and biases remain that impact the safe use of speech recognition in clinical practice.

The most advanced use of medical AI is in radiology with over 150 products for radiology containing AI algorithms cleared for use by the FDA [57]. Additionally, many sites use locally developed rather than commercial AI algorithms [58]. However, systematic reviews of medical imaging studies found little evidence that AI-based CDS improved clinician diagnostic performance [59], and there are few randomized and prospective trials behind AI claims [60]. In a 2020 survey from the American College of Radiologists with 1427 respondents, about 30% were using AI in clinical practice, but 94% of these rated the performance of AI as inconsistent [58]. The ECRI 2022 Top Ten Technology Hazards List includes AI-based image reconstruction, which can distort images reconstructed from the raw data obtained during MRI, CT, or other scans [61–63]. In addition to complex technical issues related to radiation dose, image capture, and reconstruction, radiology faces the same AI related problems found throughout medicine including unrepresentative and biased training data, workflow changes, productivity impacts, lack of external validation and validation standards, and performance deterioration over time [58, 64, 65]. A review of 62 studies to detect COVID-19 from chest X-rays and computed tomography images concluded that none were ready for clinical use due to methodological flaws and/or underlying biases [66].

Impediments to Maturity

There are impediments to the maturity of AI technology for routine use in clinical medicine that need to be recognized and directly addressed. Some of the key issues are briefly described.

Data Quality

The success of an AI algorithm is tied to the training data [67]. EHR and claim data were not developed for medical research, and there are many data quality issues related to missing data, inaccuracy, coding errors, biases, timeliness, redundancies, types of healthcare facilities, provenance or ownership trails, and lack of interoperability between vendor products [22, 68, 69]. Various factors contribute to the biases in EHR data. These include a lack of patient diversity, missing or discordant data on race and ethnicity, confounding medical interventions, oversampling of the sickest, fractured care across multiple providers, loss to follow-up, divergent processes within healthcare systems, measurement errors, and differences between recommended treatments in high versus low-resource settings [70–78].

There are special data quality concerns for psychiatry. Behavioral health related data are often missing or inaccurate in the EHR, including diagnoses, visits, and hospitalizations [79]. For example, studies have reported discrepancies and missing diagnoses in the EHR for PTSD [80–82] and a lack of documentation of suicidal ideation or attempts [83]. Stigmatizing symptoms may be underreported by the patient, and symptoms and diagnoses intentionally omitted by the physician [78, 82, 84, 85]. In large studies in the USA, 27–60% of patients prescribed psychotropic medications did not have a psychiatric diagnosis [86–89]. The transition from ICD-9-CM to ICD-10-CM in the USA in 2015 was associated with some coding changes, including reports of a decrease in the diagnosis of schizophrenia from 48 to 33 per 100,000 [90], and an abrupt increase in hospital stays for opioid and alcohol abuse [91, 92]. The electronic transfer of health information after discharge from inpatient psychiatric units occurs less often than from other areas of the hospital [93]. In an international review, less than half of studies in mental health settings reported implementing an EHR

[94]. Additionally, many people seek help for mental health problems in non-medical settings that are not integrated into the EHR [95]. The data quality issues and biases in the EHR contribute to the substantial challenges and risks for developing a ML algorithm for the prediction of suicide attempts and deaths [96, 97].

Public databases are an important resource for ML research. However, using a database published for one task to train algorithms for a different task ("off-label" use) can lead to biased results [98]. As an example, the use of reconstructed and processed MRI images from public databases to generate raw MRI data to train image recognition algorithms can result in artificial improvement in algorithm performance of 25 to 48% [98]. This is due to the implicit filtering and smoothing of the reconstructed MRI image data that is used to recreate the raw image data.

Dataset Shift

When an AI algorithm is deployed in a setting where the production data differs from the training data, the algorithm often does not perform well. This is referred to as dataset shift [99•, 100, 101]. Dataset shift may be the result of a wide range of differences between the training dataset and the production population, including population demographics, treatments available, standard of care, measurement technology, practice settings, disease classification, and disease prevalence. Since healthcare practices change over time, temporal dataset shifts occur, and the size of the shifts vary with the clinical outcome being predicted [102, 103]. Dataset shifts also occur after changing from one EHR system to another [104].

For example, gender imbalance in training datasets led to decreased performance for diagnosis of thoracic diseases from X-ray images for the underrepresented gender [105]. The diagnostic performance of a ML algorithm to detect tuberculosis that was developed using a chest X-ray training dataset of one population fell when used with another population [106]. Population diversity in age, sex, and brain scanning site substantially affected the predictive accuracy of ML neuroimaging studies, including for autism spectrum disorder [107]. An algorithm to predict clinical orders by hospital admission diagnosis performed better when trained on a small amount of recent data (one month) than when trained on larger amounts of older data (12 months of 3-year-old data) due to changing practice patterns [108]. The performance of an ML algorithm to predict the risk of sepsis in ICU patients decreased over time, related to the shift from ICD-9 to ICD-10, and to hospital expansion that reshaped the population served [109]. AI algorithms to diagnose lesions in dermatology that were trained predominantly using white populations may underperform in patients with skin of color [110].

Dataset shift can also occur when a sensor in a measurement device, such as a wearable or smartphone, is different from the sensor used to create the training data [111, 112]. The sensors embedded in wearables and smartphones differ between manufacturers and between makes and models from the same manufacturer, resulting in measurement inaccuracies and inconsistencies [113–115]. Patient behaviors such as placement of the wearable may also contribute to measurement inconsistencies and dataset shifts [111, 112].

Before AI can be safely integrated into clinical practice, the many difficult issues related to data quality, EHR, dataset shift, and public databases must be addressed.

Physician Attitudes About AI

The success of AI technology in clinical settings depends on the physicians that use it.

Physician attitudes towards the use of AI in clinical medicine are generally positive, although there are concerns about ethical and legal issues, and perspectives often vary by specialty [116]. In an international survey of 791 psychiatrists, only 17% thought a computer could replace a human in providing empathetic care, while 75% thought a computer could replace a human in documentation tasks [117]. The overall acceptance of several new technologies by 515 psychiatrists in France was moderate, with 79.6% describing them as risky [118]. In a survey of 303 physicians of all specialties in Germany, the overall attitude toward AI in medicine was positive, but only 20.5% thought AI would help with the diagnosis of psychiatric disease [119]. The dominant perspective held by 720 general practitioners in the UK was that AI would have a limited impact on primary care, with the major benefits due to reducing administrative burdens [120]. Of 121 dermatologists in the USA, in a survey about AI screening tools, 49 (42%) were worried about human deskilling [121]. In a survey of 100 physicians in all specialties in the USA, although over 70% thought chatbots could assist patients with administrative tasks such as scheduling appointments and locating clinics, over 70% also thought that chatbots cannot effectively provide care for all patient needs or display emotion and could be a risk to patients due to incorrect self-diagnosis [122].

Several studies noted that most physicians lack education in AI. Although 71% of 632 radiologists, dermatologists, and ophthalmologists in Australia and New Zealand felt AI would improve their field, 80.9% had not used AI in clinical practice and 86.2% thought there was a need for improved education and guidelines to prepare for the introduction of AI [123]. In a survey of 699 physicians and medical students in South Korea, while 83.4% thought AI would be useful in medicine, only 6% said they had good familiarity with AI [124]. A survey of 210 postgraduate trainee physicians in

the UK rated the current level of AI training as insufficient [125]. In an international survey of 209 psychiatrists, only 23.9% had any formal training in technology [126]. Physicians also have varied levels of formal training and knowledge of statistics [127, 128]. These survey responses highlight the importance of quality education in AI for clinical medicine. Physicians will need to understand how to critically assess the capabilities, benefits, limitations, and risks of AI in clinical practice, and AI training must be integrated across the wide range of medical education [67, 129, 130]. Education must also emphasize that the physician remains the primary decision maker, and the ongoing importance of human intelligence and skills in patient care [68, 131].

Safety Challenges

There are many reported safety challenges that need to be understood and addressed before AI can be routinely used in a clinical setting. The safety issues with AI are especially troubling, given the disconnect between the exuberant claims and the current maturity of AI technology.

Automation Bias and Deskilling

The interaction of humans and an automated decision support tool often leads to automation bias. Automation bias occurs when a user attributes more authority to an automated tool than to other sources of advice [132, 133]. This can result in the user following incorrect advice despite contradictory evidence or prior training, and the user failing to act without explicit prompting. There are examples of automation bias across medicine including for interpretation of electrocardiograms [134, 135], e-prescribing [136], whole slide image classification in pathology [137], and diagnosis of skin cancers [138]. Although the least experienced physicians may be most susceptible to automation bias [134], a major concern is that incorrect decision support misleads clinicians of all experience levels [134, 135, 138]. For example, incorrect AI has reduced the accuracy of expert physicians in the diagnosis of skin cancers [138], and the histopathologic classification of liver cancer [137]. To reach the potential for AI products to improve decision making in clinical practice, the vulnerability of even experienced physicians to faulty AI must be understood and addressed.

A possible long-term consequence of the overreliance on technology is deskilling of the physician workforce, due to a loss of individual skills and a reduction in skill development [139–142]. This is of particular concern given the frequently promoted perspective that AI is inherently exceptional, will outperform other technologies, and will outperform physicians [143, 144]. Another risk of overreliance on technology is that even when a failure is detected, some users do not want to proceed without the AI system [145]. Implementation plans for AI in medicine should include long-term efforts to reduce deskilling of physicians and other medical personnel.

Black-box Opacity

The black-box nature of AI algorithms poses a major obstacle for routine use in clinical medicine. Beyond the many technical issues, the opacity of AI algorithms is often due to intentional secrecy by private commercial organizations [146, 147]. Modern AI techniques were originally developed for low risk decisions such as online advertising and search engines [148]. In sharp contrast, where patient safety is at risk in medicine, physicians need to understand why an AI algorithm made a prediction [149]. A lack of interpretability will undermine trust in an AI algorithm, and the explanations must be presented to physicians in a manner that is clearly understandable. There are many ongoing efforts to provide interpretability of AI algorithms, with research often focused on healthcare. Although there are various approaches to provide interpretability, each method currently has important technical challenges and limitations [149–154]. Another problem is that explanations may contribute to inappropriate trust in the capability of an AI algorithm [155]. Due to automation bias, even incorrect explanations may increase trust in AI algorithms [156]. In a study presenting patient vignettes to 220 clinicians including 195 psychiatrists, ML recommendations did not improve selection of an antidepressant drug, and incorrect ML recommendations paired with easily understood explanations decreased selection accuracy [157•]. Additionally, physicians may not understand the limitations of the explanatory methods with respect to individual treatment decisions [150]. System design for the safe use of AI in medicine must focus on improving human–computer interactions [158].

Unanticipated Safety Challenges

Complex automated systems typically fail due to the unanticipated and unintended consequences of the design, even if the nominal purpose is achieved [159]. Adding any new technology will change the workflow, often profoundly, including the creation of new failure paths [68]. In medicine, a failing AI system can result in entirely new and unexpected types of safety hazards that physicians have not seen previously [160•]. Some failure modes for AI systems may be less obvious and harder to detect than those of conventional systems [161]. The study of AI system failures should also include identification of the worst possible mistake [162]. Any AI system that automatically initiates actions must provide explicit human alerting and override abilities [160•]. Significant human oversight of AI systems

is especially important in safety–critical situations [27, 163]. When unexpected automation errors occur, a human must solve the problem [68, 164]. The more complex the automated system, the more essential the role of humans as the exception handlers, and the greater the negative impacts of deskilling. The safety of any software system, including AI, must be thoroughly evaluated in the specific environment, workflow, and context in which it is used [165].

Validation Issues

When using an AI product, the physician relies on the validation and regulatory approval process to confirm the product works as promised, and to understand the limitations and risks. There are many challenges to validating AI algorithms including data quality and dataset shift issues, brittleness and fragility of algorithms, black-box opacity, human factors, overall system context and complexity, and software errors [166–168]. The acceptable level of accuracy for the AI algorithm must be determined, given the intended use. The inappropriate choice of internal validation method can artificially inflate estimates of ML algorithm performance [169]. Notably, it is more difficult to validate AI algorithms than traditional statistical models since the results may change over time as the algorithm learns [167]. The reproducibility of ML in healthcare compares poorly to ML used in other fields, as only 23% of over 200 studies between 2017 and 2019 used multiple datasets to establish results [170]. Additionally, multiple testing approaches are required even for high performing algorithms, since unexpected and potentially harmful errors may appear when using different methodologies [162, 171].

Although the number of approved AI-based medical devices has increased in the USA and EU in recent years [172], the current state of USA regulation of medical AI demonstrates that many problems remain [173]. From a regulatory perspective, traditional medical device regulation was not designed for adaptive AI/ML technologies, and continual learning poses many challenges [174, 175]. The guidelines for regulatory approval of medical AI devices are not finalized in either the USA or EU, and new regulatory frameworks are being proposed [174, 176–178]. Post-market surveillance of approved medical AI devices is also needed [178, 179]. The validation requirements for AI algorithms that fall outside of regulatory frameworks, such as hospital developed AI, also need clarification [180].

Validation Examples from Radiology

The regulatory problems are readily apparent in recent studies of validated imaging products in the USA and Europe. Recent studies of FDA approved AI algorithms in medical devices based on publicly available summary information emphasize various validation problems. Of 130 medical devices with AI algorithms approved between 2015 and 2020, 126 devices were evaluated by the FDA using only retrospective studies, and 59 (45%) did not include the sample size [181]. Of the 130 approved devices, only 37 (28%) reported evaluation at more than one site [181]. In another study of 118 FDA approved AI algorithms across imaging modalities approved between 2008 and 2021, only 66 reported the sample size, with 45/66 (68%) having a sample size less than 500 patients [182]. Most FDA summary documents available to the public do not provide the demographics or details of the sample studied [182]. Of 100 European Conformity marked AI radiology products, of which 51 were also cleared by the FDA, only 36 had peer-reviewed evidence of efficacy [183•]. Of 237 studies obtained for the 100 products, 192 were retrospective, and only 71/237 (30%) included multicenter data [183•]. With considerable heterogeneity in deployment and pricing, most products brought to market in recent years, and with clinical impacts unclear, the authors concluded that "artificial intelligence in radiology is still in its infancy" [183•]. Validation and reporting standards must be improved to increase safety, and physician product knowledge and trust before AI is routinely used in clinical medicine.

Discussion

The promise of AI in medicine is real, but the current technical maturity of AI is low. AI is not routinely used in clinical practice. In the USA, knowledge of AI-related skills is not a standard requirement for employment in the healthcare field [184]. Between 2015 and 2018, only 1 in 1250 online job postings for skilled jobs in hospitals required some AI-related skills, lower than in other skilled industries [184]. It is important to have realistic expectations given the wide ranging problems confronting the successful adoption of AI in routine clinical medicine. The many complex technical, validation, regulatory, implementation, maintenance, and monitoring issues need to be solved carefully and rigorously. There needs to be a strong emphasis on the human–computer interface, understanding how the introduction of AI products will modify the workflow in specific clinical settings, and training for unexpected safety hazards. Physicians need education in AI fundamentals, and should be involved in the entire process of AI software development, implementation, training, and ongoing monitoring throughout the life of the system. Additionally, psychiatrists should be involved in understanding the behavioral issues related to automation bias and deskilling in medicine, as well as in the development of AI technology that predicts human behavior [185].

Many challenges for physicians to successfully augment decision making with AI are unique to medicine. Physicians must understand and trust AI sufficiently to incorporate the advice in the treatment of individual patients. The physician must interpret the AI prediction given the overall clinical context, including patient-specific characteristics, comorbid medical conditions, unique medication regimens, and socioeconomic issues. Today, the physician hears promises of accuracy of AI tools that may not be validated and sees explanations of AI tools that may not be clear. For example, the performance measures used to describe an ML algorithm may hide the uncertainty in the predictions [186]. This is especially relevant in psychiatry given the frequent use of categorical, probabilistic diagnoses in training data [186]. Another concern is that AI output may be plausible but incorrect and potentially dangerous for an individual patient [161]. Patients frequently have comorbid illnesses, yet the separate predictive algorithms being developed for each comorbid condition could provide conflicting advice [187]. The impacts of AI on the humanistic aspects of medicine, including the doctor-patient relationship, patient trust, and communications, need to be understood [188, 189]. AI technologies will become an important source of medical knowledge for physicians, but human inductive reasoning, situational awareness, and creative problem solving will remain fundamental for individual patient care, as exemplified by psychiatry [68, 163]. The successful deployment of AI in clinical medicine will require coordinated and ongoing efforts of physicians working with professionals with a wide range of skills from a number of disciplines.

Limitations

There are many limitations to this review. Technical details about the problems noted in AI software development, validation, and interpretability were not discussed. The risks of cyber attacks on AI systems, including ML vulnerability to adversarial attacks [190, 191], and the difficulty in detecting and tracing software errors were not mentioned [165]. Approaches to select appropriate tasks for AI in medicine, enhance integration of AI tools into EHR and other connected hospital systems, improve data quality, or for ongoing safety monitoring were not discussed. The wide range of collaborative skills needed for successful AI development and implementation were not included [158]. The many ethical [192, 193] and legal issues related to AI including fairness and inclusion, privacy, physician liability, algorithm failure, and vendor contract terms were not discussed [194, 195]. Patient perspectives of the use of AI in clinical medicine [196, 197], and the economic impacts of implementing AI in healthcare were not mentioned [198]. Finally, detailed recommendations to address the many noted problems are outside the scope of this review.

Conclusion

AI for clinical medicine has enormous potential but lacks technical maturity. The safe and effective implementation of AI technology to augment medical decision making poses wide-ranging challenges involving technical and human factors, regulatory issues, and safety risks. These challenges must be recognized and methodically addressed to maximize the benefits from AI technology in psychiatry in the future. It is important to set reasonable expectations. The solutions are complex and will take time to discover, develop, validate, and implement, but will lead to the safe and beneficial use of AI to augment medical decision making in psychiatry.

Declarations

Conflict of Interest EA has served on advisory boards or consulted for Alkermes, Atheneum, Janssen, Karuna, Lundbeck/Otsuka, Roche, Sunovion, and Teva and reports previous stock holdings in AstraZeneca, Johnson & Johnson, Moderna, and Pfizer. EA has received research support from Alkermes, Astellas, Biogen, Boehringer-Ingelheim, InnateVR, Janssen, National Network of Depression Centers, Neurocrine Biosciences, Novartis, Otsuka, Pear Therapeutics, and Takeda. EA serves as an advisor to CAPNOS Zero, the World Psychiatric Association and Clubhouse International, and the SMI Adviser LAI Center of Excellence (all unpaid). The other authors declare no competing interests.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Papers of particular interest, published recently, have been highlighted as:
• Of importance

1. Solow RM. "We'd better watch out" review of manufacturing matters: the myth of the post-industrial economy, by Stephen S. Cohen and John Zysman, New York Times, 1987.
2. Brynjolfsson E, Hitt LM. Beyond the productivity paradox. Commun ACM. 1998;41:49–55. https://dl.acm.org/doi/pdf/10.1145/280324.280332. Accessed 4 Sept 2022.
3. Bronsoler A, Doyle JJ Jr, Van Reenen J. The impact of healthcare IT on clinical quality, productivity and workers. Natl Bureau Econ Res. 2021. https://www.nber.org/papers/w29218. Accessed 4 Sept 2022.
4. Bui QN, Hansen S, Liu M, Tu Q. The productivity paradox in health information technology. Commun ACM. 2018;61:78–85.
5. Schweikl S, Obermaier R. Lessons from three decades of IT productivity research: towards a better understanding of IT-induced productivity effects. Management Review Quarterly. 2020;70:461–507.
6. Brynjolfsson E, Benzell S, Rock D. Understanding and addressing the modern productivity paradox. MIT Work of the Future. 2020. https://workofthefuture.mit.edu/wp-content/uploads/

2020/11/2020-Research-Brief-Brynjolfsson-Benzell-Rock.pdf. Accessed 4 Sept 2022.
7. Brynjolfsson E, Rock D, Syverson C. Artificial Intelligence and the modern productivity paradox: a clash of expectations and statistics. In: The Economics of Artificial Intelligence: An Agenda; University of Chicago Press: Chicago, IL, USA, 2019; pp. 23–57.
8. Pretz K. Stop calling everything AI, machine-learning pioneer says. IEEE Spectrum. 2021. https://spectrum.ieee.org/stop-calling-everything-ai-machinelearning-pioneer-says. Accessed 4 Sept 2022.
9. Larson EJ. The myth of artificial intelligence. Cambridge, MA: Harvard University Press; 2021.
10. Marcus G. Deep learning: a critical appraisal. 2018. https://arxiv.org/abs/1801.00631.
11.• Jordan MI. Artificial intelligence–the revolution hasn't happened yet. Harvard Data Sci Rev. 2019. https://hdsr.mitpress.mit.edu/pub/wot7mkc1/release/9. **Commentary from a professor and pioneer in AI, ML, and computer science.**
12. Smith B, Linden G. Two decades of recommender systems at Amazon.com. IEEE Internet Comput. 2017;21:12–8.
13. Barocas S, Selbst AD. Big data's disparate impact. Calif L Rev. 2016;104:671.
14. Gandy OH. Engaging rational discrimination: exploring reasons for placing regulatory constraints on decision support systems. Ethics Inf Technol. 2010;12:29–42.
15. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319:1317–8. https://doi.org/10.1001/jama.2017.18391.
16. Deo RC. Machine learning in medicine. Circulation. 2015;132:1920–30. https://doi.org/10.1161/CIRCULATIONAHA.115.001593.
17. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med. 2019;25:24–9. https://doi.org/10.1038/s41591-018-0316-z.
18. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nat Methods. 2018;15:233–4. https://doi.org/10.1038/nmeth.4642.
19. Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. NPJ Digit Med. 2020;3:126. https://doi.org/10.1038/s41746-020-00333-z.
20. Harrell F. Road map for choosing between statistical modeling and machine learning. In: Statistical Thinking blog. 2021. https://www.fharrell.com/post/stat-ml/.
21. Romano R, Gambale E. Statistics and medicine: the indispensable know-how of the researcher. Transl Med UniSa. 2013;5:28–31.
22. Monteith S, Glenn T, Geddes J, Bauer M. Big data are coming to psychiatry: a general introduction. Int J Bipolar Disord. 2015;3:21. https://doi.org/10.1186/s40345-015-0038-9.
23. Monteith S, Glenn T, Geddes J, Whybrow PC, Bauer M. Big data for bipolar disorder. Int J Bipolar Disord. 2016;4:10. https://doi.org/10.1186/s40345-016-0051-7.
24. 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. https://doi.org/10.1002/wps.20882.
25. Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. Biol Psychiatry Cogn Neurosci Neuroimaging. 2018;3:223–30. https://doi.org/10.1016/j.bpsc.2017.11.007.
26. Lin E, Lin CH, Lane HY. Precision psychiatry applications with pharmacogenomics: artificial intelligence and machine learning approaches. Int J Mol Sci. 2020;21:969. https://doi.org/10.3390/ijms21030969.
27. Cummings ML. Rethinking the maturity of artificial intelligence in safety-critical settings. AI Mag. 2021;42:6–15.
28. Mankins JC. Technology readiness levels. A White Paper, NASA, Washington, DC, 1995.
29. Mankins JC. Technology readiness assessments: a retrospective. Acta Astronaut. 2009;65:1216–23.
30. Olechowski A, Eppinger SD, Joglekar N. Technology readiness levels at 40: a study of state-of-the-art use, challenges, and opportunities. In: 2015 Portland international conference on management of engineering and technology (PICMET) 2015 Aug 2 (pp. 2084–2094). IEEE.
31. Fleuren LM, Thoral P, Shillan D, Ercole A, Elbers PWG, Right Data Right Now Collaborators. Machine learning in intensive care medicine: ready for take-off? Intensive Care Med. 2020;46:1486–8. https://doi.org/10.1007/s00134-020-06045-y.
32.• van de Sande D, van Genderen ME, Huiskens J, Gommers D, van Bommel J. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. Intensive Care Med. 2021;47:750–60. https://doi.org/10.1007/s00134-021-06446-7. **Review article showing the lack of technological maturity of AI developed for the ICU.**
33. Butler D. Translational research: crossing the valley of death. Nature. 2008;12(453):840–2. https://doi.org/10.1038/453840a.
34. Kampers LFC, Asin-Garcia E, Schaap PJ, Wagemakers A, Martins Dos Santos VAP. From innovation to application: bridging the valley of death in industrial biotechnology. Trends Biotechnol. 2021;39:1240–2. https://doi.org/10.1016/j.tibtech.2021.04.010.
35. McIntyre RA. Overcoming "the valley of death." Sci Prog. 2014;97:234–48. https://doi.org/10.3184/003685014X14079421402720.
36. Heaven D. Deep trouble for deep learning. Nature. 2019;574:163–6.
37. Karmon D, Zoran D, Goldberg Y. LaVAN: localized and visible adversarial noise. In: Proceedings of the 35th International Conference on Machine Learning. 2018 (pp. 2507–2515). PMLR.
38. Su J, Vargas DV, Sakurai K. One pixel attack for fooling deep neural networks. IEEE Trans Evol Comput. 2019;23:828–41.
39. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 2018;15:e1002683. https://doi.org/10.1371/journal.pmed.1002683.
40. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–8. https://doi.org/10.1038/nature21056.
41. Murphree DH, Puri P, Shamim H, et al. Deep learning for dermatologists: part I. Fundamental concepts. J Am Acad Dermatol. 2020:S0190–9622(20)30921-X. https://doi.org/10.1016/j.jaad.2020.05.056.
42. Badgeley MA, Zech JR, Oakden-Rayner L, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ Digit Med. 2019;2:31. https://doi.org/10.1038/s41746-019-0105-1.
43. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. Proc ACM Conf Health Inference Learn (2020). 2020;2020:151–9.
44. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. Science. 2017;356:183–6. https://doi.org/10.1126/science.aal4230.
45. Harwell D. The Accent Gap. 2018. The Washington Post. https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/. Accessed 4 Sept 2022.
46. Kitashov F, Svitanko E, Dutta D. Foreign English accent adjustment by learning phonetic patterns. arXiv preprint 2018. arXiv:1807.03625.
47. Koenecke A, Nam A, Lake E, et al. Racial disparities in automated speech recognition. Proc Natl Acad Sci U S A. 2020;117:7684–9. https://doi.org/10.1073/pnas.1915768117.

48. Hitczenko K, Cowan H, Mittal V, Goldrick M. Automated coherence measures fail to index thought disorder in individuals at risk for psychosis. In: Proceedings of the seventh workshop on computational linguistics and clinical psychology: improving access 2021 Jun (pp. 129–150). Association for Computational Linguistics, 2021.
49. Hitczenko K, Cowan HR, Goldrick M, Mittal VA. Racial and ethnic biases in computational approaches to psychopathology. Schizophr Bull. 2022;48:285–8. https://doi.org/10.1093/schbul/sbab131.
50. Vogel AP, Morgan AT. Factors affecting the quality of sound recording for speech and voice analysis. Int J Speech Lang Pathol. 2009;11:431–7. https://doi.org/10.3109/17549500902822189.
51. Zheng B, Hu J, Zhang G, Wu Y, Deng J. Analysis of noise reduction techniques in speech recognition. In 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) 2020 (Vol. 1, pp. 928–933). IEEE.
52. Blackley SV, Huynh J, Wang L, Korach Z, Zhou L. Speech recognition for clinical documentation from 1990 to 2018: a systematic review. J Am Med Inform Assoc. 2019;26:324–38. https://doi.org/10.1093/jamia/ocy179.
53. Goss FR, Zhou L, Weiner SG. Incidence of speech recognition errors in the emergency department. Int J Med Inform. 2016;93:70–3. https://doi.org/10.1016/j.ijmedinf.2016.05.005.
54. Quiroz JC, Laranjo L, Kocaballi AB, Berkovsky S, Rezazadegan D, Coiera E. Challenges of developing a digital scribe to reduce clinical documentation burden. NPJ Digit Med. 2019;2:114. https://doi.org/10.1038/s41746-019-0190-1.
55. Kodish-Wachs J, Agassi E, Kenny P III, Overhage JM. A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech. In AMIA Ann Symp Proc. 2018 (Vol. 2018, p. 683). American Medical Informatics Association.
56. Miner AS, Haque A, Fries JA, et al. Assessing the accuracy of automatic speech recognition for psychotherapy. NPJ Digit Med. 2020;3:82. https://doi.org/10.1038/s41746-020-0285-8.
57. ACR (American College of Radiology) Data Science Institute AI Central. FDA-cleared AI algorithms. 2022. https://aicentral.acrdsi.org/. Accessed 4 Sept 2022.
58. Allen B, Agarwal S, Coombs L, Wald C, Dreyer K. 2020 ACR data science institute artificial intelligence survey. J Am Coll Radiol. 2021;18:1153–9. https://doi.org/10.1016/j.jacr.2021.04.002.
59. Vasey B, Ursprung S, Beddoe B, et al. Association of clinician diagnostic performance with machine learning-based decision support systems: a systematic review. JAMA Netw Open. 2021;4:e211276. https://doi.org/10.1001/jamanetworkopen.2021.1276.
60. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368:m689. https://doi.org/10.1136/bmj.m689.
61. Antun V, Renna F, Poon C, Adcock B, Hansen AC. On instabilities of deep learning in image reconstruction and the potential costs of AI. Proc Natl Acad Sci U S A. 2020;117:30088–95. https://doi.org/10.1073/pnas.1907377117.
62. ECRI. AI-based reconstruction can distort images, threatening diagnostic outcomes. Hazard #7—2022 top 10 health technology hazards. Device Evaluation 2022.
63. McCollough CH, Bartley AC, Carter RE, et al. Low-dose CT for the detection and classification of metastatic liver lesions: results of the 2016 Low Dose CT Grand Challenge. Med Phys. 2017;44:e339–52. https://doi.org/10.1002/mp.12345.
64. Allen B, Dreyer K, Stibolt R Jr, et al. Evaluation and real-world performance monitoring of artificial intelligence models in clinical practice: try it, buy it, check it. J Am Coll Radiol. 2021;18:1489–96. https://doi.org/10.1016/j.jacr.2021.08.022.
65. Gupta RV, Kalra MK, Ebrahimian S, et al. Complex relationship between artificial intelligence and CT radiation dose. Acad Radiol. 2021:S1076–6332(21)00489-X. https://doi.org/10.1016/j.acra.2021.10.024.
66. Roberts M, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell. 2021;3:199–217. https://doi.org/10.1038/s42256-021-00307-0.
67. Matheny M, Israni ST, Ahmed M, Whicher D. Artificial intelligence in health care: the hope, the hype, the promise, the peril. Washington, DC: National Academy of Medicine; 2019.
68. Bauer M, Monteith S, Geddes J, et al. Automation to optimise physician treatment of individual patients: examples in psychiatry. Lancet Psychiatry. 2019;6:338–49. https://doi.org/10.1016/S2215-0366(19)30041-0.
69. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013;51(8 Suppl 3):S30-37. https://doi.org/10.1097/MLR.0b013e31829b1dbd.
70. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ. 2018;30(361):k1479. https://doi.org/10.1136/bmj.k1479.
71. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med. 2018;178:1544–7. https://doi.org/10.1001/jamainternmed.2018.3763.
72. Kaplan RM, Chambers DA, Glasgow RE. Big data and large sample size: a cautionary note on the potential for bias. Clin Transl Sci. 2014;7:342–6. https://doi.org/10.1111/cts.12178.
73. Paxton C, Niculescu-Mizil A, Saria S. Developing predictive models using electronic medical records: challenges and pitfalls. AMIA Annu Symp Proc. 2013;16(2013):1109–15. PMID: 24551396.
74. Phelan M, Bhavsar NA, Goldstein BA. Illustrating informed presence bias in electronic health records data: how patient interactions with a health system can impact inference. EGEMS (Wash DC). 2017;5:22. https://doi.org/10.5334/egems.243.
75. Polubriaginof FCG, Ryan P, Salmasian H, et al. Challenges with quality of race and ethnicity data in observational databases. J Am Med Inform Assoc. 2019;26:730–6. https://doi.org/10.1093/jamia/ocz113.
76. Price WN II. Medical AI and contextual bias. Harvard Journal of Law & Technology. 2019;33:65–116.
77. Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. BMC Med Inform Decis Mak. 2014;14:51. https://doi.org/10.1186/1472-6947-14-51.
78. Walsh CG, Chaudhry B, Dua P, et al. Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence. JAMIA Open. 2020;3:9–15. https://doi.org/10.1093/jamiaopen/ooz054.
79. Madden JM, Lakoma MD, Rusinak D, Lu CY, Soumerai SB. Missing clinical and behavioral health data in a large electronic health record (EHR) system. J Am Med Inform Assoc. 2016;23:1143–9. https://doi.org/10.1093/jamia/ocw021.
80. Harper KL, Ellickson-Larew S, Bovin MJ, Keane TM, Marx BP. Discrepancies between electronic records and clinical interview diagnosis of PTSD: differences in mental health care utilization. Psychol Serv. 2021. https://doi.org/10.1037/ser0000560.10.1037/ser0000560.

81. Morgan MA, Kelber MS, O'Gallagher K, Liu X, Evatt DP, Belsher BE. Discrepancies in diagnostic records of military service members with self-reported PTSD: healthcare use and longitudinal symptom outcomes. Gen Hosp Psychiatry. 2019;58:33–8. https://doi.org/10.1016/j.genhosppsych.2019.02.006.
82. Wilk JE, Herrell RK, Carr AL, West JC, Wise J, Hoge CW. Diagnosis of PTSD by Army behavioral health clinicians: are diagnoses recorded in electronic health records? Psychiatr Serv. 2016;67:878–82. https://doi.org/10.1176/appi.ps.201500292.
83. Anderson HD, Pace WD, Brandt E, et al. Monitoring suicidal patients in primary care using electronic health records. J Am Board Fam Med. 2015;28:65–71. https://doi.org/10.3122/jabfm.2015.01.140181.
84. Dossa A, Welch LC. GPs' approaches to documenting stigmatising information: a qualitative study. Br J Gen Pract. 2015;65:e372–8. https://doi.org/10.3399/bjgp15X685273.
85. Hollister B, Bonham VL. Should electronic health record-derived social and behavioral data be used in precision medicine research? AMA J Ethics. 2018;20:E873-880. https://doi.org/10.1001/amajethics.2018.873.
86. Maust DT, Gerlach LB, Gibson A, Kales HC, Blow FC, Olfson M. Trends in central nervous system-active polypharmacy among older adults seen in outpatient care in the United States. JAMA Intern Med. 2017;177:583–5. https://doi.org/10.1001/jamainternmed.2016.9225.
87. Rhee TG, Rosenheck RA. Initiation of new psychotropic prescriptions without a psychiatric diagnosis among US adults: rates, correlates, and national trends from 2006 to 2015. Health Serv Res. 2019;54:139–48. https://doi.org/10.1111/1475-6773.13072.
88. Simon GE, Stewart C, Beck A, et al. National prevalence of receipt of antidepressant prescriptions by persons without a psychiatric diagnosis. Psychiatr Serv. 2014;65:944–6. https://doi.org/10.1176/appi.ps.201300371.
89. Wiechers IR, Leslie DL, Rosenheck RA. Prescribing of psychotropic medications to patients without a psychiatric diagnosis. Psychiatr Serv. 2013;64:1243–8. https://doi.org/10.1176/appi.ps.201200557.
90. Stewart CC, Lu CY, Yoon TK, et al. Impact of ICD-10-CM transition on mental health diagnoses recording. EGEMS (Wash DC). 2019;7:14. https://doi.org/10.5334/egems.281.
91. Heslin KC, Owens PL, Karaca Z, Barrett ML, Moore BJ, Elixhauser A. Trends in opioid-related inpatient stays shifted after the US transitioned to ICD-10-CM diagnosis coding in 2015. Med Care. 2017;55:918–23. https://doi.org/10.1097/MLR.0000000000000805.
92. Heslin KC, Barrett ML. Shifts in alcohol-related diagnoses after the introduction of International Classification Of Diseases, Tenth Revision, clinical modification coding in U.S. hospitals: implications for epidemiologic research. Alcohol Clin Exp Res. 2018;42:2205–13. https://doi.org/10.1111/acer.13866.
93. Shields MC, Ritter G, Busch AB. Electronic health information exchange at discharge from inpatient psychiatric care in acute care hospitals. Health Aff (Millwood). 2020;39:958–67. https://doi.org/10.1377/hlthaff.2019.00985.
94. Zurynski Y, Ellis LA, Tong HL, et al. Implementation of electronic medical records in mental health settings: scoping review. JMIR Ment Health. 2021;8:e30564. https://doi.org/10.2196/30564.
95. Ranallo PA, Kilbourne AM, Whatley AS, Pincus HA. Behavioral health information technology: from chaos to clarity. Health Aff (Millwood). 2016;35:1106–13. https://doi.org/10.1377/hlthaff.2016.0013.
96. Belsher BE, Smolenski DJ, Pruitt LD, et al. Prediction models for suicide attempts and deaths: a systematic review and simulation. JAMA Psychiat. 2019;76:642–51. https://doi.org/10.1001/jamapsychiatry.2019.0174.
97. Kirtley OJ, van Mens K, Hoogendoorn M, Kapur N, de Beurs D. Translating promise into practice: a review of machine learning in suicide research and prevention. Lancet Psychiatry. 2022;9:243–52. https://doi.org/10.1016/S2215-0366(21)00254-6.
98. Shimron E, Tamir JI, Wang K, Lustig M. Implicit data crimes: machine learning bias arising from misuse of public data. Proc Natl Acad Sci U S A. 2022;119:e2117203119. https://doi.org/10.1073/pnas.2117203119.
99.• Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med. 2021;385:283–6. https://doi.org/10.1056/NEJMc2104626. **Introduction to clinical dataset shift issues.**
100. Ovadia Y, Fertig E, Ren J, et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Adv Neural Inf Proces Syst. 2019;32.
101. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics. 2020;21:345–52. https://doi.org/10.1093/biostatistics/kxz041.
102. Guo LL, Pfohl SR, Fries J, et al. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. Sci Rep. 2022;12:2726. https://doi.org/10.1038/s41598-022-06484-1.
103. Nestor B, McDermott M, Chauhan G, et al. Rethinking clinical prediction: why machine learning must consider year of care and feature aggregation. arXiv preprint 2018. arXiv:1811.12583.
104. Gong JJ, Naumann T, Szolovits P, Guttag JV. Predicting clinical outcomes across changing electronic health record systems. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 2017 (pp. 1497–1505).
105. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci U S A. 2020;117:12592–4. https://doi.org/10.1073/pnas.1919012117.
106. Sathitratanacheewin S, Sunanta P, Pongpirul K. Deep learning for automated classification of tuberculosis-related chest X-Ray: dataset distribution shift limits diagnostic performance generalizability. Heliyon. 2020;6:e04614. https://doi.org/10.1016/j.heliyon.2020.e0461.
107. Benkarim O, Paquola C, Park BY, et al. Population heterogeneity in clinical cohorts affects the predictive accuracy of brain imaging. PLoS Biol. 2022;20:e3001627. https://doi.org/10.1371/journal.pbio.3001627.
108. Chen JH, Alagappan M, Goldstein MK, Asch SM, Altman RB. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. Int J Med Inform. 2017;102:71–9. https://doi.org/10.1016/j.ijmedinf.2017.03.006.
109. Ross C. AI gone astray: how subtle shifts in patient data send popular algorithms reeling, undermining patient safety. https://www.statnews.com/2022/02/28/sepsis-hospital-algorithms-data-shift/. Accessed 4 Sept 2022.
110. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. JAMA Dermatol. 2018;154:1247–8. https://doi.org/10.1001/jamadermatol.2018.2348.
111. Park C, Awadalla A, Kohno T, Patel S. Reliable and trustworthy machine learning for health using dataset shift detection. Adv Neural Inf Process Syst. 2021;6:34.
112. Simons A, Doyle T, Musson D, Reilly J. Impact of physiological sensor variance on machine learning algorithms. In:2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC) 2020 (pp. 241–247). IEEE.

113. Bauer M, Glenn T, Geddes J, et al. Smartphones in mental health: a critical review of background issues, current status and future concerns. Int J Bipolar Disord. 2020;8:2. https://doi.org/10.1186/s40345-019-0164-x.
114. Cosoli G, Spinsante S, Scalise L. Wrist-worn and chest-strap wearable devices: systematic review on accuracy and metrological characteristics. Measurement. 2020;159:107789.
115. Kos A, Tomažič S, Umek A. Evaluation of smartphone inertial sensor performance for cross-platform mobile applications. Sensors. 2016;16:477.
116. Scott IA, Carter SM, Coiera E. Exploring stakeholder attitudes towards AI in clinical practice. BMJ Health Care Inform. 2021;28:e100450. https://doi.org/10.1136/bmjhci-2021-100450.
117. Doraiswamy PM, Blease C, Bodner K. Artificial intelligence and the future of psychiatry: insights from a global physician survey. Artif Intell Med. 2020;102:101753. https://doi.org/10.1016/j.artmed.2019.101753.
118. Bourla A, Ferreri F, Ogorzelec L, Peretti CS, Guinchard C, Mouchabac S. Psychiatrists' attitudes toward disruptive new technologies: mixed-methods study. JMIR Ment Health. 2018;5:e10240. https://doi.org/10.2196/10240.
119. Maassen O, Fritsch S, Palm J, et al. Future medical artificial intelligence application requirements and expectations of physicians in German University Hospitals: web-based survey. J Med Internet Res. 2021;23:e26646. https://doi.org/10.2196/26646.
120. Blease C, Kaptchuk TJ, Bernstein MH, Mandl KD, Halamka JD, DesRoches CM. Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. J Med Internet Res. 2019;21:e12802. https://doi.org/10.2196/12802.
121. Nelson CA, Pachauri S, Balk R, et al. Dermatologists' perspectives on artificial intelligence and augmented intelligence – a cross-sectional survey. JAMA Dermatol. 2021;157:871–4. https://doi.org/10.1001/jamadermatol.2021.1685.
122. Palanica A, Flaschner P, Thommandram A, Li M, Fossat Y. Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. J Med Internet Res. 2019;21:e12887. https://doi.org/10.2196/12887.
123. Scheetz J, Rothschild P, McGuinness M, et al. A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. Sci Rep. 2021;11:5193. https://doi.org/10.1038/s41598-021-84698-5.
124. Oh S, Kim JH, Choi SW, Lee HJ, Hong J, Kwon SH. Physician confidence in artificial intelligence: an online mobile survey. J Med Internet Res. 2019;21:e12422. https://doi.org/10.2196/12422\.
125. Banerjee M, Chiew D, Patel KT, et al. The impact of artificial intelligence on clinical education: perceptions of postgraduate trainee doctors in London (UK) and recommendations for trainers. BMC Med Educ. 2021;21:429. https://doi.org/10.1186/s12909-021-02870-x.
126. Bauer R, Glenn T, Monteith S, Whybrow PC, Bauer M. Survey of psychiatrist use of digital technology in clinical practice. Int J Bipolar Disord. 2020;8:29. https://doi.org/10.1186/s40345-020-00194-1.
127. Kahwati L, Carmody D, Berkman N, Sullivan HW, Aikin KJ, DeFrank J. Prescribers' knowledge and skills for interpreting research results: a systematic review. J Contin Educ Health Prof. 2017;37:129–36. https://doi.org/10.1097/CEH.0000000000000150.
128. Swift L, Miles S, Price GM, Shepstone L, Leinster SJ. Do doctors need statistics? Doctors' use of and attitudes to probability and statistics. Stat Med. 2009;28:1969–81. https://doi.org/10.1002/sim.3608.
129. Grunhut J, Wyatt AT, Marques O. Educating future physicians in artificial intelligence (AI): an integrative review and proposed changes. J Med Educ Curric Dev. 2021;8:23821205211036836. https://doi.org/10.1177/23821205211036836.
130. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? NPJ Digit Med. 2020;3:86. https://doi.org/10.1038/s41746-020-0294-7.
131. Alrassi J, Katsufrakis PJ, Chandran L. Technology can augment, but not replace, critical human skills needed for patient care. Acad Med. 2021;96:37–43. https://doi.org/10.1097/ACM.0000000000003733.
132. Geis JR, Brady AP, Wu CC, et al. Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. J Am Coll Radiol. 2019;16:1516–21. https://doi.org/10.1016/j.jacr.2019.07.028.
133. Parasuraman R, Manzey DH. Complacency and bias in human use of automation: an attentional integration. Hum Factors. 2010;52:381–410. https://doi.org/10.1177/0018720810376055.
134. Bond RR, Novotny T, Andrsova I, Koc L, Sisakova M, Finlay D, Guldenring D, McLaughlin J, Peace A, McGilligan V, Leslie SJ, Wang H, Malik M. Automation bias in medicine: the influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms. J Electrocardiol. 2018;51(6S):S6–11.
135. Tsai TL, Fridsma DB, Gatti G. Computer decision support as a source of interpretation error: the case of electrocardiograms. J Am Med Inform Assoc. 2003;10:478–83. https://doi.org/10.1197/jamia.M1279.
136. Lyell D, Magrabi F, Raban MZ, et al. Automation bias in electronic prescribing. BMC Med Inform Decis Mak. 2017;17:28. https://doi.org/10.1186/s12911-017-0425-5.
137. Kiani A, Uyumazturk B, Rajpurkar P, et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. NPJ digital medicine. 2020;3:23. https://doi.org/10.1038/s41746-020-0232-8.
138. Tschandl P, Rinner C, Apalla Z, et al. Human-computer collaboration for skin cancer recognition. Nat Med. 2020;26:1229–34. https://doi.org/10.1038/s41591-020-0942-0.
139. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA. 2017;318:517–8. https://doi.org/10.1001/jama.2017.7797.
140. Hoff T. Deskilling and adaptation among primary care physicians using two work innovations. Health Care Manage Rev. 2011;36:338–48. https://doi.org/10.1097/HMR.0b013e31821826a1.
141. Lu J. Will medical technology deskill doctors? Int Educ Stud. 2016;9:130–4.
142. Povyakalo AA, Alberdi E, Strigini L, Ayton P. How to discriminate between computer-aided and computer-hindered decisions: a case study in mammography. Med Decis Making. 2013;33:98–107. https://doi.org/10.1177/0272989X12465490.
143. Bélisle-Pipon JC, Couture V, Roy MC, Ganache I, Goetghebeur M, Cohen IG. What makes artificial intelligence exceptional in health technology assessment? Front Artif Intell. 2021;4:736697. https://doi.org/10.3389/frai.2021.736697.
144. Dzobo K, Adotey S, Thomford NE, Dzobo W. Integrating artificial and human intelligence: a partnership for responsible innovation in biomedical engineering and medicine. OMICS. 2020;24:247–63. https://doi.org/10.1089/omi.2019.0038.
145. Parnas DL. The real risks of artificial intelligence. Commun ACM. 2017;60:27–31.
146. Burrell J. How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc. 2016;3:2053951715622512. https://doi.org/10.1177/2053951715622512.
147. Pasquale F. The black box society. The secret algorithms that control money and information. Cambridge, MA: Harvard University Press; 2015.
148. Rudin C, Radin J. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. Harv Data Sci Rev. 2019. https://doi.org/10.1162/99608f92.5a8a3a3d.

149. Petch J, Di S, Nelson W. Opening the black box: the promise and limitations of explainable machine learning in cardiology. Can J Cardiol. 2022;38:204–13. https://doi.org/10.1016/j.cjca.2021.09.004.
150. 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. https://doi.org/10.1016/S2589-7500(21)00208-9.
151. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain?. arXiv preprint. 2017. https://arxiv.org/abs/1712.09923. Accessed 4 Sept 2022.
152. Molnar C, Casalicchio G, Bischl B. Interpretable machine learning—a brief history, state-of-the-art and challenges. arXiv 2020. https://arxiv.org/abs/2010.09337. Accessed 4 Sept 2022.
153. Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning-based prediction models in healthcare. Wiley Interdiscip Rev Data Min Knowl Discov. 2020;10:e1379.
154. Watson M, Hasan BA, Al Moubayed N. Agree to disagree: when deep learning models with identical architectures produce distinct explanations. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 2022 (pp. 875–884).
155. Lockey S, Gillespie N, Holm D, Someh IA. A review of trust in artificial intelligence: challenges, vulnerabilities and future directions. Proceedings of the Annual Hawaii International Conference on System Sciences. 2021, 5463–5472. https://doi.org/10.24251/hicss.2021.664. Accessed 4 Sept 2022.
156. Kroll JA. The fallacy of inscrutability. Philos Trans A Math Phys Eng Sci. 2018;376:20180084. https://doi.org/10.1098/rsta.2018.0084.
157.• Jacobs M, Pradier MF, McCoy TH Jr, Perlis RH, Doshi-Velez F, Gajos KZ. How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. Transl Psychiatry. 2021;11:108. https://doi.org/10.1038/s41398-021-01224-x. **Experiment finding unexpected effects of ML recommendations on physician decision making.**
158. Johnson M, Vera A. No AI is an island: the case for teaming intelligence. AI Mag. 2019;40:16–28.
159. Griffin M. System engineering and the"two cultures" of engineering. NASA, The Boeing Lecture, 2007. https://www.nasa.gov/pdf/173108main_mg_purdue_20070328.pdf. Accessed 4 Sept 2022.
160.• Mongan J, Kohli M. Artificial intelligence and human life: five lessons for radiology from the 737 MAX disasters. Radiol Artif Intell. 2020;2:e190111. https://doi.org/10.1148/ryai.2020190111. **Commentary on the broad impacts of AI system failures in safety critical situations.**
161. Whitby B. Automating medicine the ethical way. In: Machine medical ethics, 2015 (pp. 223–232). Springer, Cham. van Rysewyk SP and Pontier Meds.
162. Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. Lancet Digit Health. 2022;4:e384–97. https://doi.org/10.1016/S2589-7500(22)00003-6.
163. Cummings MM. Man versus machine or man+ machine? IEEE Intell Syst. 2014;29:62–9.
164. Strauch B. Ironies of automation: still unresolved after all these years. IEEE Trans Hum-Mach Syst. 2017;48:419–33.
165. Leveson NG. The Therac-25: 30 years later. Computer. 2017;50:8–11.
166. Breck E, Polyzotis N, Roy S, Whang S, Zinkevich M. Data validation for machine learning. In: Proceedings of the 2nd SysML Conference, 2019. https://proceedings.mlsys.org/book/2019/file/5878a7ab84fb43402106c575658472fa-Paper.pdf. Accessed 4 Sept 2022.
167. Hand DJ, Khan S. Validating and verifying AI systems. Patterns (N Y). 2020;1:100037. https://doi.org/10.1016/j.patter.2020.100037.
168. Validate AI Conference White Paper. 2019 Validate AI conference, Nov. 5, 2019, London, UK. https://validateai.org/white-papers. Accessed 4 Sept 2022.
169. Jacobucci R, Littlefield AK, Millner AJ, Kleiman EM, Steinley D. Evidence of inflated prediction performance: a commentary on machine learning and suicide research. Clin Psychol Sci. 2021;9:129–34.
170. McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: still a ways to go. Sci Transl Med. 2021;13:eabb1655. https://doi.org/10.1126/scitranslmed.abb1655.
171. Oakden-Rayner L, Gale W, Bonham TA, et al. Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. Lancet Digit Health. 2022;4:e351–8. https://doi.org/10.1016/S2589-7500(22)00004-8.
172. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. Lancet Digit Health. 2021;3:e195–203. https://doi.org/10.1016/S2589-7500(20)30292-2.
173. Shah S, El-Sayed E. Medical algorithms need better regulation. Sci Am. 2021. https://www.scientificamerican.com/article/the-fda-should-better-regulate-medical-algorithms/. Accessed 4 Sept 2022.
174. FDA. Artificial intelligence and machine learning in software as a medical device. 2021. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device. Accessed 4 Sept 2022.
175. Vokinger KN, Feuerriegel S, Kesselheim AS. Continual learning in medical devices: FDA's action plan and beyond. Lancet Digit Health. 2021;3:e337–8. https://doi.org/10.1016/S2589-7500(21)00076-5.
176. EU Publications Office. Procedure 2021/0106/COD. COM (2021) 206: Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. 2021. https://eur-lex.europa.eu/procedure/EN/2021_106?uri=PROCEDURE:2021_106.
177. Larson DB, Harvey H, Rubin DL, Irani N, Tse JR, Langlotz CP. Regulatory frameworks for development and evaluation of artificial intelligence-based diagnostic imaging algorithms: summary and recommendations. J Am Coll Radiol. 2021;18:413–24. https://doi.org/10.1016/j.jacr.2020.09.060.
178. Niemiec E. Will the EU medical device regulation help to improve the safety and performance of medical AI devices? Digit Health. 2022;8:20552076221089080. https://doi.org/10.1177/20552076221089079.
179. Dreyer KJ, Allen B, Wald C. Real-world surveillance of FDA-cleared artificial intelligence models: rationale and logistics. J Am Coll Radiol. 2022;19:274–7. https://doi.org/10.1016/j.jacr.2021.06.025.
180. Weissman GE. FDA regulation of predictive clinical decision-support tools: what does it mean for hospitals? J Hosp Med. 2021;16:244–6. https://doi.org/10.12788/jhm.3450.
181. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med. 2021;27:582–4. https://doi.org/10.1038/s41591-021-01312-x.

182. Ebrahimian S, Kalra MK, Agarwal S, et al. FDA-regulated AI algorithms: trends, strengths, and gaps of validation studies. Acad Radiol. 2022;29:559–66. https://doi.org/10.1016/j.acra.2021.09.002.
183.• van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol. 2021;31:3797–804. https://doi.org/10.1007/s00330-021-07892-z. **Discusses the frequent lack of evidence of efficacy for commercial AI software in radiology.**
184. Goldfarb A, Teodoridis F. Why is AI adoption in health care lagging? Brookings Inst. 2022. https://www.brookings.edu/research/why-is-ai-adoption-in-health-care-lagging/. Accessed 4 Sept 2022.
185. Monteith S, Glenn T, Geddes J, Whybrow PC, Bauer M. Commercial use of emotion artificial intelligence (AI): implications for psychiatry. Curr Psychiatry Rep. 2022;24:203–11. https://doi.org/10.1007/s11920-022-01330-7.
186. Joyce DW, Geddes J. When deploying predictive algorithms, are summary performance measures sufficient? JAMA Psychiat. 2020;1(77):447–8. https://doi.org/10.1001/jamapsychiatry.2019.4484.
187. Stetson PD, Cantor MN, Gonen M. When predictive models collide. JCO Clin Cancer Inform. 2020;4:547–50. https://doi.org/10.1200/CCI.20.00024.
188. Hatherley JJ. Limits of trust in medical AI. J Med Ethics. 2020;46:478–81. https://doi.org/10.1136/medethics-2019-105935.
189. Johnston SC. Anticipating and training the physician of the future: the importance of caring in an age of artificial intelligence. Acad Med. 2018;93:1105–6. https://doi.org/10.1097/ACM.0000000000002175.
190. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science. 2019;363:1287–9. https://doi.org/10.1126/science.aaw4399.
191. O'Brien JT, Nelson C. Assessing the risks posed by the convergence of artificial intelligence and biotechnology. Health Secur. 2020;18:219–27. https://doi.org/10.1089/hs.2019.0122.
192. Floridi L, Cowls J, King TC, Taddeo M. How to design AI for social good: seven essential factors. Sci Eng Ethics. 2020;26:1771–96. https://doi.org/10.1007/s11948-020-00213-5.
193. Murphy K, Di Ruggiero E, Upshur R, et al. Artificial intelligence for good health: a scoping review of the ethics literature. BMC Med Ethics. 2021;22:14. https://doi.org/10.1186/s12910-021-00577-8.
194. Floridi L. The European legislation on AI: a brief analysis of its philosophical approach. Philos Technol. 2021;34:215–22. https://doi.org/10.1007/s13347-021-00460-9.
195. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. In: Artificial intelligence in healthcare. 2020 (pp. 295–336). Academic Press. https://doi.org/10.1016/B978-0-12-818438-7.00012-5.
196. Kovarik CL. Patient perspectives on the use of artificial intelligence. JAMA Dermatol. 2020;156:493–4. https://doi.org/10.1001/jamadermatol.2019.5013.
197. Longoni C, Bonezzi A, Morewedge CK. Resistance to medical artificial intelligence. J Consum Res. 2019;46:629–50.
198. Wolff J, Pauling J, Keck A, Baumbach J. The economic impact of artificial intelligence in health care: systematic review. J Med Internet Res. 2020;22:e16866. https://doi.org/10.2196/16866.