PERSPECTIVE

Lessons Learned About Autonomous AI: Finding a Safe, Efficacious, and Ethical Path Through the Development Process

MICHAEL D. ABRÀMOFF, DANNY TOBEY, AND DANTON S. CHAR

Artificial intelligence (AI) describes systems capable of making decisions of high cognitive complexity; autonomous AI systems in healthcare are AI systems that make clinical decisions without human oversight. Such rigorously validated medical diagnostic AI systems hold great promise for improving access to care, increasing accuracy, and lowering cost, while enabling specialist physicians to provide the greatest value by managing and treating patients whose outcomes can be improved. Ensuring that autonomous AI provides these benefits requires evaluation of the autonomous AI's effect on patient outcome, design, validation, data usage, and accountability, from a bioethics and accountability perspective. We performed a literature review of bioethical principles for AI, and derived evaluation rules for autonomous AI, grounded in bioethical principles. The rules include patient outcome, validation, reference standard, design, data usage, and accountability for medical liability. Application of the rules explains successful US Food and Drug Administration (FDA) de novo authorization of an example, the first autonomous point-of-care diabetic retinopathy examination *de novo* authorized by the FDA, after a preregistered clinical trial. Physicians need to become competent in understanding the potential risks and benefits of autonomous AI, and understand its design, safety, efficacy and equity, validation, and liability, as well as how its data were obtained. The autonomous AI evaluation rules introduced here can help physicians understand limitations and risks as well as the potential benefits of autonomous AI for their patients. (Am J Ophthalmol 2020;■:■–■.

Supplemental Material available at AJO.com.
Accepted for publication Feb 22, 2020.

From the Department of Ophthalmology and Visual Sciences, University of Iowa, Iowa City, Iowa, USA (M.D.A.); IDx Technologies, Coralville, Iowa, USA (M.D.A.); DLA Piper, Dallas, Texas, USA (D.T.); and Department of Anesthesiology, Division of Pediatric Cardiac Anesthesia (D.S.C.) and Center for Biomedical Ethics (D.S.C.), Stanford University School of Medicine, San Francisco, California, USA.

Inquiries to Michael D. Abràmoff, Retina Service, University of Iowa Hospital and Clinics, 200 Hawkins Drive, Iowa City, IA 52242, USA; e-mail: michael-abramoff@uiowa.edu

AUTONOMOUS AI HAS THE POTENTIAL TO LESSEN PHYSICIAN BURDEN, INCREASE PATIENT ACCESS, AND LOWER COST

ARTIFICIAL INTELLIGENCE OR AUGMENTED INTELLIGENCE (AI) describes systems capable of making decisions of high cognitive complexity; autonomous AI systems in healthcare are AI systems that make such clinical decisions without human oversight, where the autonomous AI creator assumes medical liability.[1] For example, a diagnostic autonomous AI system for the point-of-care diagnosis of diabetic retinopathy provides a direct diagnostic recommendation without physician or human interpretation. Thus, it performs a cognitive, highly complex task that was previously only performed by ophthalmologists and optometrists—representing 0.02% of all Americans—after extensive, specialized training. Such rigorously validated medical diagnostic autonomous AI systems hold great promise for improving access to care, increasing accuracy, and lowering cost, while enabling specialist physicians to provide the greatest value by managing and treating those patients whose outcomes can be improved.[2,3] Ensuring that autonomous AI provides these benefits requires negotiating multiple ethical and practical challenges.

Recently, the first autonomous point-of-care diabetic retinopathy examination was de novo authorized by the US Food and Drug Administration (FDA), after a preregistered clinical trial, and became part of the American Diabetes Association's Standard of Diabetes Care.[4,5] No prior approval could serve as guidance, and there are significant ethical and legal concerns raised around introducing autonomous AI into healthcare.[6] We describe the

0002-9394
https://doi.org/10.1016/j.ajo.2020.02.022

concerns, go through the ethical and accountability principles we drew on to create evaluation rules for autonomous AI, and explain how we addressed them practically through the clinical trial and de novo authorized FDA clearance process and during ongoing implementation.

AUTONOMOUS AI EVOKES MANY PUBLIC CONCERNS

AS STATED, THERE ARE MANY POTENTIAL ADVANTAGES TO autonomous AI; but there are many concerns, as well, about the idea of "a computer making a diagnosis"—as is to be anticipated with any new technology. Thus, patients and physicians are concerned about whether the autonomous AI improves patient outcomes[3]; whether there is racial, ethnic, or other inappropriate bias[7]; whether patients' data are used appropriately[7]; or whether doctors will lose their jobs.[8]

> Will automation lower the quality of care? Or, in other words, will patients benefit from the use of autonomous AI, and will its use lead to better clinical outcomes?

To adequately address this concern, diagnostic AI must have well-defined and disease-specific indications for use, which then need to be rigorously validated in preregistered studies for safety, efficacy, and equity, involving real-world workflow. In 2007, Fenton and associates first demonstrated the importance of rigorous validation of AI in the actual workflow setting, rather than in a modeled laboratory setting.[9] In their pivotal study, the outcomes of women undergoing breast cancer screening by a radiologist assisted by a previously FDA-approved assistive AI system were compared to outcomes among women who underwent breast cancer screening by a radiologist *without* such an assistive AI. The assistive AI had previously been approved by the FDA on the basis of a study that showed that, when used in isolation, the assistive AI had high diagnostic accuracy compared to radiologists. When this assistive AI system was tested in the manner that reflected actual usage—where it assists a radiologist who makes the final clinical decision—the study showed worse outcomes for the women who underwent breast cancer screening with AI assistance. This finding and its implications are consistent with the FDA's larger trends toward real-world data and continuing assessment in the postmarket phase. Outside of healthcare, even more dramatic examples of the dangers of premature AI implementation have occurred: a recent report attributes the catastrophic events with AI-assisted air travel in part to a real-world impact of an AI system in workflow and the interaction with the pilot.[10] These rising concerns made clear that any diagnostic AI system, including autonomous AI, needs well-defined and disease-specific indications for use, which then needed to be able to be rigorously validated in preregistered studies for safety, efficacy, and equity, involving real-world workflow.

> Does the autonomous AI work equally well on the vast majority of patients, or does it exhibit inappropriate racial, ethnic, sex, or other bias?

There are now multiple examples of this concern about AI in general: a recent study showed that using medical cost as a proxy for patients' overall health needs led to inappropriate racial bias in allocating healthcare resources, as black patients were incorrectly deemed to have lower risk compared to white patients because their incurred costs were lower for a given health risk status.[11] Mitigating the risks of inappropriate bias needs to be addressed in the design, validation, and deployment of the autonomous AI, as we discuss below.

> Will doctors lose their jobs because of the introduction of autonomous AI?

As an illustration, 1 of the authors of this review (M.D.A.) received the nickname "the Retinator" in 2010, when McDonnell wrote a (somewhat tongue-in-cheek) editorial on the scientific research into autonomous AI for the diabetic retinopathy examination, stating that some ophthalmologists "disagree, resent, or even fear" autonomous AI for the diabetic eye examination for disease.[12] In consequence, the American Medical Association (AMA) has recently adopted the term "augmented intelligence"—of which autonomous AI is 1 subtype—highlighting the role of human physicians in interpreting and safeguarding the use of many forms of AI in healthcare.[1]

> Are patient-derived data used appropriately for both training and when deployed?

Typically, the development of any AI requires vast amounts of clinical data. There are many statutes and regulations covering patient-derived data, such as HIPAA and HITECH.[13] Ultimately, whether patient-derived data belong to the patient, the physician, the hospital system, or even whoever paid for acquisition has not been fully litigated, and as such can easily lead to concerns and controversy. For example, in 1 case patient data for AI training were obtained through an agreement with a health system.[14] While agreements were in place, patients and physicians were not aware of this data usage, leading to confusion, so that the Department of Health and Human Services became involved. In another example, a class action lawsuit alleging failure to adequately deidentify patient data for AI training was initiated against an academic health system.[15] Safeguarding patient data and using the data properly is clearly an important issue that reaches into the ethical considerations of the use of autonomous AI.

To realize the many potential advantages of autonomous AI, it is essential to address these and other concerns, ethical and otherwise, in an accountable and transparent fashion. If the concerns are not addressed appropriately,

TABLE 1. Autonomous AI Evaluation Rules

Evaluation Rules: Autonomous AI should be evaluated for the following	Classical Bioethical Principles and Accountability, Where Applicable[30]
Improve patient outcome, as shown by direct evidence-linked clinical literature, and aligned with evidence-based clinical standards of care/practice patterns from quality-of-care organizations, professional medical societies, and patient organizations, while accounting for safety, efficacy, and equity	Nonmaleficence Beneficence Justice
Design so the AI's operations are maximally reducible to characteristics aligned with scientific knowledge of human clinician cognition, rather than proxy characteristics	Beneficence
Maximize traceability of patient-derived data, and commensurate data stewardship, accountability, and authorization, including by adherence to accepted standards	Accountability Respect for autonomy
Validate rigorously for safety, efficacy, and equity, using preregistered clinical studies, by comparing the AI against clinical outcome (or outcome surrogates, in the case of chronic diseases) in the intended clinical workflow and usage, as shown by either direct or linked evidence	Nonmaleficence Justice
Align liability or other protections commensurate with indications for use and autonomy, without unduly burdening with liabilities beyond other comparable entities	Accountability Justice

the risk of a backlash on autonomous AI is real, as has been the case for other cutting-edge medical advances. After 2003, gene therapy effectively went through a moratorium on research funding, including closure of research institutions, after deaths and other complications from poorly overseen gene therapy studies came to light.[16] Only in 2017, almost 2 decades later, did the FDA approve the first-ever gene therapy to treat children and adults with the RPE65 variant of Leber congenital amaurosis.[17]

ETHICS-BASED DERIVATION OF PRELIMINARY AUTONOMOUS AI EVALUATION RULES FOR EVALUATING AUTONOMOUS AI

UNTIL NOW, THERE HAS BEEN A DEARTH OF ETHICAL EXAMination of actual AI systems for healthcare. In part, this lack of scrutiny is because much AI development has been occurring in private industry and has not yet been subject to multidisciplinary evaluation,[18] though studies of clinical AI systems are emerging and are likely to grow over the next few years. Solutions to ethical problems arising with AI are challenging without observation of actual applications and an understanding of the spectrum of issues arising with implementation. Evaluation of ethical problems arising with AI is also constrained by limited methodologies with which to examine ethical concerns.[19,20]

Since the potential benefits of autonomous AI risk being eclipsed by potential harms if ethical concerns are not addressed early in autonomous AI implementation to healthcare, what is a reasonable preliminary approach to ethical evaluation?

Although conceptual frameworks have been proposed to guide anticipatory ethical analyses of emerging technologies[19] or to ascertain the values inherent in design approaches,[21] there are no easily generated rules to follow for ethical human-computer interaction. These approaches are largely concerned with what constitutes a “value” or an “ethic”; ontological dilemmas over where such entities or actions might reside in people, technology, or their interaction; and questions of agency and intention in design.[21] In addition, they are limited by not encompassing the clinical contexts, regulatory and legal constraints, and healthcare economics needed to understand ethical ramifications of AI design or application choices in situ. Novel ethical challenges, unconsidered by classical bioethics approaches, which focus largely on a dyadic physician-patient relationship, are already emerging with implementations of AI to healthcare, including the enlarging responsibilities of learning healthcare systems and the capture of “big data” required to support AI approaches.[6,7]

In addition to the challenges with studying AI ethics, once identified, attempts at ethical guidance that do not engage with these multiple stakeholders are likely to be marginalized or ignored.[22] AI applications for healthcare can involve a multidisciplinary intersection of professional groups, including medicine, computer engineering, data science, regulatory and legal experts, and information technologists, who have often demonstrated skepticism about the usefulness of ethics teaching or codes of ethics to change behavior.[23–27] Moreover, these disparate disciplines may have competing interests, such as the tension between proprietary and confidential design in

TABLE 2. Aspects Requiring Consideration During the Design Stage of an Autonomous AI System, With Examples, for the Point-of-Care Diabetic Retinopathy Examination

Consideration	Example Application
The population on which it is to be used	Adults with diabetes without visual symptoms, with normal visual acuity, without known diabetic retinopathy
Narrow-use case. Potentially, multiple diseases could be designed to be diagnosed. The safety, efficacy, and equity would need validation for each of these diseases, requiring additional, equivalent clinical trials for each claimed diagnosis.	Diagnosis of diabetic retinopathy and macular edema only
The environment where it is to be used, including sensor hardware	In primary care without specific requirements the room, by operators with no previous expertise in retinal imaging and minimal training. This required the design of a robotic, easy-to-use retinal camera, coupled to an assistive AI, to help minimally proficient operators to take high-quality images
The diagnostic output. Outcome or surrogate outcome outputs allow linkage of diagnostic performance to studies of management and treatment.[31–34]	Positive output indicates ETDRS levels 35 and higher, or diabetic macular edema, or center-involved macular edema, and a negative output indicates the absence of these
Alignment with current clinical evidence and professional clinical standards	American Academy of Ophthalmology evidence-based guidelines on diabetic retinopathy management, as well as the American Diabetes Association Standard of Diabetes Care

the software space to reward innovation and deter piracy, vs an emphasis on regulatory and professional oversight, informed patient consent, and other transparencies in the medical field. Despite these and other ethical concerns, investment in AI for healthcare continues to rise, expecting to add $260 billion to healthcare by 2025.[28] Under the pressure of ongoing AI development, ethical guidance will need to be operationally relevant, and conducted and provided “on the fly,” both to not curtail innovation and to provide ethical resources to match the speed of current development. Ethical revisions to AI and evaluative frameworks will have to be iterative.

Many AI developers have already turned away from ethical analysis as unworkable or not adequately responsive to ongoing development,[29] and have instead begun to pursue an ideal of “algorithmic fairness,” or the ability to computationally demonstrate a lack of between-group bias with a machine learning application.[11,29] If latent biases can be identified, machine learning approaches might be used to correct for them or improve “fairness” .[29] However, such approaches assume a comprehensive, a priori, understanding of where and why such latent biases are occurring, or risk introducing a complex set of unintended and unforeseen biases, and second and third order effects, in attempts to correct the initial bias (Goodman 2018).

Drawing on a systematic review of ethical considerations that have so far been described for AI healthcare applications (Appendix; Supplemental Material available at AJO.com), we show in Table 1 an initial approach to create evaluation principles and accountability for autonomous AI systems in a healthcare context. Where applicable, we also note their alignment with classical bioethics principles, such as Beauchamp and Childress.[30] The rules in Table 1 are meant to be implemented practically, and have indeed been implemented in a regulatory process: some of them were quantified as clinical study endpoints. In the next sections we illustrate the principles by going through the design, validation, and deployment of an example, IDx-DR, the first autonomous AI system de novo authorized by the FDA.[4]

DESIGN OF THE AUTONOMOUS AI SYSTEM

DESIGN CONSIDERATIONS CAN HAVE UNEXPECTED AND profound ethical implications. For instance, in robotic surgery, systems that modify the tactile feedback a surgeon receives to reproduce a more “natural,” nonmechanical feel can bias the surgeon’s analysis of when to override the system, because the same AI algorithms driving the machine’s decision-making are biasing the information provided to the human monitor. In the example, the goal is to create a real-time point-of-care autonomous retinal examination for diabetic retinopathy and diabetic macular edema, available in the primary care office, that is safe, efficient, and equitable. The latter terms have recently been addressed in the AMA’s new policy on AI regulation and payment (2019 version),[1] and we discuss later what is meant more exactly by these terms. Center-involved macular edema has become such an important factor in visual loss in diabetes that we wanted to ensure that this was detectable in addition to the more classic diabetic retinopathy and clinically significant macular edema. Table 2 shows the most important aspects that required consideration during the design stage of the example autonomous AI.

DESIGN OF THE AUTONOMOUS AI DIAGNOSTIC ALGORITHM

AS FAR AS THE DIAGNOSTIC ALGORITHM IS CONCERNED, the autonomous AI evaluation principles require that physicians can understand how the autonomous AI system arrives at its clinical decision, not only to gain physician and patient trust, but also to improve AI safety. Inappropriate bias can result from incomplete or unrepresentative training data, and also from relying on complete and representative data that reflect and reproduce (at scale) preexisting bias. Using black-box or gray-box algorithm designs , where the inferences are not understoord by anyone, makes such bias harder to mitigate and detect, while the speed and scalability can multiply the effect of inappropriate bias faster than traditional enforcement efforts can react. In our example, for hundreds of years clinicians have evaluated a patient's retina for the different indicators of diabetic retinopathy, such as hemorrhages, microaneurysms, and neovascularization—indicators or biomarkers that are invariant with regard to race, ethnicity, sex, and age. Using multiple, statistically dependent detectors for such lesions,[35,36] each optimized using machine learning algorithms, addresses equity in the design phase.[37,38] Studies have shown that machine learning algorithms that align closely to the way clinicians diagnose are also more robust to small perturbations in the input, and show unexpected catastrophic failure, and are less likely to exhibit inappropriate racial and other bias[39,40]

The autonomous AI evaluation rules also require that, where possible, high-quality digital inputs and a corresponding high-validity disease state, called a reference standard or truth (such as patient outcome), are available, as well as widely accepted associations between the disease state and clinical management. For this example, in diabetes, decades of research is available regarding the diagnosis and management of diabetes and diabetic retinopathy through the Diabetes Control and Complications Trial (DCCT), the Epidemiology of Diabetes Interventions and Complications (EDIC) study, the Diabetic Retinopathy Study (DRS), the Early Treatment of Diabetic Retinopathy Study (ETDRS), and the Diabetic Retinopathy Clinical Research (DRCR) studies.[31,32,41,42]

AUTONOMOUS AI DATA STEWARDSHIP

THE AUTONOMOUS AI EVALUATION PRINCIPLES REQUIRE autonomous AI creators to be responsible stewards of patient data in order to design, develop, and monitor autonomous AI systems. Thus, autonomous AI creators have an obligation to lawfully collect data, in this case, in compliance with HIPAA/HITECH and other statutory and regulatory rules, in a manner that is transparent about the purpose and scope for which the data will be used.[13] Data used by the autonomous AI creator should be traceable to an authorization to use such data. Transparency on the part of autonomous AI creators, through written agreements, is essential to assess whether patients have adequately authorized use of data. Physicians and AI creators together are accountable directly to patients and each must take full responsibility for protecting patient rights as stewards of patient-derived data. Additionally, the rules require auditable processes and security controls to ensure that data are being used in accordance with the scope for which such use was authorized and to protect the data from unauthorized use or access. A current controversy is the desire of clinicians contributing the reference standard to patient-derived data to be rewarded or recognized for their contribution to the intellectual property of an AI system (eg, Paige.AI and Memorial Sloan Kettering) through their diagnostic work, recorded in medical records and subsequently used to train or evaluate an AI system. Such ownership collides with rising public desire for increased control over, and privacy regarding, electronic data and emerging regulations to address these (General Data Protection Regulation (EU) 2016/679 (GDPR)/California consumer privacy act AB 375), as well as increasing patient activism for inclusion in recognitions for specimen contribution to scientific advances.

CLINICAL VALIDATION OF SAFETY, EFFICACY, AND EQUITY OF AUTONOMOUS AI

THE EVALUATION RULES REQUIRE HIGH QUALITY AND rigor of autonomous AI validation studies for safety, efficacy, and equity, based on the ethical principles of nonmaleficence and justice (Table 1),[30] so as to ensure that safety and efficacy are equally valid for racial, ethnic, age, and sex subgroups, in relation to hypothesis testing study endpoints. In our example, safety was quantified using the sensitivity metric, which expresses how many patients with disease it diagnoses correctly, as a missed diagnosis can cause harm to the patient. A 100% sensitive AI can be created by always outputting "has disease" for every patient, but it would also be useless because its specificity would be 0%. Efficacy was quantified using the specificity metric, which expresses how many patients without disease it diagnoses correctly as not having the disease, since misdiagnosis of a patient without disease increases resource use without benefit to that patient. A 100% specific autonomous AI can be created by always outputting a "no disease" for every patient, but it would also be useless because its sensitivity would be 0%. Obviously, the challenge is to create an autonomous AI with the right balance between sensitivity and specificity given a particular use case. Equity was quantified with a combination of a diagnosability metric and a statistical analysis of subgroup validity. The diagnosability metric expresses how many patients receive a valid diagnostic result, rather than an indeterminate result. If only a small subset of patients with disease can be adequately diagnosed, then equity is diminished.

The hypothesis tested was that all 3 outcome parameters of sensitivity, specificity, and diagnosability exceed a preset threshold in the entire US population of people with diabetes, against surrogate outcome.[43] Earlier studies showed that experienced clinicians had 30%-40% sensitivity, 95% specificity, and 80%-90% diagnosability against this same surrogate outcome, and thus were unlikely to meet the 3 endpoints.[44,45] The autonomous AI exceeded all 3 superiority endpoints, at 87%, 91%, and 96% for sensitivity, specificity, and diagnosability, respectively. This confirmed the hypothesis of safety, efficacy, and equity. In addition, subgroup validity statistical analysis determines the amount of inappropriate diagnostic biases—as these can lead to clinical outcome disparities—and includes stratification of sensitivity, specificity, and diagnosability by race, ethnicity, sex, and age, as well as any other relevant group characteristics. In our example, safety, efficacy, and equity principles were implemented as 3 study endpoints of sensitivity, specificity, and diagnosability, with subgroup validity analysis, all of which had to be met to satisfy hypothesis testing.[43]

As we saw, the autonomous AI evaluation rules require comparing the autonomous AI to clinical outcome to estimate safety, efficacy, and equity. Given that clinical outcome is central to patient benefit of the autonomous AI, outcome or surrogate outcome (in the case of chronic diseases, where actual outcome only appears sometimes decades later) are clearly optimal. For our example, many foundational studies for diabetic retinopathy treatment and management performed over the past 5 decades were available, so that the so-called ETDRS severity scale, as well as the DRCR's center-involved macular edema scale, were available as robust surrogate outcomes.[31,32,41,42] It is obviously of great importance that surrogate outcomes are stable over time as well as consistent, as multiple studies have shown for our example.[33,46] Such surrogate outcomes are fundamentally different from clinicians' agreement with the AI or even each other: rarely has the diagnostic performance of such individual clinicians been validated against outcome. In many cases, when initiating development of (autonomous) AI, the typical process involves a single clinician evaluating a patient image for clinical purposes, and there are no data on how their evaluation relates to a clinical outcome. Where widely accepted clinical outcome surrogates are not available for chronic diseases, as is for example the case for glaucoma, they should be established to determine the safety, efficacy, and equity, commensurate with patient risk of harm.

The autonomous AI evaluation rules require preregistered studies, which is consistent with US federal regulation. Without preregistration, autonomous AI performance tends to be overestimated, while successful study replication becomes less likely: in fact when comparing trials with and without preregistration, the trial effect sizes are larger when lacking preregistration.[43,47] Other considerations for correct validation related to good clinical practice[4] include a hypothesis-testing design with predefined endpoints, a predefined method for statistical analysis, predefined inclusion and exclusion criteria, a predefined sampling protocol, a plan for handling of the trial data by an independent contract research organization or third party, and prohibition of access by the researchers to the subject-level results before finalizing the statistical analysis.

The autonomous AI evaluation rules require validation in the envisioned context, environment, and workflow, in "locked" form, so that its performance is known and persists in clinical practice—thus avoiding the unanticipated effect of AI in the Fenton study.[9] Such a locked autonomous AI, once validated, cannot be automatically updated based on new inputs, as then the safety efficacy and equity are not known. This precludes, for now, "continuous learning" AI systems that automatically update as they process new inputs, in chronic disease. In our example, this required the trial to be performed in primary care clinics, in the standard diabetes management workflow, without modifications to the clinic environment, and with operators recruited from existing staff without prior experience or training.

LEGAL CONSIDERATIONS OF AUTONOMOUS AI

AUTONOMOUS AI—EVEN WHEN EMBEDDED IN THE HEALTHcare system and Medical Home models—introduces several accountability and medicolegal considerations. While these are not ethical considerations per se, it is worthwhile to evaluate them in this context.

The autonomous AI evaluation principles require creators of autonomous AI to assume liability for harm caused by the diagnostic output of the device when used properly and on-label. This is essential for adoption—it is inappropriate for clinicians, using an autonomous AI to make a diagnosis they are not comfortable making themselves, to have full medical liability for harm caused by that AI. This view was recently endorsed by the AMA in its 2019 AI Policy.[1] Just like a physician that would be held legally responsible for his or her diagnosis or other clinical decision, IDx, as creators of autonomous AI products, assume similar liability and have obtained medical malpractice insurance. This paradigm shifts medical liability for a medical diagnostic from the provider managing the patient's diabetes, who orders the autonomous point of care retinal examination, to the autonomous AI creator.

However, medical decisions by autonomous AI on *individual patients* typically cannot be unequivocally labeled as correct or incorrect, especially in chronic diseases where outcomes may emerge years later. On *populations of patients*, however, the medical decisions can be compared statistically to the desired decisions, for example to claimed correctness, and it is thus where the liability will be focused. Another issue is that, while autonomous AI is preferably compared to patient outcome or surrogate outcome, this requires enormous resources that will not be available for the individual patient where liability is at stake. Then, the autonomous AI decision will be compared to an individual physician or group of physicians, lacking validation and, thus, with unknown correspondence to outcome or surrogate outcome. As an aside, this can be an issue also for so-called continuous learning AI systems.

These distinctions will need to be resolved as various AI applications move forward. The legal responsibility for an AI system built in partnership with a large learning healthcare system and intended to be used on its patient population is, by definition, more diffuse and likely to vest in the sponsoring healthcare system or with some comparative or contributory analysis of fault. A privately designed system, sold as a finished product, will need to bear its own responsibility for autonomous output, absent superseding or intervening causation.

What is clear is that the bedside provider using the AI output does not have the proficiency to make the particular clinical decision for which they need the AI. Responsibility for proper use and maintenance of the device, consistent with terms of service and FDA labeling, remains with the providers. This situation is different, however, for assistive AI, where a physician is able to make an independent evaluation of the AI system's recommendation output and remains fully liable.

The output of the autonomous AI system, though valid as a diagnostic record from a regulatory perspective, is not currently defined as a medical record when it is not signed off on by a physician. What is and is not, and who can and cannot create, a medical record is determined primarily by the state medical boards. At present, most state medical boards do not consider an autonomous AI output to have the same medicolegal status as physician documentation. The legal status of reports generated by AI has been brought to the attention of the Federation of State Medical Boards.

SUMMARY

THE AUTONOMOUS AI EVALUATION RULES DISCUSSED IN this article both implicitly and explicitly played a role in ongoing discussions with the FDA and the medical and legal community, to establish the safety, efficacy, and equity of autonomous AI. It led to the first preregistered, prospective clinical trial—the standard in FDA drug trials—and *de novo* FDA authorization of an autonomous AI system, in a real-world clinical setting, with primary care–based operators. The surrogate outcome reference standard to which the autonomous AI performance was compared was derived from an independent reading center with validated published protocols and reproducibility and repeatability metrics. The trial showed that autonomous AI exceeded the 3 prespecified superiority endpoint goals of sensitivity, specificity, and diagnosability, and that there was no significant effect of race or ethnicity on these 3 endpoints.

As physicians, we are likely to be asked more and more to evaluate the clinical value and scientific evidence for autonomous AI, just as we are now asked for novel drug treatments. Thus, physicians need to become competent in understanding the limitations and risks as well as the potential benefits of autonomous AI, and in understanding the design; its liability; its safety, efficacy, and equity; and how its data were obtained. The autonomous AI evaluation rules introduced here may support this process.

FUNDING/SUPPORT: M.D.A.: IDX, CORALVILLE, IOWA, USA; RESEARCH TO PREVENT BLINDNESS, NEW YORK, NEW YORK, USA; National Eye Institute, Bethesda, Maryland, USA, R01 EY016822; The Robert C. Watzke Professorship for Retina Research. D.S.C.: National Human Genome Research Institute of the National Institutes of Health K01HG008498. Financial Disclosures: Michael D. Abramoff: Investor, Director, Employee of IDx Technologies, Coralville, Iowa, USA; patents and patent applications assigned to the University of Iowa as well as IDx that are relevant to the subject matter of this manuscript. Danny Tobey and Danton S. Char have no financial disclosures.

REFERENCES

1. American Medical Association (AMA). Augmented intelligence in healthcare 2019. AMA Board of Trustees Policy Summary. https://www.ama-assn.org/system/files/2019-08/ai-2018-board-policy-summary.pdf. Accessed April 14, 2020.
2. Helmchen LA, Lehmann HP, Abramoff MD. Automated detection of retinal disease. *Am J Manag Care* 2014;20(11 Spec No. 17):eSP48–eSP52.
3. Centers for Medicare and Medicaid Services. Artificial Intelligence (AI) Health Outcomes Challenge 2019; https://innovation.cms.gov/initiatives/artificial-intelligence-health-outcomes-challenge/. Accessed April 14, 2020.
4. US Food and Drug Administration. FDA Permits Marketing of Artificial Intelligence-Based Device to Detect Certain Diabetes-Related Eye Problems 2018. Washington, DC. Available at: https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm604357.htm. Accessed April 14, 2020.

5. American Diabetes Association. 11. Microvascular Complications and Foot Care: Standards of Medical Care in Diabetes-2020. *Diabetes Care* 2020;43(Suppl 1):S135–S151.
6. Char DS, Shah NH, Magnus D. Implementing machine learning in health care-addressing ethical challenges. *N Engl J Med* 2018;378(11):981–983.
7. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf* 2019;28(3):231–237.
8. Sarwar S, Dent A, Faust K, et al. Physician perspectives on integration of artificial intelligence into diagnostic pathology. *NPJ Digit Med* 2019;2:28.
9. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med* 2007;356(14):1399–1409.
10. Langewiesche W. What really brought down the Boeing 737 Max 2019. New York, NY: NYT Magazine. https://www.nytimes.com/2019/09/18/magazine/boeing-737-max-crashes.html. Accessed April 14, 2020.
11. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366(6464):447–453.
12. McDonnell PJ. 'The Retinator': revenge of the machines. *Ophthalmol Times* 2010;35(13):4.
13. Blumenthal D. Launching HITECH. *N Engl J Med* 2010;362(5):382–385.
14. Copeland R, Needleman S. Google's "Project Nightingale" triggers federal inquiry 2019. New York, NY: Wall Street Journal. https://www.wsj.com/articles/behind-googles-project-nightingale-a-health-data-gold-mine-of-50-million-patients-11573571867. Accessed April 14, 2020.
15. Dinerstein v. Google LLC et al.. Available at ; 2019. https://dockets.justia.com/docket/illinois/ilndce/1:2019cv04311/366172; Accessed April 14, 2020.
16. Chandler RJ, Venditti CP. Gene therapy for metabolic diseases. *Transl Sci Rare Dis* 2016;1(1):73–89.
17. Russell S, Bennett J, Wellman JA, et al. Efficacy and safety of voretigene neparvovec (AAV2-hRPE65v2) in patients with RPE65-mediated inherited retinal dystrophy: a randomised, controlled, open-label, phase 3 trial. *Lancet* 2017;390(10097):849–860.
18. O'Neil C. The ivory tower can't keep ignoring tech 2017. New York, NY: The New York Times. https://www.nytimes.com/2017/11/14/opinion/academia-tech-algorithms.html. Accessed April 14, 2020.
19. Brey P. Anticipatory ethics for emerging technologies. *Nanoethics* 2012;6:1–13.
20. Shilton K. Values levers: building ethics into design. *Sci Technol Human Values* 2013;38(3):374–397.
21. Shilton K. Values and ethics in human-computer interaction. *Found Trends Human Comput Interact* 2018;12(2):107–171.
22. AI, Ethics & Society @ Yale. Conference April 9, 2019. Materials accessed online 29 April 2019 at https://aiethicsyale.wordpress.com/. Accessed April 14, 2020.
23. McGinn R. The Ethical Engineer: Contemporary Concepts and Cases. Princeton, NJ: Princeton University Press; 2018.
24. Murrell V. The failure of medical education to develop moral reasoning in medical students. *Int J Med Educ* 2014;27(5):219–225.
25. Carrese J, Malek J, Watson K, et al. The essential role of medical ethics education in achieving professionalism: the Romanell Report. *Acad Med* 2015;90(6):744–752.
26. Fleischmann K, Hui C, Wallace W. The societal responsibilities of computational modelers: human values and professional codes of ethics. *J Assoc Inf Sci Technol* 2017;68:543–552.
27. IEEE. Code of Ethics. Available at ; 2015. https://www.ieee.org/about/corporate/governance/p7-8.html;. Accessed April 14, 2020.
28. McKinsey Global Institute. Featured Insights. VIsualizing the uses and potential impact of AI and other analytics. 2018; https://www.mckinsey.com/featured-insights/artificial-intelligence/visualizing-the-uses-and-potential-impact-of-ai-and-other-analytics. Accessed April 14, 2020.
29. Rajkomar A, Hardt M, Howell MD, et al. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018;169(12):866–872.
30. Beauchamp TL, Childress JF. Principles of Biomedical Ethics. 8th ed. New York: Oxford University Press; 2019.
31. Fundus photographic risk factors for progression of diabetic retinopathy. ETDRS report number 12. Early Treatment Diabetic Retinopathy Study Research Group. *Ophthalmology* 1991;98(5 Suppl):823–833.
32. Diabetic Retinopathy Clinical Research Network, Beck RW, Edwards AR, et al. Three-year follow-up of a randomized trial comparing focal/grid photocoagulation and intravitreal triamcinolone for diabetic macular edema. *Arch Ophthalmol* 2009;127(3):245–251.
33. Grading diabetic retinopathy from stereoscopic color fundus photographs–an extension of the modified Airlie House classification. ETDRS report number 10. Early Treatment Diabetic Retinopathy Study Research Group. *Ophthalmology* 1991;98(5 Suppl):786–806.
34. Photocoagulation for diabetic macular edema. Early Treatment Diabetic Retinopathy Study report number 1. Early Treatment Diabetic Retinopathy Study research group. *Arch Ophthalmol* 1985;103(12):1796–1806.
35. Hubel DH, Wiesel TN. Receptive fields of single neurones in the cat's striate cortex. *J Physiol* 1959;148:574–591.
36. Ts'o DY, Frostig RD, Lieke EE, Grinvald A. Functional organization of primate visual cortex revealed by high resolution optical imaging. *Science* 1990;249(4967):417–420.
37. Friedenwald J, Day R. The vascular lesions of diabetic retinopathy. *Bull Johns Hopkins Hosp* 1950;86(4):253–254.
38. MacKenzie S. A Case of Glycosuric Retinitis, with Comments. (Microscopical Examination of the Eyes by Mr. Nettleship). *Roy London Ophthal Hosp Rep* 1879;9(134).
39. Shah A, Lynch S, Niemeijer M, et al. "Susceptibility to misdiagnosis of adversarial images by deep learning based retinal image analysis algorithms". Washington, DC: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018); 2018:1454–1457.
40. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science* 2019;363(6433):1287–1289.
41. Diabetes Control and Complications Trial Research Group. Progression of retinopathy with intensive versus conventional treatment in the Diabetes Control and Complications Trial. *Ophthalmology* 1995;102(4):647–661.

42. Browning DJ, Glassman AR, Aiello LP, et al. Optical coherence tomography measurements and analysis methods in optical coherence tomography studies of diabetic macular edema. *Ophthalmology* 2008;115(8):1366–1371. 1371.e1361.
43. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med* 2018;1(1):39.
44. Pugh JA, Jacobson JM, Van Heuven WA, et al. Screening for diabetic retinopathy. The wide-angle retinal camera. *Diabetes Care* 1993;16(6):889–895.
45. Lin DY, Blumenkranz MS, Brothers RJ, Grosvenor DM. The sensitivity and specificity of single-field nonmydriatic monochromatic digital fundus photography with remote image interpretation for diabetic retinopathy screening: a comparison with ophthalmoscopy and standardized mydriatic color photography. *Am J Ophthalmol* 2002;134(2):204–213.
46. Scott IU, Bressler NM, Bressler SB, et al. Agreement between clinician and reading center gradings of diabetic retinopathy severity level at baseline in a phase 2 study of intravitreal bevacizumab for diabetic macular edema. *Retina* 2008;28(1):36–40.
47. Kaplan RM, Irvin VL. Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLoS One* 2015;10(8):e0132382.