Journal of the American Medical Informatics Association, 30(2), 2023, 361–366
https://doi.org/10.1093/jamia/ocac218
Advance Access Publication Date: 15 November 2022
Review

Review

Diachronic and synchronic variation in the performance of adaptive machine learning systems: the ethical challenges

Joshua Hatherley * and Robert Sparrow

Philosophy Department, School of Philosophical, Historical and International Studies, Monash University, Clayton, Victoria 3800, Australia

*Corresponding Author: Joshua Hatherley, MBioethics, Philosophy Department, School of Philosophical, Historical and International Studies, Monash University, Level 6, 20 Chancellor's Walk (Menzies Building), Wellington Road, Clayton, VIC 3800, Australia; joshua.hatherley@monash.edu

Received 26 July 2022; Revised 19 September 2022; Editorial Decision 28 October 2022; Accepted 1 November 2022

ABSTRACT

Objectives: Machine learning (ML) has the potential to facilitate "continual learning" in medicine, in which an ML system continues to evolve in response to exposure to new data over time, even after being deployed in a clinical setting. In this article, we provide a tutorial on the range of ethical issues raised by the use of such "adaptive" ML systems in medicine that have, thus far, been neglected in the literature.
Target audience: The target audiences for this tutorial are the developers of ML AI systems, healthcare regulators, the broader medical informatics community, and practicing clinicians.
Scope: Discussions of adaptive ML systems to date have overlooked the distinction between 2 sorts of variance that such systems may exhibit—diachronic evolution (change over time) and synchronic variation (difference between cotemporaneous instantiations of the algorithm at different sites)—and underestimated the significance of the latter. We highlight the challenges that diachronic evolution and synchronic variation present for the quality of patient care, informed consent, and equity, and discuss the complex ethical trade-offs involved in the design of such systems.

Key words: artificial intelligence, bioethics, update problem, medicine, federated learning

INTRODUCTION

Machine learning (ML) has the potential to facilitate "continual learning" in medicine, in which an ML system continues to adapt and evolve in response to exposure to new data over time, even after being deployed in a clinical setting. Leveraging this "adaptive" potential of medical ML could generate significant benefits for patient health and well-being. Recent engagements with the ethical issues generated by the use of adaptive ML systems in medicine have typically been limited to discussions of "the update problem": how should systems that continue to change and evolve postregulatory approval be regulated? In this article, we draw attention to an important set of ethical issues raised by the use of adaptive ML systems in medicine that have, thus far, been neglected and are highly deserving of further attention.

Discussions of adaptive ML systems to date have overlooked the distinction between 2 sorts of variance that such systems may exhibit—diachronic evolution (change over time) and synchronic variation (difference between cotemporaneous instantiations of the algorithmic system at different sites)—and underestimated the significance of the latter. Both diachronic evolution and synchronic variation will complicate the hermeneutic task of clinicians in interpreting the outputs of AI systems, and will therefore pose significant challenges to the process of securing informed consent to treatment. Equity issues may occur where synchronic variation is permitted, as the quality of care may vary significantly across patients or between hospitals. However, the decision as to whether to allow or eliminate synchronic variation involves complex trade-offs between accuracy and generalizability, as well as a number of other values, including justice and nonmaleficence. In some contexts, preventing synchronic variation from emerging may only be possible at the expense of the wellbeing, and the quality of care available to, particular patients or classes of patients. Designers and regulators of adaptive ML systems will need to confront these issues if the potential benefits of adaptive ML in medical care are to be realized.

ADAPTIVE MACHINE LEARNING IN MEDICINE

ML is a form of AI that involves "programming computers to optimize a performance criterion using example data or past experience".[1] The application of ML in medicine could significantly improve the delivery of medical care, and expand the availability of medical knowledge and expertise, among other benefits.[2–5] ML systems can be either "locked" or "adaptive". *Locked* ML systems have parameters fixed prior to clinical deployment, and do not continue to learn from new data over time. While, to date, regulatory approvals of medical AI systems have been limited to locked systems the U.S. Food and Drug Administration (FDA) is considering regulatory approval for *adaptive* ML systems, which evolve as they are exposed to new data ("continuous learning"), even after the system has been deployed in a clinical setting.[6,7] We will refer to these sorts of ML devices as (medical) adaptive machine learning system(s) (MAMLS).

The use of MAMLS could have a number of benefits for patients. In some applications, MAMLS can continuously "tune" their algorithms to individual patients' physiology, along with any changes that occur in a patient's physiology over their lifetime, thereby contributing to the realization of "personalised medicine". The use of ML to deliver personalized medicine is already being explored via the combination of ML with a variety of other new and emerging technologies.[8] For example, ML-enabled wearables and implantables have been developed to enable personalized identification of ventricular arrythmias and hypoglycemic events for diabetic patients, and also to predict the onset of seizures in patients with drug-resistant epilepsy.[9–12]

Additionally, MAMLS could be trained on data collected from particular cohorts of patients to tune their performance to the features of the cohorts of each particular clinical site or institution.[13] For example, MAMLS could be used to predict risk of hospital readmission for outpatients, or to identify patients at a high-risk of heart attack within particular communities.[14] Some researchers are already seeking to enable such site-specific training of medical ML systems by making the source codes of their algorithms freely available online.[15]

THE UPDATE PROBLEM

While there has been some engagement with the ethical issues raised by MAMLS, it has mostly been confined to discussions of "the update problem". Existing regulatory approaches in healthcare and medicine were designed to address products that do not evolve over time, such as pharmaceuticals. Consequently, the capacity for ongoing evolution in MAMLS presents a serious challenge for regulators. As Babic and coauthors have written: "After evaluating a medical AI/ML technology and deeming it safe and effective, should the regulator limit its authorization to market only the version of the algorithm that was submitted, or permit marketing of an algorithm that can learn and adapt to new conditions?".[16] If they approve MAMLS, regulators may be exposing patients to risks that have developed in the system postdeployment. However, restricting regulatory approvals to locked systems places a strong limit on the potential benefits that ML could generate for patient health outcomes.

In their recent *Proposed Regulatory Framework for Modifications to ML-Based Software as a Medical Device (SaMD)*[6] and subsequent *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device Action Plan,*[7] the US FDA has attempted to address the update problem. A key feature of the FDA's preferred approach is the requirement for manufacturers of MAMLS to provide algorithmic change protocols (ACPs) as part of their applications for premarket approval. ACPs are supposed to outline how a MAMLS will change over time and what the limits of these changes will be. They will also require manufacturers to state how they will mitigate any risks that these changes will present. The FDA suggests that this approach could allow MAMLS to be approved and deployed in clinical settings without a need for ongoing regulatory review.

A number of serious criticisms have been raised against the FDA's proposed framework.[16,17] For instance, the proposal gives little indication as to how the performance of MAMLS will be monitored in practice, even suggesting that manufacturers could monitor these systems themselves. We are sympathetic to many of the concerns that have been expressed in the literature. However, we believe that the current focus on regulatory challenges that MAMLS present has led researchers to overlook the broader set of ethical issues that the use of these systems in medicine will present.

TWO TYPES OF VARIATION: DIACHRONIC AND SYNCHRONIC

The literature on the ethics of MAMLS is cognizant that these systems will evolve over time—a phenomenon that we shall call *diachronic evolution*. As MAMLS continue learning from new data, their parametric weightings will change from update to update. They will respond differently to identical input data at different times. Their accuracy and performance will evolve over time, for better or worse. They may even adopt different classes of algorithmic bias as they continue to learn and evolve.

However, it is less often recognized that variation will emerge between copies of a MAMLS that have been implemented across different sites.

Synchronic variation refers to the differences that will emerge between copies of a MAMLS implemented at different sites or in different patients. MAMLS will be deployed across diverse clinical settings with different data collection policies, organizational procedures, user behaviors, data infrastructures, and patient demographics, each of which will affect the datasets upon which these systems learn. Even small variations in the datasets on which an algorithm learns can have significant effects on what it learns. If each copy of a MAMLS learns from data collected from the site at which it has been deployed, either exclusively, or even just to fine tune its parameters after initial learning from a training dataset, then these differences between site-specific datasets mean that copies of a MAMLS deployed at different sites (or devices implanted in different individuals) are likely to diverge over time. Eventually, identical data entered into different copies of a MAMLS will likely cause these systems to generate different outputs.

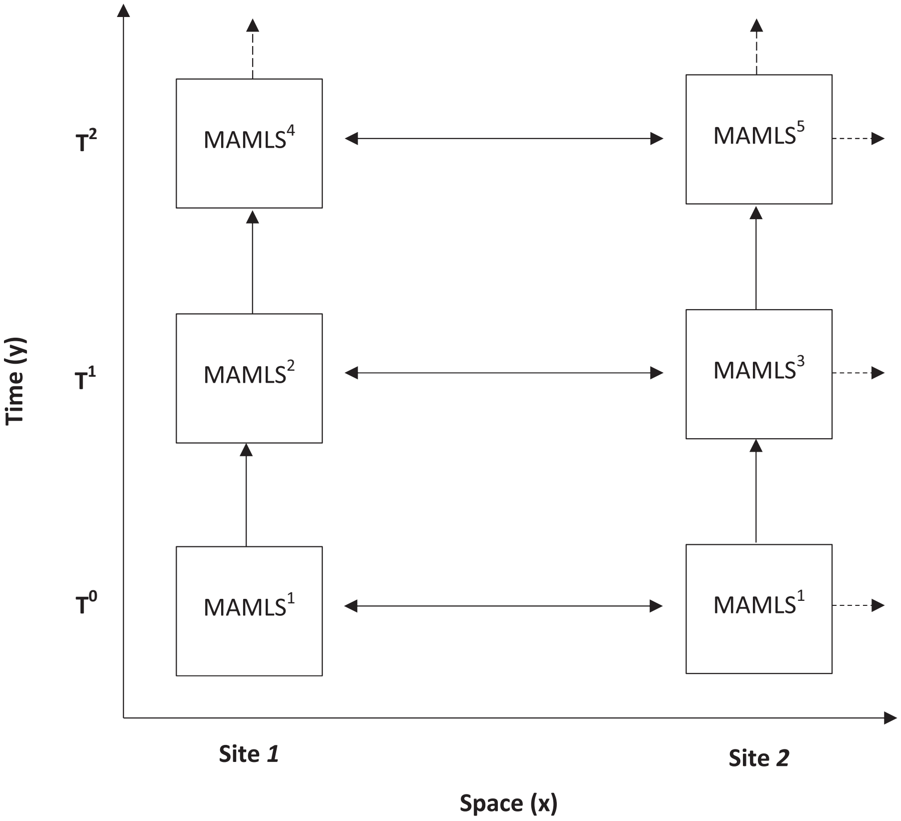

Figure 1. Schematic representation of diachronic evolution (y-axis) and synchronic variation (x-axis) in a MAMLS deployed across 2 sites.

Figure 1 illustrates the relation between these 2 types of variation.

Diachronic evolution alone would seem to require that regulators monitor the evolution of each MAMLS over time and conduct postmarket surveillance of the product's performance, errors, and instances of patient harm as it evolves. The possibility of synchronic variation suggests that, in addition regulators should monitor each *copy* of a MAMLS, due to the gradual divergence that may occur *between* each copy of the product over time. The additional administrative burden that this could entail would be expensive and time-consuming for both manufacturers, purchasers, and regulatory agencies. Moreover, the more instantiations there are, the more likely it is that one will go catastrophically wrong and undercut support for all of them. These factors could increase costs for manufacturers and chill the incentive to develop these systems in the first place.

FEDERATED LEARNING TO PREVENT SYNCHRONIC VARIATION

In some cases, however, manufacturers may have the option of eliminating synchronic variation entirely via the adoption of "federated learning" (FL), which "involves training statistical models over remote devices or siloed data centers, such as mobile phones or hospitals, while keeping data localized".[18] To date, interest in FL in healthcare has mostly been driven by their potential to maintain privacy.[4,19] However, if FL can be used to train MAMLS, they could allow each iteration of a MAMLS to learn on the same pool of data, which would preclude the emergence of synchronic variation.

Eliminating synchronic variation would have a number of benefits for stakeholders. For instance, it would reduce the burden of regulators and manufacturers by eliminating divergence and variation between copies of a MAML and thus the risk that different copies might require distinct regulatory evaluation and approval. In some cases, it might improve the generalizability of MAMLS by enabling these systems to learn from larger, more heterogenous datasets collected across multiple clinical sites. As we will argue below, it could also eliminate the potential for inequities in standards of care to emerge across clinical sites.

However, the decision to allow or eliminate synchronic variation carries practical trade-offs (along with some ethical trade-offs: see "Costs to particular cohorts" in the following section). Implementing FL in MAMLS may require updates to existing digital infrastructure that may be prohibitively expensive for many clinics and hospitals. Federated learning, for instance, "require[s] investment in on-premise computing infrastructure or private-cloud service provision and adherence to standardised and synoptic data formats so that ML models can be trained and evaluated seamlessly".[20] Moreover, if data collected at one hospital cannot easily be transferred to new sites without first being processed, this processing may itself introduce a form of synchronic variation by virtue of adding different extra layers of code to the AI system at different sites. Where implanted medical devices include MAMLS, FL will only be possible if these devices can transmit the result of training on local data back to other instantiations of the learning algorithm *and* can update the local algorithm in the light of the results of the training of other versions there-of, which increases the risk to patient privacy and of iatrogenic harm, including as a result of hacking.

In an important set of cases, then, users will face a choice between either allowing synchronic variation to occur, or not using MAMLS at all.

ETHICAL CONSIDERATIONS

The deployment and use of MAMLS generates a number of ethical concerns relating to the quality of patient care and doctor–patient relations, informed consent to treatment, threats to health equity and problems of obsolescence, and harms to particular cohorts of patients.

Impact on quality of care

The use of MAMLS presents a number of risks to the quality of patient care.

Left to continue learning postdeployment MAMLS may adopt erroneous and potentially dangerous associations from new input data that could jeopardize patient health. In one well-known case, a mortality-prediction ML system learned to classify asthmatic patients presenting to the emergency department with pneumonia as "low risk" due to a true but misleading correlation in the training data.[21] If operationalized, the system would have presented critical risks to patient safety. With MAMLS there is a risk that such errors will emerge postdeployment as a result the process of continual learning. Moreover, MAMLS are susceptible to the phenomenon of "catastrophic forgetting", in which a MAMLS overwrites what it has previously learned during the process of learning from new data, leading to sudden poor performance that could significantly jeopardize the quality of physician judgments and the health and safety of patients.[22] Finally, MAMLS are susceptible to hacking and adversarial attacks, including by "data poisoning". In locked ML systems, adversarial attacks can only affect individual outputs.[23] In MAMLS, however, adversarial attacks could interfere with the performance of the system (or systems) in all future uses. These possibilities highlight the urgency of the "update problem".

Challenges to clinical interpretation

Furthermore, achieving downstream benefits from the use of ML in medicine is critically dependent upon clinicians' ability to understand, interpret, and act on the outputs of these systems.[24] For instance, clinicians must decide how much epistemic weight they ought to give the outputs of algorithms in their clinical decision-making. Placing too much or too little weight on the outputs of an algorithmic system can result in patient harm, even death. An example of clinicians placing too little epistemic weight in the output(s) of an algorithmic system is "alert fatigue", which refers to "declining clinician responsiveness to a particular type of alert as the clinician is repeatedly exposed to that alert over a period of time, gradually becoming 'fatigued' or desensitized to it".[25] Alert fatigue can lead clinicians to ignore important alerts, potentially resulting in patient harm or death (for a particularly egregious instance of patient harm caused by alert fatigue, see reference 26). An example of patient harm caused by clinicians placing too much epistemic weight in the outputs of an algorithmic system is "automation bias", which "refers to errors resulting from the use of automated cues as a heuristic replacement for vigilant information seeking and processing".[27] The presence of diachronic evolution and synchronic variation in MAMLS will pose a significant challenge to clinicians being able to reliably interpret and act upon the outputs of these systems. If every time clinicians encounter a MAMLS it is a subtly (or occasionally not so subtly) different system—different both to previous iterations, and between patients and across clinical sites—it may be exceedingly difficult for them to be confident how it is functioning and how much they should trust it. These challenges are further complicated by the opacity of ML systems, which make it difficult to understand how or why a system works or has produced a certain output.[28]

Admittedly, that the performance of MAMLS will change over time, and will differ between sites and/or patients, does not in-and-of-itself distinguish them radically from other systems with which clinicians must engage in the course of their professional practice. Clinicians who work across different institutions often have to take account of differences in the way things are done, or particular devices are set up, in each institution. The fact that diagnostic tools and treatments are evolving all the time is, after all, why continuing medical education is so important. However, the key virtue of MAMLS is their ability to continuously improve at a faster rate than existing diagnostic tools without human intervention or oversight. The speed with which MAMLS evolve may outpace clinician's abilities to adapt to these changes.

Impact on doctor–patient relations

Where clinicians make use of MAMLS for the purpose of clinical decision-support, both diachronic evolution and synchronic variation will pose challenges to communication between doctors and patients and reduce the capacity for shared decision making. If clinicians are themselves not able to understand precisely what has changed between each update to a MAMLS system, or how, precisely, the system they are dealing with at this site, or in this patient, differs from other iterations, they may struggle to identify and explain the factors that are casually relevant to their ultimate decision about a diagnosis and/or treatment plan. In particular, they may find it difficult to provide the patient with counter factual information that might be relevant to shared decision-making about treatment. Importantly, this effect may occur even if the clinician is in fact justified—and can explain to the patient that they are justified—in relying on the MAMLS because of its superior accuracy relative to the alternatives.

It is sometimes argued that the use of AI and ML could allow clinicians more time to spend with their patients.[29,30] However, the various tasks associated with maintaining AI and ML systems could equally lead to increased administrative burdens for clinicians that could further interfere with the quality of care and empathy in the doctor–patient relationship.[31,32] This risk seems particularly acute in the case of MAMLS, because healthcare institutions will likely need to significantly expand the scope of their data collection policies and procedures to be able to provide the continuous stream of new data that training MAMLS will require.

The potential of MAMLS to evolve over time may also be expected to exacerbate the issues related to computers being "the third party in the room" in clinical consultations. As Christopher Pearce and others have noted, the introduction of computers into healthcare settings has transformed what was originally a dyadic relationship, between the doctor and patient into a triadic relationship between the doctor, the patient, and the doctor's computer.[33,34] Both doctor and patient now spend some—perhaps even much—of their time "together" looking at and relating to the computer: information provided by the computer shapes the course of the consultation. If the doctor's computer is—or accesses—a MAMLS then this will add an important temporal dimension to the relationship between the doctor and the computer and the patient and the computer. What the computer "says" may change over time. This alone may be sufficient to draw more of the doctor's and the patient's attention to the computer. However, the fact that the operations and the outputs of the MAMLS may change also opens up the possibility that doctors will become involved in trying to manage or shape those changes in order to meet their, and their patients, expectations. One might imagine clinicians trying to influence the evolution of the MAMLS by curating the data that they input into it, in the same way many of us now try to manage the recommendation engines of Spotify or Netflix. Clinicians' relationships with MAMLS will evolve along with the MAMLS and we should expect that at least some clinicians may want to be active in shaping the former evolution—and, thus, the latter.

Challenges to informed consent

Insofar as it is not typically considered necessary for clinicians to inform patients about the technologies that they have used to inform their clinical recommendations, the use of MAMLS for decision support is unlikely to have implications for informed consent. However, where MAMLS assist in the delivery of medical *treatment* (eg, robotic surgery or, hypothetically, AI-guided radiation therapy) their nature may well be relevant to the process of securing informed consent to treatment.

The use of ML in treatments already involves new risks that may need to be disclosed to patients, such as the threat of cyberattack.[23,35] Adaptive learning will introduce additional risks, including the risk of catastrophic forgetting and of algorithmic biases developing postdeployment, which may need to be disclosed to patients in order to allow them to make an informed decision of the use of such systems. Moreover, the potential of MAMLS to evolve and to differ between sites and patients means that the provision of general or "standard" information about treatments guided by MAMLS may not be sufficient to secure informed consent to treatment. A patient who returns to a medical clinic for treatment involving a MAMLS after some time will undergo treatment that may differ subtly, or even significantly, from that they received in their previous visit. Similarly, a patient who moves from one hospital to another, which has implemented a version of the same MAMLS, may be subject to different levels of risk—indeed, different risks—in each location. Fully informed consent, then, may require that patients are made aware of the risks associated with treatment by the particular MAMLS that is involved in their treatment. However, diachronic evolution and synchronic variation, coupled with the characteristic opacity of ML systems, mean that it may not always be possible for manufacturers to provide information about the specific risks associated with a particular iteration of a MAMLS.

Equity and obsolescence

One hopes that, with appropriate regulation, the continuous learning of MAMLS will lead to improved outcomes for patients over time. In-and-of itself, then, diachronic evolution in the performance of MAMLS should not raise issues of equity.

The ability of MAMLS to adapt to specific patient cohorts and improve the performance of the system among these cohorts has the potential to promote health equity by better serving the needs of minority groups that are often under-represented in the training data used to train locked models. However, where synchronic variation is permitted, it is also possible that the difference in the performance of MAMLS at different sites or in different patients may become so pronounced as to generate serious issues of justice in relation to the quality of healthcare available to different cohorts. Particular instantiations of a MAMLS may have biases that are more pronounced, more numerous, or more consequential within the patient cohort that they serve, than other iterations of the product deployed at different sites. Moreover, it is possible that some iterations of a MAMLS product may become stuck in local minima during the learning process, such that their performance stagnates while others continue to improve. In some cases, these performance disparities may become so large that the MAMLS available to particular sites/patients are effectively obsolete.

Costs to particular cohorts

The challenges that synchronic variation presents for equity may serve as another incentive for manufacturers and regulators to try to eliminate synchronic variation. However, although FL is likely to enhance the generalizability of MAMLS, as Futoma and coauthors have noted, "the demand for universal rules—generalisability—often results in [ML] systems that sacrifice strong performance at a single site for systems with mediocre or poor performance at many sites" [reference 36, see also reference 37]. Disease, symptoms, side-effects, and so on occur with differing probabilities across lines of race, sex, gender, ability, and so on, and the application of a one-size-fits-all model across different subpopulations will often result in a system having differing utility for members of different cohort. Indeed, it can result in a model that is suboptimal for all groups, or optimal only for the dominant subpopulation—a phenomenon known as "aggregation bias".[38] For this reason, the decision to prevent synchronic variation in MAMLS involves an ethical and political trade-off between prioritizing the health and well-being of dominant groups and the prioritization of the health and well-being of marginalized groups.

CONCLUSION

We have argued that the implementation of MAMLS raises a number of challenging ethical issues that have thus far received little attention. We distinguished between 2 sorts of variance that such systems may exhibit—diachronic evolution (change over time) and synchronic variation (difference between cotemporaneous instantiations of the algorithm at different sites). Diachronic evolution complicates the hermeneutic task of clinicians and could interfere with downstream patient health benefits. Maintaining the digital infrastructure and data collection requirements necessary to enable continual learning in MAMLS may generate greater administrative burdens for human physicians, resulting in compromised relations of care and empathy between doctors and patients. Synchronic variation has the potential to generate inequities between clinical sites using the same MAMLS. The choice between site-specific and FL approaches involves a trade-off between pursuing generalizability or local impact, and may be to the detriment of particular cohorts of patients. These ethical issues require sustained attention if we are to realize the benefits of continuous learning in medicine.

FUNDING

This research was supported by the Australian Government through the Australian Research Council's Centres of Excellence funding scheme (project CE140100012). RS is also an Associate Investigator in the ARC Centre of Excellence for Automated Decision-Making and Society (CE200100005) and contributed to this article in that role. JH was supported by an Australian Government Research Training Program scholarship. The views expressed herein are those of the authors and are not necessarily those of the Australian Government or Australian Research Council.

AUTHOR CONTRIBUTIONS

RS and JH were jointly responsible for the research design. JH completed the literature search and analysis. JH wrote the original draft, which was revised and edited by RS. Both authors then contributed to a further round of revisions and approved the final version of the manuscript for publication.

ACKNOWLEDGMENTS

The authors thank Mark Howard for drawing their attention to relevant literature.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

No new data were generated or analyzed in support of this research.

REFERENCES

1. Alpaydin E. *Introduction to Machine Learning*. 3rd ed. Cambridge, MA: The MIT Press; 2014.
2. Esteva A, Robicquet A, Ramsundar B, *et al.* A guide to deep learning in healthcare. *Nat Med* 2019; 25 (1): 24–9.
3. Rajkomar A, Dean J, Kohane I, Van Calster B, Wynants L. Machine learning in medicine. *N Engl J Med* 2019; 380 (14): 1347–58.
4. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med* 2022; 28 (1): 31–8.
5. Sparrow R, Hatherley J. The promise and perils of AI in medicine. *IJCCPM* 2019; 17 (2): 79–109.
6. FDA. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) – discussion paper and request for feedback; 2019.
7. FDA. Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan; 2021.
8. Banaei N, Moshfegh J, Mohseni-Kabir A, Houghton JM, Sun Y, Kim B. Machine learning algorithms enhance the specificity of cancer biomarker detection using SERS-based immunoassays in microfluidic chips. *RSC Adv* 2019; 9 (4): 1859–68.
9. Porumb M, Stranges S, Pescapè A, Pecchia L. Precision medicine and artificial intelligence: a pilot study on deep learning for hypoglycemic events detection based on ECG. *Sci Rep* 2020; 10 (1): 1–16.
10. Jia Z, Wang Z, Hong F, Ping L, Shi Y, Hu J. Personalized deep learning for ventricular arrhythmias detection on medical loT systems. In: *IEEE/ACM international conference on computer-aided design*, Digest of Technical Papers, ICCAD; November 2020.
11. Cook MJ, O'Brien TJ, Berkovic SF, *et al.* Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study. *Lancet Neurol* 2013; 12 (6): 563–71.
12. Pinto MF, Leal A, Lopes F, Dourado A, Martins P, Teixeira CA. A personalized and evolutionary algorithm for interpretable EEG epilepsy seizure prediction. *Sci Rep* 2021; 11 (1): 1–12.
13. Ong CS, Reinertsen E, Sun H, *et al.* Prediction of operative mortality for patients undergoing cardiac surgical procedures without established risk scores. *J Thoracic Cardiovasc Surg* 2021; doi: 10.1016/j.jtcvs.2021.09.010.
14. Yu S, Farooq F, van Esbroeck A, Fung G, Anand V, Krishnapuram B. Predicting readmission risk with institution-specific prediction models. *Artif Intell Med* 2015; 65 (2): 89–96.
15. Hong JC, Niedzwiecki D, Palta M, Tenenbaum JD. Predicting emergency visits and hospital admissions during radiation and chemoradiation: an internally validated pretreatment machine learning algorithm. *JCO Clin Cancer Inform* 2018; (2): 1–11.
16. Babic B, Gerke S, Evgeniou T, Glenn Cohen I. Algorithms on regulatory lockdown in medicine. *Science* 2019; 366 (6470): 1202–4.
17. Gerke S, Babic B, Evgeniou T, Cohen IG. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *NPJ Digit Med* 2020; 3 (1): 1–4.
18. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. *IEEE Signal Process Mag* 2020; 37 (3): 50–60.
19. Usynin D, Ziller A, Makowski M, *et al.* Adversarial interference and its mitigations in privacy-preserving collaborative machine learning. *Nat Mach Intell* 2021; 3 (9): 749–58.
20. Rieke N, Hancox J, Li W, *et al.* The future of digital health with federated learning. *NPJ Digit Med* 2020; 3 (1): 1–7.
21. Caruana R, Lou Y, Microsoft JG, Koch P, Sturm M, Elhadad N. Intelligible models for health care: predicting pneumonia risk and hospital 30-day readmission. In: proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. New York, NY: ACM; 2015; pp. 1721–30. http://dx.doi.org/10.1145/2783258.2788613.
22. van de Ven GM, Tolias AS. Three scenarios for continual learning. *arXiv.org* 2019; 1904.07734: 1–18.
23. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science (1979)* 2019; 363(6433): 1287–90.
24. Hatherley J, Sparrow R, Howard M. The virtues of interpretable medical AI. *Camb Q Healthc Ethics* 2022; doi: 10.1017/S0963180122000305.
25. Embi PJ, Leonard AC. Evaluating alert fatigue over time to EHR-based clinical trial alerts: findings from a randomized controlled study. *J Am Med Inform Assoc* 2012; 19 (E1): e145–48.
26. Wachter RM. *The Digital Doctor: Hope Hype, and Harm at the Dawn of Medicine's Computer Age*. New York, NY: McGraw-Hill Education; 2015.
27. Mosier K, Skitka LJ, Heers S, Burdick M. Automation bias: decision making and performance in high-tech cockpits. *Int J Aviat Psychol* 1997; 8 (1): 47–63.
28. Hatherley JJ. Limits of trust in medical AI. *J Med Ethics* 2020; 46 (7): 478–81.
29. Topol EJ. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York, NY: Basic Books; 2019.
30. Israni ST, Verghese A. Humanizing artificial intelligence. *JAMA* 2019; 321 (1): 29–30.
31. Sparrow R, Hatherley J. High hopes for "Deep Medicine"? AI, economics, and the future of care. *Hastings Cent Rep* 2020; 50 (1): 14–7.
32. Maddox TM, Rumsfeld JS, Payne PRO. Questions for artificial intelligence in health care. *JAMA* 2019; 321 (1): 31–2.
33. Pearce C, Arnold M, Phillips C, Trumble S, Dwan K. The patient and the computer in the primary care consultation. *J Am Med Inform Assoc* 2011; 18 (2): 138–42.
34. Pearce C, Sandoval M. Consulting with a computer: new frontiers. *Aust J Gen Pract* 2020; 49 (9): 612–4.
35. Kiener M. Artificial intelligence in medicine and the disclosure of risks. *AI Soc* 2021; 36 (3): 705–13.
36. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health* 2020; 2 (9): e489–92.
37. Burns ML, Kheterpal S. Machine learning comes of age local impact versus national generalizability. *Anesthesiology* 2020; 132 (5): 939–41.
38. Suresh H, Guttag JV. A framework for understanding unintended consequences of machine learning. *arXiv.org* 2019; 1901.10002. http://arxiv.org/abs/1901.10002