Cell Reports Medicine

CellPress
OPEN ACCESS

Perspective

Artificial intelligence in medicine: Overcoming or recapitulating structural challenges to improving patient care?

Alex John London[1,2,*]
[1]Department of Philosophy and Center for Ethics and Policy, Carnegie Mellon University, Pittsburgh, PA 15228, USA
[2]Twitter: @alexjohnlondon
*Correspondence: ajlondon@andrew.cmu.edu
https://doi.org/10.1016/j.xcrm.2022.100622

SUMMARY

There is considerable enthusiasm about the prospect that artificial intelligence (AI) will help to improve the safety and efficacy of health services and the efficiency of health systems. To realize this potential, however, AI systems will have to overcome structural problems in the culture and practice of medicine and the organization of health systems that impact the data from which AI models are built, the environments into which they will be deployed, and the practices and incentives that structure their development. This perspective elaborates on some of these structural challenges and provides recommendations to address potential shortcomings.

INTRODUCTION

Artificial intelligence (AI) has captured medicine's imagination, but its long-term prospects hinge on how it grapples with a paradox that reaches to the very heart of the healing art. On the one hand, the central and defining moral duty of healthcare providers is to use medical knowledge, clinical skill, and available interventions to safeguard and advance the health and well-being of patients. On the other, medical knowledge is incomplete, and the learning environment is noisy. The ongoing coronavirus disease 2019 (COVID-19) pandemic illustrates the dramatic challenges that arise from novel pathogens for which safe and effective interventions are unknown. However, in many areas of medicine, our understanding of the ecology and pathophysiology of sickness and disease are incomplete and evolving. As a result, expert opinion about which practices, procedures, or interventions will advance the duty of care can be uncertain, disputed, or simply mistaken. This paradox sometimes leads to clinicians acting from benevolent intent grounded solely in clinical experience to deliver care, the actual result of which is ineffective or harmful. In the face of uncertainty, conflicting judgment, or novel circumstances, the duty to care can only be realized in practice if it is accompanied by a duty to learn.[1]

The widespread enthusiasm for AI in medicine expresses the ambition of harnessing vast repositories of medical information to create resources and practices within health systems that enable stakeholders to learn more efficiently and thereby increase the likelihood that caregivers will produce beneficial outcomes when they act with therapeutic intent. Part of the challenge, however, is that some of the very deficiencies that AI is intended to overcome permeate the data on which those systems rely for learning, the practices that structure the development of such systems, and the environments in which AI systems must be deployed to produce better outcomes for patients. These challenges create the prospect that, without thoughtful reforms, the introduction of AI into medicine could have three unintended consequences. First, it might expand, rather than reduce, unwarranted variation in medical practice. Second, it might impede our ability to ensure that innovation produces significant social value. Third, it might alter the distribution of effort and resources in clinical practice without improving outcomes for patients or making health systems more effective (better ability to diagnose, prevent or treat sickness, injury, or disease), efficient (better ability to achieve these goals more quickly with fewer resources), or equitable (better ability to effectively meet the health needs of the diverse populations health systems serve). After outlining these challenges, this perspective offers some recommendations for addressing them and better ensuring that AI systems will help discharge the duty to learn in a way that enhances the equity, effectiveness, and efficiency of health systems.

THE DATA WE NEED FOR SOCIAL VALUE

The ambition of organizing the practices and resources within health systems to learn more efficiently derives from an important moral imperative. This is the imperative to ensure that health systems can effectively, efficiently, and equitably address important health needs of the individuals and populations they serve.[2] Learning initiatives that enhance the capacity of stakeholders within health systems to better function on one or more of these dimensions are supported by a strong claim to generate social value—an important requirement for ethical research and learning.[3–5]

Please cite this article in press as: London, Artificial intelligence in medicine: Overcoming or recapitulating structural challenges to improving patient care?, Cell Reports Medicine (2022), https://doi.org/10.1016/j.xcrm.2022.100622

Relying heavily on data gathered from current practices poses a challenge for this ambition. In particular, it may be difficult to ensure that learning activities are directed at questions of importance for patients and health systems if the problems pursued are dictated by the availability of datasets rather than by determinations regarding which knowledge gaps have the most significant impact on priority health needs.[6] The availability of large datasets may provide useful grist for the mill for computer scientists and researchers interested in developing AI technology, but in order to produce significant value for patients and health systems, learning initiatives must align with open questions that impact clinically important aspects of provider practice and patient outcomes. If the goals of AI system development are driven by objectives that are not clearly connected to meaningful patient outcomes, research productivity (measured by grant funding, papers published, citations, etc.) is unlikely to translate into advances that promote health priorities for patients or health systems.

Relying on data generated from current practice also poses challenges to the goal of creating practices and health systems that function equitably in the sense of providing effective care to the full diversity of individuals and groups who rely on those health systems. Inequalities in access to high-quality healthcare often translate into disparities as to which populations are represented in medical databases. For example, genome-wide association studies play an increasingly important role in drug development, but a 2016 study of 2,500 studies involving 35 million samples showed that 81% of participants were of European ancestry, while people of African ancestry accounted for only 3% of participants.[7] Similar problems have been seen in other medical databases.[8,9] Under-representation of this kind can lead to the development of AI systems whose utility deteriorates when applied to groups already under-served in current health systems—thereby widening pre-existing disparities in quality of care and health outcomes.

Even when diverse populations are represented in data routinely generated within health systems, concerns about whether those data can support research with high social and clinical value are amplified by the extent to which data are entangled with the practices from which, and the purposes for which, they are gathered.[10] Although some of these challenges might be addressed through technical solutions, others require adjustments to the practices and the health systems from which health data are generated.

In the United States, patterns of exclusion and oppression that target individuals and groups on the basis of characteristics such as race, ethnicity, sex, or gender have been encoded into a range of basic social structures. These social structures include banking and access to capital, employment, housing, law enforcement, and the penal system. Injustice in the operation of these social structures has produced broad disparities in access to the social determinants of health which, in turn, have produced higher rates of avoidable morbidity and mortality in targeted social groups.[11–14] Data generated within United States health systems thus reflect complex interactions between human biology, socio-economic conditions, and the prior practices of providers and health systems at a given time. Nevertheless, these widespread and persistent health disparities have sometimes been treated as reflecting diversity in the baseline biological functioning of different classes of people rather than as the physical effects of cumulative burdens of social and economic inequalities grounded in the unfair or unjust operation of important social structures. This in turn has resulted in practices such as adjusting diagnostic or treatment criteria across a wide range of specialties on the basis of race, a social construct that is often reified as demarcating biologically distinct groups.[15,16] It has been estimated that race-based adjustments to equations used to calculate estimated glomerular filtration rate (eGFR) levels for determining the severity of kidney disease alone results in the under-treatment of 3.3 million Black Americans who, without these corrections, would be more likely to receive earlier treatment for a range of complicating conditions.[17]

When health disparities are the product of unequal access to the social determinants of health, hardcoding them into clinical practice is likely to systematically produce worse outcomes for groups who are already under-served by health systems and who are more likely to experience higher rates of morbidity and mortality. Health data produced from such a system will reflect this biased view of the biology of different groups and the biased treatment practices these views have engendered. This is one example where bias in the data on which healthcare providers rely for decision-making is shaped by legacies of injustice within a range of background social structures that influence the life prospects of people. Healthcare systems are also social structures with histories of unequal treatment for marginalized groups, and biased practices and attitudes of providers can influence the care of patients more directly. In a recent analysis of over 40,000 medical histories and chart notes from over 18,000 patients, Sun and colleagues found that "Black patients had 2.54 times the odds of having at least one negative descriptor in the history and physical notes."[18] Uncritical reliance on such data thus risks further recapitulating within medicine some of the background health inequalities that disproportionately affect members of marginalized or oppressed social groups.

Background racial, social, and economic factors that shape the life prospects of patients and the operation of health systems are widespread confounders in medical data.[19] Increasing the volume of such data will be of limited value if it continues to reflect these common background conditions. Such weaknesses can sometimes be addressed using techniques from causal reasoning to better distinguish the relative contributions of biological, social, or other influences on health.[20,21] However, these techniques often require that we already understand important elements of the underlying causal system.[22] For example, instrumental variables can be used to distinguish the relative causal contributions of variables that are associated with an outcome,[23] but treating a variable as instrumental requires that we understand its relationship to the other variables in question. In medicine, this kind of granular causal knowledge is often absent. In the absence of such background knowledge, however, it is particularly difficult to make full and effective use of powerful tools from causal discovery. In such cases, uncertainty about the underlying causal relationships between various influences on health limits our ability to use the data that we generate

Please cite this article in press as: London, Artificial intelligence in medicine: Overcoming or recapitulating structural challenges to improving patient care?, Cell Reports Medicine (2022), https://doi.org/10.1016/j.xcrm.2022.100622

from ordinary medical practice to answer questions of deep importance to the effective and equitable practice of medicine.

The value of data produced from the normal operation of health systems is further limited by the fact that healthcare data reflect the purposes for which it was gathered. Despite efforts to improve data gathering in medicine, electronic health records (EHRs) continue to reflect administrative goals, such as billing, that do not necessarily align with the information needs of science and learning. For variables to appear in data tables or models, we must collect the relevant data. But in areas where medical uncertainty is the greatest, such as central nervous system and brain disorders, it may not be clear whether our current theories of disease pathophysiology are adequately informative about the features of patients that we need to measure. It is likely that many features that are relevant to disease course or treatment success are not captured in existing datasets.

Alternatively, even in cases where we are capturing features of patients and treatments that are relevant to particular scientific questions, that data might not be gathered with the frequency, granularity, or bandwidth necessary to distinguish relevant relationships. This is a critical limitation since many bodily systems are homeostatic, with compensatory feedback mechanisms that can make it difficult to discover the direction of causation among relevant variables. In such cases, without a rich time series of data, variables that are causally related can appear to be statistically independent. For example, if you drive a car uphill at a constant speed, the slope of the hill and the position of the pedal might appear to be independent of speed if the speed remains constant as these variables change. In the body, the dilation of vasculature and the delivery of a particular medication may appear to be independent of blood pressure if successful treatment keeps the latter constant. Without sufficiently granular time-series data, standard machine-learning techniques can fail to find dependencies between variables or can impute incorrect relationships.[24] Here again, capturing data at appropriate intervals and with sufficient granularity depends on prior knowledge of the underlying system. It also requires a practice and infrastructure for data collection and storage that can be more easily tailored to the specific goals of research and learning.

Finally, uncoordinated practices for assembling and disseminating data result in datasets that are not well suited to training AI systems. Some collections of images generated during the COVID-19 pandemic were expanded, and then the collection was renamed and recirculated, creating the prospect that the datasets used to validate AI system performance were not independent of the datasets used to train those systems.[25] Common practices for resizing or formatting images create potential confounders on which AI systems might condition when associating images with outcome labels.[26] Common clinical practices also create potential confounders such as when the sickest patients in a hospital are most likely to have radiographs taken from portable machines, or from a particular angle, allowing algorithms to improve their performance by conditioning on these arbitrary features rather than aspects of the underlying pathology.[27] In this respect, aspects of health data that humans might overlook in their focus on pathology can pose a fundamental challenge to AI systems. Creating datasets that better control for such confounding artifacts requires changes to practices across the life cycle of data production, acquisition, storage, and transmission.

Ensuring that AI development addresses questions that are critical to producing the most significant benefits for patients and health systems may necessitate a development pipeline in which research questions derived from these goals determine the nature, frequency, granularity, and bandwidth of data that need to be gathered and the standardization necessary to ensure that they are most likely to support these specific learning tasks. Such initiatives, however, would likely require alterations to clinical, data-gathering, and data-management practices.

THE INTERVENTIONAL AMBITION OF LEARNING HEALTH SYSTEMS

The ambition of improving the ability of stakeholders within health systems to learn is fundamentally transformative and interventional; the goal, frequently, is not simply to understand what is happening, or to predict what is likely to happen in the future, but to alter and reconfigure clinical, organizational, and institutional policies, practices, and settings to make healthcare more effective, efficient, and equitable. This transformative ambition is in deep tension with the limitations of current AI technologies.

The most mature AI systems are well suited to two kinds of tasks: (1) classification, such as detecting or diagnosing pathology and (2) predicting or prognosticating outcomes that are likely to occur under the assumption that current practices remain consistent with the practices that are reflected in data. Predictions generally take place under this assumption because they are derived from complex patterns of association between the variable of interest and the potentially vast range of features or variables whose past values have been measured and tracked. The resulting models have been described as atheoretic or theory free, in the sense that models of the relationships between variables in the domain of interest are constructed from relationships in data and need not reflect the substantive domain knowledge of any expert.[28,29] For example, if tobacco-stained fingers in patients with various symptoms have been associated with cancer, or with a poor prognosis, then we might use that association to estimate the probability that a current patient with yellow fingers also has cancer or will have a poor prognosis.

The complex statistical relationships that AI systems use for diagnostic or prognostic purposes often cannot provide a guide to intervention because intervention involves altering the state of the world with the goal of changing the variable of interest. Formally speaking, there is a gap between models that entail the probability of the value of the target variable t on the basis of the value m of the measured variables ($p(t|m)$) and models that entail the probability of the value t if we intervene to make m the value of the measured variables $p(t|\ \text{do}\ (m))$.[30] In other words, even if tobacco-stained fingers at age 50 are a reliable predictor of the probability of lung cancer by age 60, intervening to clean the tobacco stains from the fingers of patients at 50 is not going to reduce their probability of lung cancer by age 60.

Predictive models can be a sound guide to intervention under special circumstances. This is the case with Google's Alpha Go, where the structure of the board is known, the current board

Please cite this article in press as: London, Artificial intelligence in medicine: Overcoming or recapitulating structural challenges to improving patient care?, Cell Reports Medicine (2022), https://doi.org/10.1016/j.xcrm.2022.100622

positions are known, and the pieces can only move according to known and clearly determined laws. Under these conditions, reinforcement learners can explore the way that millions of board configurations constrain future moves and the probability of victory.[31] The critical point is that few problems in medicine have this structure. In many cases, the set of relevant variables might be unknown, especially when we are contemplating novel interventions. Similarly, the laws by which these variables interact might remain unknown, so assumptions about how small changes in one configuration of variables (analogous to the present state of the board in Go) will constrain the future states of those variables might be unclear or mistaken.

As a result, hypotheses about counterfactuals such as what will happen if we allocate resources differently, change treatment practices, or alter the criteria for assigning treatments often lie outside the scope of questions that existing data can address. The reason is that, unlike the game of Go, proposals to alter medical practices are often not simply choosing a strategy that will exploit a set of relationships that are already well represented in the data we have. Rather, counterfactuals in medicine frequently involve introducing policies, actions, or interventions that are novel in the sense that we do not have prior experience with their effects. If the permutations of actions and outcomes represented in health data do not capture the alternatives that are relevant to interventions derived from new theories of organizational behavior, disease pathophysiology, or drug mechanism, then AI systems will struggle to identify new practices that are superior to the status quo.[32]

Recent advances in causal inference and causal structure learning seek to overcome deficiencies in the more limited predictive approaches of traditional AI techniques.[33] Work in this area is promising, but it is constrained by many of the same limitations that face any data-driven approach to learning. In particular, deficits in the completeness of measured variables and in the granularity, frequency, and bandwidth of those measurements limit the usefulness of these techniques. These approaches to learning face the additional challenge that the models they employ rely on more substantive assumptions about relationships of variables in the domain under study. In cases where medical uncertainty is the greatest, we may be most constrained in our ability to supply the knowledge needed to most effectively utilize advances in this area.

As a result, the clinical utility of diagnostic and predictive systems hinges on whether improving the accuracy or speed of diagnosis or prediction can improve patient outcomes by, for example, reducing unnecessary delays in access to independently established effective care without creating new inefficiencies by over-diagnosing disease that is not clinically meaningful.

The ambition of streamlining diagnosis and expediting access to independently established effective care is illustrated by FDA-approved AI systems for diagnosing more than mild diabetic retinopathy from retinal-fundus images.[34] These systems seek to improve patient outcomes by eliminating delays in screening while maintaining uniformly high degrees of sensitivity and specificity to better optimize the rate at which high-risk patients are referred to specialists. The advantage of such systems is that they seek to streamline the process of identifying patients at high risk of vision loss and then to link them to care whose clinical merits have been independently established.

In contrast, many hospitals sought to use predictions about the likely cost of care for newly admitted patients to reallocate clinical services in the hope of improving patient outcomes. However, using a system that predicts risk in the domain of cost to support interventions in the domain of care had the effect of systematically giving lower priority to the health needs of Black patients who were equally as sick as their White counterparts.[35] The reason is that the widespread operation of social forces discussed above produces both health outcomes and patterns of medical care in which Black patients who fall into a projected expense category are likely to be sicker than their White counterparts in the same expense category.

In this example, a system that might be actuarially correct at predicting patient costs was used to alter patient care. The gap between prediction and intervention was bridged by the tacit, and ultimately false, background assumption that patients in the same cost category would likely have equivalent health status. Rather than reducing unwarranted variation in care and improving the practice of evidence-based medicine, using current AI systems to advance the fundamentally interventional goals of medicine is likely to increase the proliferation of the very kind of unwarranted variation in clinical practice that learning health systems in general, and AI systems in particular, are supposed to reduce.

If prediction models are not integrated into clinical pathways that direct high-risk patients to health services that have been established as effective at reducing or eliminating those risks, then the rapid development and proliferation of AI systems can work against the fundamental ambitions of learning health systems. Rapid advances in prediction that do not link patients to independently established effective care for the condition in question might encourage variation in medical practices whose efficacy is unknown and that are carried out in clinical contexts in which it is difficult to learn. In such cases, AI systems can play a valuable role in generating hypotheses about the likely value of new interventions, practices, policies, and organizational or institutional configurations that might better address the health needs of patients. But testing those hypotheses will likely require reconfiguring the delivery of healthcare to facilitate the conduct of randomized controlled trials.

An approach to learning that centers patient needs and gives priority to hypotheses about issues that require departures from current practice or that do not map neatly onto current datasets can only be advanced through the creation of novel datasets. To control for confounding and to isolate the effects of alterations to clinical practice, it may be important to generate such data using randomized trials employing pragmatic designs, such as cluster randomization. This approach has the additional advantage of seeking to evaluate proposed changes in practice or organizational behavior before they become widespread and entrenched into organizational habit or culture. In such cases, AI may have a valuable role to play, especially with respect to transfer learning—our ability to extrapolate relationships in one domain to another—but this requires abandoning the heady ambition of replacing traditional research methods with AI systems. It requires, instead, incorporating AI into a more diverse portfolio of

Please cite this article in press as: London, Artificial intelligence in medicine: Overcoming or recapitulating structural challenges to improving patient care?, Cell Reports Medicine (2022), https://doi.org/10.1016/j.xcrm.2022.100622

tools for learning, some of which require changes to practices and systems for delivering care, controlling confounders, and gathering data.

FRAGMENTATION, REDUNDANCY, AND UNWARRANTED VARIATION

The push to promote the ability of stakeholders within health systems to learn is motivated, in part, by the need to overcome fragmented, uncoordinated, and unwarranted medical practices, but the ability of AI systems to advance these goals is frustrated by the degree to which the ecosystems in which these systems are being developed suffer from these same shortcomings.

The current ecosystem of development for AI systems in medicine incentivizes the proliferation of models designed to perform similar tasks.[36] The result is a dispersion of effort across development teams that creates a high rate and volume of published papers and a high degree of redundancy in models, most of which reflect shallow development trajectories.[37]

The development trajectory of an AI system can be shallow in several respects. First, AI models tend to be "validated" *in silico*, with few tested on external datasets and still fewer subject to prospective testing in real-world environments. Wessler et al. note that 58% of the cardiovascular prediction models (CPMs) in the Tufts Predictive Analytics and Comparative Effectiveness CPM Registry had never been externally validated.[37] McDermott et al. note that "whereas ~80% of computer vision studies and ~58% of natural language processing studies used multiple datasets to establish their results, only ~23% of [machine learning applied to health] papers did this." [19] Wu and colleagues report that public documents show that all but 4 of 130 AI devices to receive FDA approval appear to have been evaluated solely on the basis of retrospective studies, and none of the 54 high-risk devices appear to have been evaluated prospectively. They also note that the performance of 93 of these devices appears not to have been evaluated across multiple implementation sites.[38] Although validation of a model on independent datasets is important, prospective studies of complete technologies (AI models plus the practices and procedures and training necessary to implement those models in actual practice) under real-world conditions provide a more accurate picture of what to expect from the use of AI technologies in practice. However, the development of AI systems is often treated as quality improvement rather than as research, and so high-quality prospective studies that control for bias and have greater relevance to clinical practice remain rare.[39,40]

Second, showing that an AI model can perform a desired task on a range of fixed datasets may shed light on the efficacy of that model under idealized conditions, but the clinically relevant question for patients, care providers, and health systems is whether an AI model can be incorporated into a system that can be deployed in clinical practice with sufficient effectiveness in the real world as to offer a net benefit to stakeholders. Models that perform well in idealized laboratory environments may fail to provide a net benefit to patients for many reasons. A recent review of 65 randomized controlled trials (RCTs) found that two-fifths of the prediction tools evaluated in these trials that "achieved good performance in observational model development and/or validation studies failed to show clinical benefit for patients compared to routine clinical treatment."[41]

Even if AI systems are deployed on the basis of solid evidence that they can provide a net advantage to patients at the time of deployment, changes in clinical practice, background rates of concurrent disease, and other social and environment factors can shift the distribution of patient attributes in ways that require AI systems to be retrained or reevaluated.[42,43] This was illustrated during the current pandemic by an increase in sepsis alerts in health systems using AI systems to monitor clinical care.[44] Establishing that AI systems can promote clinical value is not a threshold that is crossed in development, which can then be taken for granted once a system is incorporated into clinical workflows. System performance requires constant monitoring, a process that may require alterations to clinical workflows and the mix of expertise on healthcare teams.

One ambition for the use of AI in medicine is to capture the data necessary to reduce unwarranted variation in clinical practice, thereby weeding out ineffective practices and stewarding scarce resources toward more effective alternatives. Ironically, the landscape of AI in medicine reflects a similar proliferation of systems that perform the same task but whose relative clinical merits are not well characterized and not easily compared. Rather than reducing unnecessary variation, the current state of AI has the potential to simply add another dimension to this range of variation in clinical care.[36]

AN UNHEALTHY KNOWLEDGE ECOSYSTEM

Finally, the fact that the knowledge ecosystem surrounding AI in medicine is cluttered with hype and misinformation jeopardizes the prospect that AI can help medicine give practical substance to the idea of a learning health system. By a "knowledge ecosystem," I mean the shared understanding of the medical profession relating to the capabilities and limitations of various medical technologies. In a healthy knowledge ecosystem, stakeholders such as clinicians and health-system administrators should have sufficiently clear understanding of the advantages and limitations of various technologies that they can interface fruitfully with specialists who command a deeper and more substantial knowledge of that technology. A healthy knowledge ecosystem is thus essential to a fruitful division of social and epistemic labor, in which the stakeholders who provide care, manage health systems, and set policy for treatment and payment can use policies, practices, and interventions to improve patient outcomes, the efficiency of healthcare delivery, and the fairness with which scarce resources are shared.

The knowledge ecosystem surrounding AI in medicine is inflated with hype. Machine-learning systems are described as approaching "problems as a doctor progressing through residency might: by learning rules from data."[45] Articles in top journals speculate about the prospect that AI will render doctors obsolete,[46] and major medical centers have invested millions in initiatives such as IBM's Watson.[47] Yet, systematic reviews of AI applications in medicine find that few are ready for clinical deployment.[25,39,40,48] The Watson program in health, once described as the "future of healthcare," has been "sold for scraps" in what some commentators describe as the "total

Please cite this article in press as: London, Artificial intelligence in medicine: Overcoming or recapitulating structural challenges to improving patient care?, Cell Reports Medicine (2022), https://doi.org/10.1016/j.xcrm.2022.100622

failure'' of a program that over-promised and under-delivered from the start.[49,50] Descriptions of AI systems as learning from experience in the way that medical students learn fosters the impression that the current generation of systems—often brittle and trained to perform a single task—have access to the kind of background knowledge and global intelligence that allow humans to learn across domains, from few examples. This mismatch between the inflated rhetoric surrounding AI capabilities and practical realities on the ground reflects, in part, the extent to which stakeholders in medicine have been unprepared for, and relatively naive in their response to, the self-serving promotional activities of entrepreneurial entities looking to capture a portion of the $3 trillion United States healthcare market. It also creates a dynamic in which it can be difficult for clinicians, administrators, policy makers, and other stakeholders to use a detailed understanding of the capabilities and, most importantly, the limits of current AI systems to counter the bandwagon effects of inflated enthusiasm.

An unhealthy knowledge ecosystem surrounding AI is a recipe for misapplication and miscalibrated trust.[51] This can take the form of over-trusting systems whose performance in real-world settings has not been adequately verified or encouraging the use of systems trained to make predictions in one domain to support intervention decisions in another.[28] Alternatively, miscalibrated trust can take the form of under-trusting well-designed and carefully validated systems merely because they are the application of an algorithm.[52] It also makes it difficult for care providers and health systems to navigate the proliferation of systems, to make informed decisions about which systems can be usefully incorporated into clinical workflows, and to clearly and readily assess what steps might need to be taken to ensure that such systems provide a net benefit to patients and health systems.

RECOMMENDATIONS

To ensure that AI initiatives in medicine are better aligned with the moral imperative to ensure that health systems function effectively, efficiently, and equitably for individuals and communities, a wide range of stakeholders should take steps to improve four features of the knowledge ecosystem relating to medical AI.

First, funding agencies, health systems, clinicians, and AI developers should invest in efforts to identify areas where the type of tasks that can be performed by current AI systems would add significant value to the equity, effectiveness, or efficiency of health systems. Frameworks from a value-sensitive design can facilitate this goal through a process of stakeholder engagement in which technical aspects of system development are evaluated within a context structured by clinical goals and their relative value to stakeholders.[53] Fostering design practices that foreground questions of clinical value at multiple points in system development might better align pipelines for AI development in medicine with the goal of ensuring that innovation targets questions of high clinical and social value.

Second, a significant part of these initiatives should include an assessment of the clinical data required to develop these systems and whether existing data are fit for purpose. Greater emphasis should be placed on funding and coordinating initiatives designed to gather data with the standardization, frequency, granularity, bandwidth, and attention to reducing or controlling confounders necessary to support the development of accurate and equitable AI systems. Initiatives like the NIH's Bridge2AI, which seeks to gather datasets specifically designed to support the development of AI systems, might be a step in the right direction, depending on the extent to which data-collection efforts are aligned with priority health needs and with the requirements of the specific scientific questions that such datasets are anticipated to support.

Third, funding agencies, professional societies, academic institutions, and health systems should develop incentives for researchers to engage in the sustained, cross-disciplinary collaborations that are often necessary to deepen development trajectories for AI systems. This might be facilitated through training programs focused on expanding the knowledge base of key stakeholders. This might include training for clinicians and health system administrators about the capabilities and limits of AI and clinical rotations to expose researchers in computer science to the complexities of clinical decision-making and healthcare delivery.

Fourth, stakeholders from AI and medicine, including researchers, journal editors, and regulators, should develop standards for better elucidating the maturity of AI systems and the level of evidence supporting specific claims to clinical utility. A framework for indicating level of maturity might utilize categories that characterize the development trajectory for a system, from basic science, proof of concept, and concept refinement through confirmatory testing in real-world contexts and post-deployment monitoring. It should include a clear statement on the intended clinical use case for a system along with possible uses for which the merits of the system have not been validated. It should also include a statement on the extent to which system performance has been externally validated, the nature of that validation, whether the conditions necessary to replicate system performance in actual practice have been established, and whether the performance of this system, with specified training for operators, has been evaluated in real-world settings (M. McCradden, S. Joshi, J.A. Anderson, and A.J.L., unpublished data).

Finally, implementation science should play a greater role in structuring and evaluating proposals to implement AI systems in healthcare settings. Implementation scientists might serve as a bridge between AI researchers and healthcare teams and facilitate the evidence-based, longitudinal assessment of such systems.

CONCLUSION

Many of the ambitions driving enthusiasm about AI in medicine are worthwhile and important. They include the goals of promoting decisions that are informed by a comprehensive assessment of available data, ensuring that the relative merits of policies, practices, procedures, and interventions are clearly assessed before they are implemented in practice and continuously monitoring clinical workflows and practice patterns so that they can be adjusted in light of changing circumstances. Applications of AI are likely to play a valuable role in advancing each of these

goals. In that sense, the point of the present analysis is not to convey techno-pessimism.

The point of the present analysis, rather, is to emphasize that when it comes to the goals of promoting the ability of stakeholders to learn within health systems, AI is neither autonomous nor exceptional. It is not autonomous in that it cannot decide which questions to pursue, whether datasets are adequately complete or fit for purpose, which training and testing methods promote both accuracy and equity, and which practices for incorporating AI systems into clinical decision-making increase the efficiency of health systems. It is not exceptional in that, like all other methods, tools, and interventions, its ability to advance important social and clinical objectives hinges critically on decisions that are made by a wide range of stakeholders, each of whom faces a series of local incentives that may cohere with and support or conflict with and undermine socially desirable ends to varying degrees.[1]

Uncoordinated decisions by stakeholders pursuing their own interests in such an environment do not necessarily ensure that novel technologies will advance important social and clinical objectives. In our fascination with the potential for AI, we too often overlook these mundane limitations. In that regard, our ability to use AI to promote the goals of learning health systems will depend on our ability to reshape aspects of health systems on which this new technology depends, into which it will be inserted, and from which it is being developed.

ACKNOWLEDGMENTS

The author thanks four anonymous referees for their incisive and constructive feedback on an early draft of this manuscript.

AUTHOR CONTRIBUTIONS

A.J.L. conceived, wrote, revised, and edited this manuscript.

DECLARATION OF INTERESTS

The author declares no competing interests.

REFERENCES

1. London, A.J. (2021). For the Common Good: Philosophical Foundations of Research Ethics (Oxford University Press).
2. London, A.J. (2021). Self-defeating codes of medical ethics and how to fix them: failures in COVID-19 response and beyond. Am. J. Bioeth. *21*, 4–13.
3. Wendler, D., and Rid, A. (2017). In defense of a social value requirement for clinical research. Bioethics *31*, 77–86.
4. Wenner, D.M. (2017). The social value of knowledge and the responsiveness requirement for international research. Bioethics *31*, 97–104.
5. London, A.J. (2019). Social value, clinical equipoise, and research in a public health emergency. Bioethics *33*, 326–334.
6. Passi, S., and Barocas, S. (2019). Problem formulation and fairness. In Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 39–48.
7. Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nat. News *538*, 161.
8. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., and Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc. Natl. Acad. Sci. U S A. *117*, 12592–12594, Epub 2020 May 26. PMID: 32457147; PMCID: PMC7293650. https://doi.org/10.1073/pnas.1919012117.
9. Daneshjou, R., Smith, M.P., Sun, M.D., Rotemberg, V., and Zou, J. (2021). Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. JAMA Dermatol. Published online September 22. https://doi.org/10.1001/jamadermatol.2021.3129.
10. Cabitza, F., Ciucci, D., and Rasoini, R. (2019). A giant with feet of clay: on the validity of the data that feed machine learning in medicine. In Organizing for the Digital World (Springer), pp. 121–136.
11. Yearby, R. (2020). Structural racism and health disparities: reconfiguring the social determinants of health framework to include the root cause. J. Law Med. Ethics *48*, 518–526.
12. Bailey, Z.D., Feldman, J.M., and Bassett, M.T. (2021). How structural racism works—racist policies as a root cause of US racial health inequities. N. Engl. J. Med. *384*, 768–773.
13. Bailey, Z.D., Krieger, N., Agénor, M., Graves, J., Linos, N., and Bassett, M.T. (2017). Structural racism and health inequities in the USA: evidence and interventions. Lancet *389*, 1453–1463.
14. Phelan, J.C., and Link, B.G. (2015). Is racism a fundamental cause of inequalities in health? Annu. Rev. Sociol. *41*, 311–330.
15. Vyas, D.A., Eisenstein, L.G., and Jones, D.S. (2020). Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. N. Engl. J. Med. *383*, 874–882.
16. Ioannidis, J.P., Powe, N.R., and Yancy, C. (2021). Recalibrating the use of race in medical research. JAMA *325*, 623–624.
17. Tsai, J.W., Cerdeña, J.P., Goedel, W.C., Asch, W.S., Grubbs, V., Mendu, M.L., and Kaufman, J.S. (2021). Evaluating the impact and rationale of race-specific estimations of kidney function: estimations from US NHANES, 2015-2018. EClinicalMedicine *42*, 101197.
18. Sun, M., Oliwa, T., Peek, M.E., and Tung, E.L. (2022). Negative patient descriptors: documenting racial bias in the electronic health record: study examines racial bias in the patient descriptors used in the electronic health record. Health Aff. *41*, 10–1377.
19. McDermott, M.B., Wang, S., Marinsek, N., Ranganath, R., Foschini, L., and Ghassemi, M. (2021). Reproducibility in machine learning for health research: still a ways to go. Sci. Transl. Med. *13*, eabb1655.
20. Pearl, J., and Mackenzie, D. (2018). The Book of Why: The New Science of Cause and Effect (Basic Books).
21. Spirtes, P., Glymour, C.N., Scheines, R., and Heckerman, D. (2000). Causation, Prediction, and Search (MIT Press).
22. McCradden, M.D., Joshi, S., Mazwi, M., and Anderson, J.A. (2020). Ethical limitations of algorithmic fairness solutions in health care machine learning. Lancet Digital Health *2*, e221–e223.
23. Athey, S. (2017). Beyond prediction: using big data for policy problems. Science *355*, 483–485.
24. Danks, D., and Plis, S. (2019). Amalgamating evidence of dynamics. Synthese *196*, 3213–3230.
25. Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., and Schönlieb, C.B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat. Mach. Intell. *3*, 199–217.
26. Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., and Oermann, E.K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. *15*, e1002683.
27. Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., and King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Med. *17*, 1–9.
28. London, A.J. (2019). Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Cent. Rep. *49*, 15–21.

29. Chen, J.H., and Asch, S.M. (2017). Machine learning and prediction in medicine—beyond the peak of inflated expectations. N. Engl. J. Med. *376*, 2507.

30. Pearl, J. (2009). Causality (Cambridge University Press).

31. Hernán, M.A., Hsu, J., and Healy, B. (2019). A second chance to get causal inference right: a classification of data science tasks. Chance *32*, 42–49.

32. Gottesman, O., Johansson, F., Meier, J., Dent, J., Lee, D., Srinivasan, S., Zhang, L., Ding, Y., Wihl, D., Peng, X., et al. (2018). Evaluating reinforcement learning algorithms in observational health settings. Preprint at arXiv, 1805.12298.

33. Prosperi, M., Guo, Y., Sperrin, M., Koopman, J.S., Min, J.S., He, X., Rich, S., Wang, M., Buchan, I.E., and Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nat. Mach. Intell. *2*, 369–375.

34. Abràmoff, M.D., Lavin, P.T., Birch, M., Shah, N., and Folk, J.C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit. Med. *1*, 1–8.

35. Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science *366*, 447–453.

36. Adibi, A., Sadatsafavi, M., and Ioannidis, J. (2020). Validation and utility testing of clinical prediction models: time to change the approach. JAMA *324*, 235–236. https://doi.org/10.1001/jama.2020.1230.

37. Wessler, B.S., Nelson, J., Park, J.G., McGinnes, H., Gulati, G., Brazil, R., Van Calster, B., van Klaveren, D., Venema, E., Steyerberg, E., et al. (2021). External validations of cardiovascular clinical prediction models: a large-scale review of the literature. Circ. Cardiovasc. Qual. Outcomes, CIRCOUTCOMES121007858. https://doi.org/10.1161/CIRCOUTCOMES.121.007858.

38. Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, D.E., and Zou, J. (2021). How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat. Med. *27*, 582–584.

39. Topol, E.J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. *25*, 44–56.

40. Nagendran, M., Chen, Y., Lovejoy, C.A., Gordon, A.C., Komorowski, M., Harvey, H., Topol, E.J., Ioannidis, J.P., Collins, G.S., and Maruthappu, M. (2020). Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ *368*, m689.

41. Zhou, Q., Chen, Z.H., Cao, Y.H., and Peng, S. (2021). Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. NPJ Digit. Med. *4*, 1–12.

42. Nestor, B., McDermott, M.B., Boag, W., Berner, G., Naumann, T., Hughes, M.C., Goldenberg, A., and Ghassemi, M. (2019). Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. In Machine Learning for Healthcare Conference, PMLR, pp. 381–405.

43. Finlayson, S.G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., and Saria, S. (2020). The clinician and dataset shift in artificial intelligence. N. Engl. J. Med. *385*, 283–286.

44. Wong, A., Cao, J., Lyons, P.G., Dutta, S., Major, V.J., Ötleş, E., and Singh, K. (2021). Quantification of sepsis model alerts in 24 US hospitals before and during the COVID-19 pandemic. JAMA Netw. Open *4*, e2135286.

45. Obermeyer, Z., and Emanuel, E.J. (2016). Predicting the future - big data, machine learning, and clinical medicine. N. Engl. J. Med. *375*, 1216–1219, PMID: 27682033; PMCID: PMC5070532. https://doi.org/10.1056/NEJMp1606181.

46. Goldhahn, J., Rampton, V., and Spinas, G.A. (2018). Could artificial intelligence make doctors obsolete? BMJ *363*, k4563, PMID: 30404897. https://doi.org/10.1136/bmj.k4563.

47. Herper, M. (2017). MD Anderson Benches IBM Watson in setback for artificial intelligence in medicine, Forbes Feb. 19th at. https://www.forbes.com/sites/matthewherper/2017/02/19/md-anderson-benches-ibm-watson-in-setback-for-artificial-intelligence-in-medicine/?sh=e7023a377485.

48. Wynants, L., Van Calster, B., Collins, G.S., Riley, R.D., Heinze, G., Schuit, E., and van Smeden, M. (2020). Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ *369*, m1328.

49. O'Leary, L. (2022). How IBM's Watson went from the future of health care to sold off for parts. Slate. https://slate.com/technology/2022/01/ibm-watson-health-failure-artificial-intelligence.html.

50. Strickland, E. (2019). How IBM Watson overpromised and underdelivered on AI health care-IEEE spectrum. IEEE spectrum: technology, engineering, and science news. https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care.

51. Lee, J.D., and See, K.A. (2004). Trust in automation: designing for appropriate reliance. Hum. Factors *46*, 50–80.

52. Sieck, W.R., and Arkes, H.R. (2005). The recalcitrance of overconfidence and its contribution to decision aid neglect. J. Behav. Decis. Mak. *18*, 29–53.

53. Friedman, B., and Hendry, D.G. (2019). Value Sensitive Design: Shaping Technology with Moral Imagination (MIT Press).