THE AMERICAN JOURNAL OF BIOETHICS
2020, VOL. 20, NO. 11, 7–17
https://doi.org/10.1080/15265161.2020.1819469

Taylor & Francis
Taylor & Francis Group

TARGET ARTICLE

Identifying Ethical Considerations for Machine Learning Healthcare Applications

Danton S. Char[a], Michael D. Abràmoff[b,c], and Chris Feudtner[d,e]

[a]Stanford University School of Medicine; [b]University of Iowa; [c]Digital Diagnostics; [d]The University of Pennsylvania; [e]The Children's Hospital of Philadelphia

ABSTRACT

Along with potential benefits to healthcare delivery, machine learning healthcare applications (ML-HCAs) raise a number of ethical concerns. Ethical evaluations of ML-HCAs will need to structure the overall problem of evaluating these technologies, especially for a diverse group of stakeholders. This paper outlines a systematic approach to identifying ML-HCA ethical concerns, starting with a conceptual model of the pipeline of the conception, development, implementation of ML-HCAs, and the parallel pipeline of evaluation and oversight tasks at each stage. Over this model, we layer key questions that raise value-based issues, along with ethical considerations identified in large part by a literature review, but also identifying some ethical considerations that have yet to receive attention. This pipeline model framework will be useful for systematic ethical appraisals of ML-HCA from development through implementation, and for interdisciplinary collaboration of diverse stakeholders that will be required to understand and subsequently manage the ethical implications of ML-HCAs.

KEYWORDS

Artificial intelligence; effectiveness; ethics; machine learning; safety; test characteristics

> There is an old saying that a problem well put is half solved. This much is obvious. What is not so obvious, however, is how to put a problem well.
>
> Churchman, Ackoff, Arnoff
>
> Introduction to Operations Research, 1957, page 67.

With the FDA authorization of an autonomous artificial intelligence diagnostic system based on machine learning (ML), which employs algorithms that can learn from large data sets and make predictions without being explicitly programmed, ML healthcare applications (ML-HCAs) have transitioned from being an enticing future possibility to a present clinical reality (Abràmoff et al. 2018; Commissioner Office of the FDA 2020). Almost certainly, ML-HCAs will have a substantial impact on healthcare processes, quality, cost, and access, and in so doing will raise specific and perhaps unique ethical considerations and concerns in the healthcare context (Obermeyer and Emanuel 2016; (Rajkomar et al. 2019; Maddox et al. 2019; Matheny et al. 2019, 2020). This has been the case in non-healthcare contexts (Char et al. 2018; Bostrom and Yudkowski 2011), where ML implementation has generated toughening scrutiny due to scandals regarding how large repositories of private data have been sold and used (Rosenberg and Frenkel 2018), how the ML design of algorithmic flight controls resulted in accidents (Nicas et al. 2019), and how computer-assisted prison sentencing guidelines perpetuate racial bias (Angwin et al. 2016), to name but a few of the growing number of examples. Regarding specifically ML-HCAs, our review of the literature (see appendix for review methods) identified a variety of ethical considerations and concerns that have been cited, such as bias arising from the training data set (Challen et al. 2019), the privacy of personal data in business arrangements (Comfort 2016; Hern 2017), ownership of the data used to train ML-HCAs (Ornstein and Thomas 2018) and accountability for ML-HCA's failings (Ross and Swelitz 2017).

Notably, no systematic approach has yet emerged regarding how to survey the landscape of ML-HCA conception, development, calibration, implementation, evaluation, and oversight. Berefit of any conceptual map of this landscape, the identification of ethical concerns arising from this emerging, complex, cross-disciplinary technology that potentially affects many aspects of healthcare has thus far been reactive, ad

CONTACT Danton S. Char dchar@stanford.edu Department of Anesthesiology, Stanford University Medical Center, 300 Pasteur Drive, Stanford, CA 94305, USA.

hoc, and fragmented. This is problematic, especially for so-called "wicked" problems, which unlike more straightforward and "tame" technical problems, typically defy a singular formulation of the problem, are nested within systems that have interrelated problems, and have social values woven into their fabric such that solutions are not simply true or false but rather better or worse (Rittel and Webber 1973). In such circumstances, problem solvers are better served by approaches that enable taking a step back at the outset to assure that the problem is as "well put" (Churchman et al. 1957) as possible. Although this fundamental step for the analysis of any problem is often overlooked, a variety of problem structuring methods exist (Rosenhead and Mingers 2008). A common attribute across these methods is creating and clarifying (ideally with a diverse group of stakeholders) a shared conceptual mental map of the problem, which often evolves over time. Equipped with such a map, problem solvers may identify more decisions and their interconnected consequences, which in turn may advance value-focused thinking (Keeney 1992) and improve ethical decision-making (Stenmark et al. 2011).

In this paper, we aim to enhance our ability to identify—proactively, systematically, and in a more thoroughgoing and integrated manner—the variety of ethically relevant decisions and their ethically relevant consequences regarding ML-HCA. Specifically, we propose framing this problem of identifying ethical issues as occurring within and across the entire pipeline of activities that comprise the development, implementation, and ongoing evaluation of any ML-HCA (Figure 1, top 2 rows). Onto this conceptual structure can be mapped an overlay of questions that raise values-based issues and ethical considerations (Figure 1, bottom 2 rows). This pipeline schematic can serve as overview map not only to help us spot novel ethical concerns, but also to recognize familiar ethical considerations of healthcare technology and interventions, such as promoting benefit while protecting against harm, clarifying the values that are inexorably built into test calibration cut-points, and ensuring benefit and burdens are equitably distributed across populations of individuals.

We should raise three caveats before proceeding. First, our pipeline framework, and in particular our mapping from the ML-HCA process to sets of ethical considerations, is undoubtedly incomplete. More work will need to be done (as mentioned below) by diverse stakeholders to flesh out this framework and mapping. Indeed, by laying out an overview as we do, gaps are likely to stand out. In no small sense, this is one of the prime values of the approach we propose. Second, this framework does not address the issue of who should be responsible for what, but instead is intended to help anyone who wishes (or is required) to be ethically thoughtful to do so in a more systematic manner. Third, our chief goal is how to identify ML-HCA ethical concerns and considerations. This is necessary but not sufficient. A subsequent process of evaluating these considerations and confronting the likely tradeoffs to resolve them is needed, which we will not be emphasizing. These subsequent tradeoff decisions will always require detailed content and context-specific knowledge. Nevertheless, such decisions would be flawed if the broader process of first identifying the range of relevant considerations is not thorough.

CHALLENGES TO IDENTIFYING ETHICAL CONSIDERATIONS

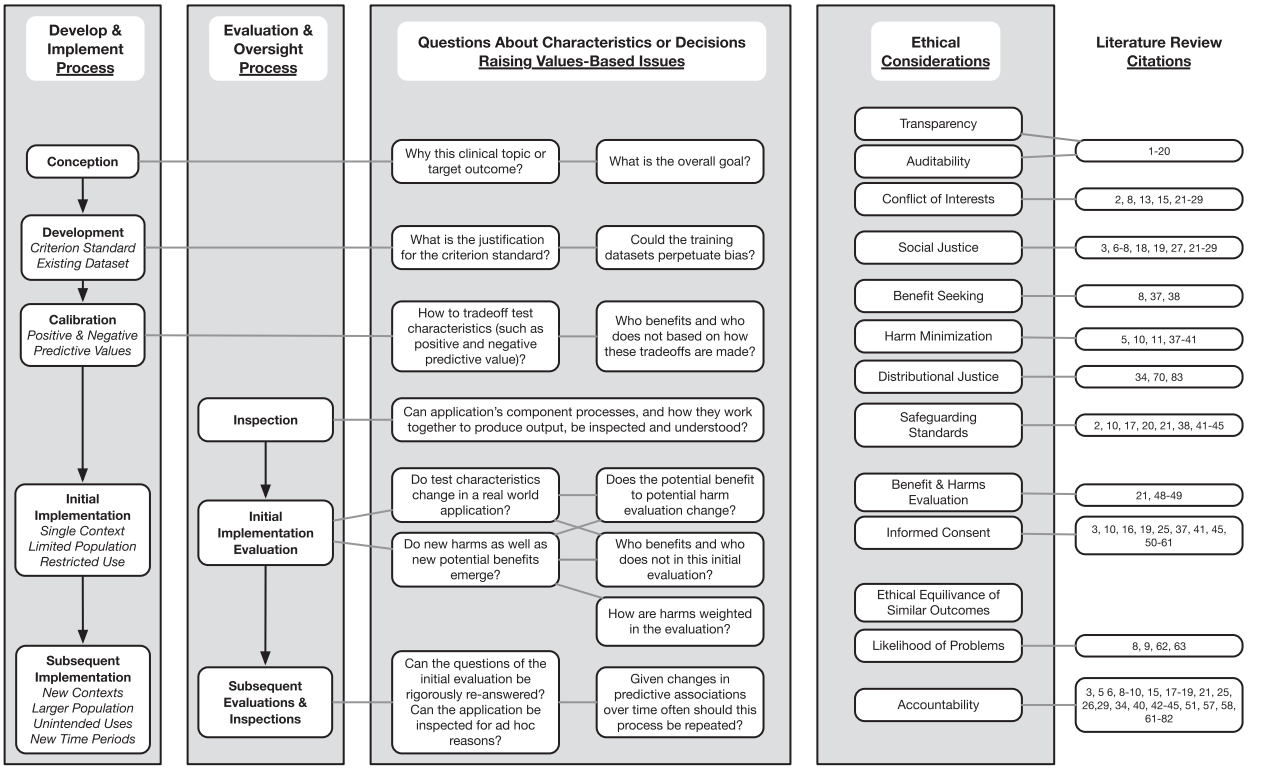
Before laying out the pipeline model, we need to clarify five significant challenges to identifying ethical considerations arising from ML-HCAs design, implementation, and evaluations, as any approach to the identification task should be designed to meet these challenges.

Uncertain Impact of Emerging Technologies

ML-HCAs, like all new technologies, present uncertainty regarding their future impact. Ethical frameworks that focus on articulating guiding principles without first systematically identifying potential problems (Challen et al. 2019; Matheny et al. 2019, 2020) do not specifically address this uncertainty. While various conceptual frameworks have been proposed to guide anticipatory ethical analyses of emerging technologies (Brey 2012) or to ascertain the values inherent in design approaches (Shilton 2018), a common general feature of these methods is the importance of having a systematic approach guided by an underlying evaluative framework to identify key considerations across as full a range as is possible of potential impacts. This feature does not reduce the uncertainty per se, but represents a strategy to manage it by casting a broad and thorough net.

Machine Learning and Artificial Intelligence Exceptionalism

As advanced as ML-HCAs are, built with cutting edge technology, no sound reason as yet exists to believe



Citation References

1. Avati, A. *et al.* Improving palliative care with deep learning. *BMC Med. Inform. Decis. Mak.* **18**, 122 (2018).
2. Berner ES. Ethical and legal issues in the use of clinical decision support systems. *J. Healthc. Inf. Manag. JHIM* **16**, 34–7 (2002).
3. Cohen, I. G., Amarasingham, R., Shah, A., Xie, B. & Lo, B. The Legal And Ethical Concerns That Arise From Using Complex Predictive Analytics In Health Care. *Health Aff. (Millwood)* **33**, 1139–1147 (2014).
4. Feudtner, C., Schall, T., Nathanson, P. & Berry, J. Ethical Framework for Risk Stratification and Mitigation Programs for Children With Medical Complexity. *Pediatrics* **141**, S250–S258 (2018).
5. Fiske A., Henningsen P. & Buyx A. Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy. *J. Med. Internet Res.* **21**, e13216 (2019).
6. Ho, A. Deep Ethical Learning: Taking the Interplay of Human and Artificial Intelligence Seriously. *Hastings Cent. Rep.* **49**, 36–39 (2019).
7. Kulikowski, C. A. Beginnings of Artificial Intelligence in Medicine (AIM): Computational Artifice Assisting Scientific Inquiry and Clinical Art – with Reflections on Present AIM Challenges. *Yearb. Med. Inform.* **28**, 249–256 (2019).
8. Liyanage, H. *et al.* Artificial Intelligence in Primary Health Care: Perceptions, Issues, and Challenges: Primary Health Care Informatics Working Group Contribution to the Yearbook of Medical Informatics 2019. *Yearb. Med. Inform.* **28**, 041–046 (2019).
9. London, A. J. Artificial Intelligence and Black-Box Medical Decisions: *Accuracy versus Explainability*. *Hastings Cent. Rep.* **49**, 15–21 (2019).
10. O'Sullivan, S. *et al.* Anatomy 101 for AI-driven robotics: Explanatory, ethical and legal frameworks for development of cadaveric skills training standards in autonomous robotic surgery/autopsy. *Int. J. Med. Robot.* e2020 (2019). doi:10.1002/rcs.2020
11. Peirce AG. *et al.* 'Knowledge development, technology and questions of nursing ethics'. *Nurs. Ethics* 969733019840752 (2019). doi:10.1177/0969733019840752
12. Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W. & Wallach, H. Manipulating and Measuring Model Interpretability. *ArXiv180207810 Cs* (2018).
13. Powles, J. & Hodson, H. Google DeepMind and healthcare in an age of algorithms. *Health Technol.* **7**, 351–367 (2017).
14. Price, W. N. Medical Malpractice and Black-Box Medicine. in *Big Data, Health Law, and Bioethics* (eds. Cohen, I. G., Lynch, H. F., Vayena, E. & Gasser, U.) 295–306 (Cambridge University Press, 2018). doi: 10.1017/9781108147972.027
15. Shaw, J., Rudzicz, F., Jamieson, T. & Goldfarb, A. Artificial Intelligence and the Implementation Challenge. *J. Med. Internet Res.* **21**, e13659 (2019).
16. Springer, A. Making Transparency Clear. *Los Angel.* 5 (2019).
17. Taljaard M. *et al.* Cardiovascular Disease Population Risk Tool (CVDPoRT): predictive algorithm for assessing CVD risk in the community setting. A study protocol. *BMJ Open* **4**, e006701 (2014).
18. Vayena, E. & Blasimme, A. Health Research with Big Data: Time for Systemic Oversight. *J. Law. Med. Ethics* **46**, 119–129 (2018).
19. Vellido, A. Societal Issues Concerning the Application of Artificial Intelligence in Medicine. *Kidney Dis.* **5**, 11–17 (2019).
20. Watson DS. *et al.* Clinical applications of machine learning algorithms: beyond the black box. *BMJ* **364**, l886 (2019).
21. Amarasingham R. *et al.* Consensus Statement on Electronic Health Predictive Analytics: A Guiding Framework to Address Challenges. *EGEMS Wash. DC* **4**, 1163 (2016).
22. Banja, J. Welcoming the "Intel-ethicist". *Hastings Cent. Rep.* **49**, 33–36 (2019).
23. Bauer M. *et al.* Ethical perspectives on recommending digital technology for patients with mental illness. *Int. J. Bipolar Disord.* **5**, 6 (2017).
24. Hagendorff, T. The Ethics of AI Ethics -- An Evaluation of Guidelines. *ArXiv190303425 Cs Stat* (2019).
25. Jaremko JL. *et al.* Canadian Association of Radiologists White Paper on Ethical and Legal Issues Related to Artificial Intelligence in Radiology. *Can. Assoc. Radiol. J. J. Assoc. Can. Radiol.* **70**, 107–118 (2019).
26. Kohli M., Prevedello LM., Filice RW. & Geis JR. Implementing Machine Learning in Radiology Practice and Research. *AJR Am. J. Roentgenol.* **208**, 754–760 (2017).
27. Meskó B., Hetényi G. & Győrffy Z. Will artificial intelligence solve the human resource crisis in healthcare? *BMC Health Serv. Res.* **18**, 545 (2018).
28. SFR-IA Group. Artificial intelligence and medical imaging 2018: French Radiology Community white paper. *Diagn. Interv. Imaging* **99**, 727–742 (2018).
29. Zuboff, S. *Big Other: Surveillance Capitalism and the Prospects of an Information Civilization*. (Social Science Research Network, 2015).
30. Bauer, M. *et al.* Ethical perspectives on recommending digital technology for patients with mental illness. *Int. J. Bipolar Disord.* **5**, 6 (2017).
31. Feuerstein G. Allocation by computer: ethical problems of prognostic scoring systems. *Biomed. Ethics Tubingen Ger.* **4**, 56–8 (1999).
32. Fisher, S. *et al.* Dementia Population Risk Tool (DemPoRT): study protocol for a predictive algorithm assessing dementia risk in the community. *BMJ Open* **7**, e018018 (2017).
33. Flahault A. *et al.* Precision global health in the digital age. *Swiss Med. Wkly.* **147**, w14423 (2017).
34. Ho CWL., Soon D., Caals K. & Kapur J. Governance of automated image analysis and artificial intelligence analytics in healthcare. *Clin. Radiol.* **74**, 329–337 (2019).
35. Kuziemsky C. *et al.* Role of Artificial Intelligence within the Telehealth Domain. *Yearb. Med. Inform.* (2019). doi:10.1055/s-0039-1677897
36. Mittelstadt, B., Russell, C. & Wachter, S. Explaining Explanations in AI. in *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19* 279–288 (ACM Press, 2019). doi:10.1145/3287560.3287574
37. Suominen H. *et al.* Applying language technology to nursing documents: pros and cons with a focus on ethics. *Int. J. Med. Inf.* **76 Suppl 2**, S293–301 (2007).
38. Nebeker, C., Torous, J. & Bartlett Ellis, R. J. Building the case for actionable ethics in digital health research supported by artificial intelligence. *BMC Med.* **17**, 137 (2019).
39. McMinn MR., Ellens BM. & Soref E. Ethical perspectives and practice behaviors involving computer-based test interpretation. *Assessment* **6**, 71–8 (1999).
40. Nadler, S., Raz, O. & Zalmanovici, M. Governance and Regulations Implications on Machine Learning. 15
41. Nabi J. How Bioethics Can Shape Artificial Intelligence and Machine Learning. *Hastings Cent. Rep.* **48**, 10–13 (2018).
42. Artificial Intelligence Standards. *Federal Register* (2019). Available at: https://www.federalregister.gov/documents/2019/05/01/2019-08818/artificial-intelligence-standards. (Accessed: 22nd July 2019)
43. Bezemer T. *et al.* A Human(e) Factor in Clinical Decision Support Systems. *J. Med. Internet Res.* **21**, e11732 (2019).
44. Moscatelli M. *et al.* An infrastructure for precision medicine through analysis of big data. *BMC Bioinformatics* **19**, 351 (2018).
45. Pesapane F., Volonté C., Codari M. & Sardanelli F. Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. *Insights Imaging* **9**, 745–753 (2018).
46. Brown GS. & Kornmayer K. Expert systems restructure managed care practice: implementation and ethics. *Behav. Healthc. Tomorrow* **5**, 31–4 (1996).
47. Stahl, B. C. & Coeckelbergh, M. Ethics of healthcare robotics: Towards responsible research and innovation. *Robot. Auton. Syst.* **86**, 152–161 (2016).
48. Di Nucci E. Should we be afraid of medical AI? *J. Med. Ethics* (2019). doi:10.1136/medethics-2018-105281
49. McDougall, R. J. Computer knows best? The need for value-flexibility in medical AI. *J. Med. Ethics* **45**, 156–160 (2019).
50. Citizens Juries | Connected Health Cities | Health Data Trial. *Connected Health Cities* Available at: https://www.connectedhealthcities.org/chc-hub/public-engagement/citizens-juries-chc/citizens-juries/. (Accessed: 26th July 2019)
51. Fischer T., Brothers KB., Erdmann P. & Langanke M. Clinical decision-making and secondary findings in systems medicine. *BMC Med. Ethics* **17**, 32 (2016).
52. Grssun D. [Ethics and artificial intelligence in healthcare, towards positive regulation]. *Soins Rev. Ref. Infirm.* **64**, 54–57
53. Henshall C. *et al.* A web-based clinical decision tool to support treatment decision-making in psychiatry: a pilot focus group study with clinicians, patients and carers. *BMC Psychiatry* **17**, 265 (2017).
54. Kluge EH. Medical narratives and patient analogs: the ethical implications of electronic patient records. *Methods Inf. Med.* **38**, 253–9 (1999).
55. Mamzer MF. *et al.* Partnering with patients in translational oncology research: ethical approach. *J. Transl. Med.* **15**, 74 (2017).
56. McKernan LC., Clayton EW. & Walsh CG. Protecting Life While Preserving Liberty: Ethical Recommendations for Suicide Prevention With Artificial Intelligence. *Front. Psychiatry* **9**, 650 (2018).
57. Panchmatia JR., Visenio MR. & Panch T. The role of artificial intelligence in orthopaedic surgery. *Br. J. Hosp. Med. Lond. Engl. 2005* **79**, 676–681 (2018).
58. Price, W. N. & Cohen, I. G. Privacy in the age of medical big data. *Nat. Med.* **25**, 37–43 (2019).
59. Racine E., Boehlen W. & Sample M. Healthcare uses of artificial intelligence: Challenges and opportunities for growth. *Healthc. Manage. Forum* 840470419843831 (2019). doi:10.1177/0840470419843831
60. Rigby M., Draper R. & Hamilton I. Finding ethical principles and practical guidelines for the controlled flow of patient data. *Methods Inf. Med.* **38**, 345–9 (1999).
61. Sethi P. & Theodos K. Translational bioinformatics and healthcare informatics: computational and ethical challenges. *Perspect. Health Inf. Manag.* **6**, 1h (2009).
62. Ngiam, K. Y. & Khor, I. W. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol.* **20**, e262–e273 (2019).
63. Santoro, E. [Artificial intelligence in medicine: limits and obstacles.]. *Recenti Prog. Med.* **108**, 500–502 (2017).
64. Dickens BM. & Cook RJ. Legal and ethical issues in telemedicine and robotics. *Int. J. Gynaecol. Obstet. Off. Organ Int. Fed. Gynaecol. Obstet.* **94**, 73–8 (2006).
65. Luxton DD. Recommendations for the ethical use and design of artificial intelligent care providers. *Artif. Intell. Med.* **62**, 1–10 (2014).
66. Ameneiros-Lago, E., Carballada-Rico, C., Garrido-Sanjuan, J. A. & Garcia Martinez, A. [Adequacy of clinical interventions in patients with advanced and complex disease. Proposal of a decision making algorithm]. *Rev. Calid. Asist. Organo Soc. Espanola Calid. Asist.* **30**, 38–43 (2015).
67. Bali J., Garg R. & Bali RT. Artificial intelligence (AI) in healthcare and biomedical research: Why a strong computational/AI bioethics framework is required? *Indian J. Ophthalmol.* **67**, 3–6 (2019).
68. Buiten, M. C. Towards Intelligent Regulation of Artificial Intelligence. *Eur. J. Risk Regul.* **10**, 41–59 (2019).
69. Enrico Coiera et al: First compute no harm. *The BMJ* (2017). Available at: https://blogs-bmj-com.stanford.idm.oclc.org/bmj/2017/07/19/enrico-coiera-et-al-first-compute-no-harm/. (Accessed: 16th September 2019)
70. Evans, B. J. Big Data and Individual Autonomy in a Crowd. *Big Data, Health Law, and Bioethics* (2018). doi:10.1017/9781108147972.003
71. Gesley, J. *Regulation of Artificial Intelligence in Selected Jurisdictions*. 138
72. Grsson D., Petrellaon J., Mehl J., Burgun A. & Garcelon N. [Ethical, legal and operational issues of artificial intelligence]. *Rev. Prat.* **68**, 1145–1148 (2018).
73. Lawrence, B. L. *Text - H.Res.153 - 116th Congress (2019-2020): Supporting the development of guidelines for ethical development of artificial intelligence*. (2019).
74. Smith, H. L. S. *et al. ARTIFICIAL INTELLIGENCE: WITH GREAT POWER COMES GREAT RESPONSIBILITY JOINT HEARING BEFORE THE SUBCOMMITTEE ON RESEARCH AND TECHNOLOGY & SUBCOMMITTEE ON ENERGY COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY HOUSE OF REPRESENTATIVES*. 157
75. Thune, J. *et al. THE DAWN OF ARTIFICIAL INTELLIGENCE HEARING BEFORE THE SUBCOMMITTEE ON SPACE, SCIENCE, AND COMPETITIVENESS OF THE COMMITTEE ON COMMERCE, SCIENCE, AND TRANSPORTATION UNITED STATES SENATE*. 61
76. Vayena, E. & Tasioulas, J. The dynamics of big data and human rights: the case of scientific research. *Philos. Transact. A Math. Phys. Eng. Sci.* **374**, (2016).
77. Vayena, E., Blasimme, A. & Cohen, I. G. Machine learning in medicine: Addressing ethical challenges. *PLOS Med.* **15**, e1002689 (2018).
78. Abouelj, S., Liu, R. H., Desta, B. N., Chaurasia, A. & Ebrahim, S. The Use of Artificially Intelligent Self-Diagnosing Digital Platforms by the General Public: Scoping Review. *JMIR Med. Inform.* **7**, e13445 (2019).
79. Begoli, E., Bhattacharya, T. & Kusnezov, D. The need for uncertainty quantification in machine-assisted medical decision making. *Nat. Mach. Intell.* **1**, 20–23 (2019).
80. Cath C. Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philos. Transact. A Math. Phys. Eng. Sci.* **376**, (2018).
81. Veinot, T. C., Mitchell, H. & Ancker, J. S. Good intentions are not enough: how informatics interventions can worsen inequality. *J. Am. Med. Inform. Assoc.* **25**, 1080–1088 (2018).
82. Žliobaitė, I., Pechenizkiy, M. & Gama, J. An Overview of Concept Drift Applications. *Big Data Anal. New Algorithms New Soc.* 91–114 (2016). doi:10.1007/978-3-319-26989-4_4
83. Deist TM. *et al.* Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. *Clin. Transl. Radiat. Oncol.* **4**, 24–31 (2017).

Figure 1. Pipeline model for identifying ethical considerations for machine learning healthcare applications.

that the health applications powered by ML are, in and of themselves, exceptional. The clinical applications all seek to perform, in novel and hopefully better ways, standard healthcare tasks, such as diagnosis, generating a prognosis, or assisting with treatment decision-making. These tasks each have already identified ethical considerations that likely apply to ML-HCAs. The technology itself is also built from essentially standard clinical information, such as patient demographic or clinical information, such as laboratory values or diagnostic images, and while this information is being analyzed in remarkable ways, standard ethical considerations about these data also likely apply to ML-HCAs. Accordingly, a framework to guide identifying ethical considerations does not need to be focused on exceptions, even as it should leave space for exceptional considerations to be identified.

Breadth of Applications

The breadth of emerging ML-HCAs, regarding what they aim to do, how they are constructed, and where they are being applied, is remarkably broad. ML-HCAs range from fully autonomous artificial intelligence diagnosis of diabetic retinopathy in primary care settings to non-autonomous mortality predictions to guide insurance and allocation of healthcare resources (Ching et al. 2018). The analytic framework guiding the identification of ethical considerations should therefore ideally be sufficiently generic to be useful across a wide variety of ML-HCAs. For the ethical appraisal of any given ML-HCA, detailed content and context-specific knowledge will always be needed to provide more thorough and precise ethical evaluation, and this will require cross-disciplinary collaborations. A framework for identification of ethical considerations, one that can accommodate a broad range of ML-HCAs, would help such collaboration.

Allure of Highly Restricted Focus

Many ML-HCA computer scientists have already turned away from ethical analysis as unworkable or not adequately responsive to ongoing ML-HCA development, have instead focused exclusively on the ethical consideration of fairness and emerging concerns regarding bias, and have begun to pursue an ideal of "algorithmic fairness," or the ability to computationally demonstrate a lack of between-group bias with an ML application (Rajkomar et al. 2018). They reason that if latent biases can be identified, ML approaches might be used to correct for them or improve "fairness" (Rajkomar et al. 2018). Highly focused approaches such as this assume an *a priori* comprehensive understanding of where and why such biases are occurring; if this assumption is wrong, these approaches risk introducing a complex set of unintended biases in attempts to correct the initial bias (Goodman et al. 2018). More generally, a highly restrictive focus and limited framework may be applicable for ultimately addressing a specific ethical consideration and set of concerns, but will not suffice to manage the uncertainty regarding other potential ethical considerations.

Diverse Stakeholders

Finally, ML-HCAs are likely to have a broad range of stakeholders, from patients and health care practitioners, to computer scientists, engineers, and entrepreneurial developers, to healthcare organizations and payers, to oversight bodies charged with regulating medical practice. Any framework to help identify ethical considerations should provide for potential perspectives and concerns of each of these diverse stakeholders, commensurate with their expertise.

PIPELINE FRAMEWORK TO IDENTIFY ETHICAL CONSIDERATIONS

We propose using the developmental pipeline of ML-HCAs, from conception to implementation, with a parallel pipeline of ML-HCA evaluation and oversight, as a framework to help identify ethical considerations (Figure 1). This pipeline framework is neither too narrow nor too broad, applies across a wide variety of ML-HCAs, and accommodates the perspectives and concerns of different groups of stakeholders. Along this pipeline, key questions can be asked to uncover values-based issues, which in turn can be linked to both standard and potentially novel ethical considerations (which we have annotated with citations based on our literature search).

Conception: Auditability, Transparency Standards, and Conflicts of Interest

When designers and implementers of a ML-HCA clearly declare the intentions, indications for use, and goals for an application, clinicians, patients, regulators, and other stakeholders are better enabled to exercise their own evaluative and decisional autonomy. Without transparency about intentions or specific goals, stakeholders will not be able to decide for themselves whether they want to support these intentions, or whether they believe that the ML-HCA will advance these intentions and the stated goals (Feudtner et al. 2018). Stakeholders do not need to understand in detail the inner working of an ML-HCA in order to achieve "auditability."

To support evaluative autonomy, transparency will require "auditability": ML systems in medicine must have an explainable architecture, designed to align with human cognitive decision-making processes familiar to physicians, and directly tied to clinical evidence. Any ML-HCA's functioning and output will need to be interpretable to any stakeholder who uses the output to inform clinical decisions so that they can evaluate whether the ML-HCA is likely to live up to the stated intentions. This would include auditability of aspects of the development phase (such as algorithm design, the training data, training process testing, and validation methods) and in the initial

clinical implementation phase (where, as is now the case for clinical trials, pre-specification of study design, outcome measures, and analysis are required to enable a potential audit regarding whether the trial was conducted according to the pre-specified plan).

A simple but key aspect of determining the safety of any healthcare application depends upon the ability to inspect the application—to literally disassemble and examine a physical application to determine how the parts work together, to see the mechanisms at work, and thus better understand how the application might fail. The process is similar for software applications and, by analogy, to the components and physiologic mechanisms of medications or mechanical devices. ML-HCAs, however, can present a "black box" problem, with workings that are not inspectable by evaluators, clinicians, and patients. Unlike MRI scanners, where the clinician-user may not understand how the MRI functions but an engineer or designer could take apart the machine and explain its inner workings, for certain ML approaches (such as neural networks) the learning methods of the system can be opaque even to system designers. Even when post hoc explainability can be provided, such black box, neural-network based systems are vulnerable to "catastrophic failures" and implicit biases in the training sets compared to more explainable ML architectures (Finlayson et al. 2019; Shah et al. 2018). A non-inspectable, autonomous system poses a higher risk of patient harm, raises questions about the responsibility of the system in situations of harm (and the need for the system to have malpractice insurance), and could engender significant backlash against autonomous systems. Transparency, however, needs to be balanced against protection of the intellectual property of ML-HCA design.

Transparency standards should also clarify whether a ML-HCA is "locked" or "continuously learning." Continuous learning ML-HCAs automatically update using inputs during use, as opposed to locked ML-HCAs, which are deterministic (Daniel 2019). Transparency about whether the ML-HCA is locked or continuously learning is critical because evaluating the safety, efficiency, and equity for a continuous learning ML-HCA is more challenging, and therefore understanding ethical considerations and addressing concerns is more difficult.

Some have argued that continuous ML learning in healthcare contexts may be harmful (Challen et al. 2019). With continuous learning, "distributional shift" can occur, if target training data does not match ongoing patient data (such as if the ML-HCA is applied to a population with higher pretest probability of disease than the training population data), leading an ML-HCA to begin to draw inaccurate conclusions. Even if a ML-HCA underwent exemplary development and rigorous initial evaluation, subsequent evaluations of accuracy will be necessary over time due to what can be thought of as association half-life. The associations between the data elements that underwrote the outcome prediction are likely to change over time, due to changes in populations, technology, and processes of care. In addition, in many cases a goal of ML-HCA is lowering cost, yet for certain conditions (such as most chronic diseases, where costs are driven by long-term adverse outcomes), obtaining high quality long-term outcome data needed for validation and subsequent updating may require more not less financial resources.

Transparency standards should also specify whether a ML-HCA is assistive or autonomous. Assistive ML-HCAs aid healthcare providers by supplying "recommendations" regarding treatment, diagnosis, or management, while relying on user interpretation of any recommendations to make decisions. Autonomous ML-HCAs provide direct diagnosis and management statements without any clinician's or any other human interpretation or supervision. Since the developer's choice of a ML-HCA's level of autonomy has clear implications for assumption of responsibility and liability, this autonomy level needs to be apparent.

Last but not least, with growing understanding that mores and values can intentionally or unintentionally become embedded in the design of engineered systems (Manders-Huits 2011) transparency will be required regarding any potential conflicts of interest. These potential conflicts of interest include individual financial interests (such as payment for services or personal ownership of stocks) as well as any operational interests of the organization that may not be aligned with the duty of clinicians and health care delivery organizations to advance the best interest of each patient under their care (Kohli et al. 2017; Fischer et al. 2016; Jaremko et al. 2019). Transparency on the part of ML-HCA developers allows clinicians, patients, and society as a whole to independently assess potential conflicts of interest and other harms that may have negative consequences outside of the AI developer's direct control.

Development: Perpetuation of Bias within Training Data, Risk of Harm Due to Group Membership, and Obtaining Training Data

An important and acknowledged concern (Char et al. 2018; Rajkomar et al. 2019) in the development of

ML-HCAs relates to the possibility of bias, particularly whether latent biases in training data may be perpetuated or even amplified. Examples already exist of predictive scores failing both because of poorly composed training data and because, when expanded to broader populations, racially discriminatory outcomes occurred (Char et al. 2018; Obermeyer et al. 2019). For example, ML programs designed to aid judges in sentencing by predicting an offender's risk for recidivism have shown a disturbing propensity for racial discrimination (Angwin et al. 2016). In healthcare, when used to predict cardiovascular event risk in non-Caucasian populations, Framingham study data has shown bias both over- and under-estimating risk for different specific populations (Gijsberts et al. 2015).

Furthermore, any perpetuated biases incorporated into a ML-HCA may subsequently impact clinical decisions and support self-fulfilling prophesies. For example, if clinicians currently routinely de-escalate or withhold interventions in patients with specific severe injuries or progressive conditions, ML systems may classify such clinical scenarios as nearly always fatal, and any ML-HCA built on such a classification would likely result in an even higher likelihood of de-escalation or withholding, thereby reducing the opportunity to improve outcomes for such conditions (Begoli et al. 2019; Fiske et al. 2019; Nabi 2018; Cohen et al. 2014; Ho 2019; Taljaard et al. 2014). Training of ML-HCAs against real world data, rather than high-quality research-grade data, may simply perpetuate suboptimal clinical practices that are not aligned with the best scientific evidence. Conversely, an algorithm's over-reliance on research-grade data alone may miss important clinically relevant sources of knowledge, lowering the quality of care delivered (Fenton et al. 2007).

A related concern is obtaining needed training data, and questions of data ownership, pricing and protecting privacy. Machine learning requires large amounts of training data. The aggregation and curation of these large datasets raises not only issues regarding specifying the standards that high-quality reference standard data must achieve, but also issues regarding data privacy and data ownership (Aboueid et al. 2019; Amarasingham et al. 2016; Cohen et al. 2014; Gruson et al. 2018; Henshall et al. 2017; Jaremko et al. 2019; Nicholson Price and Glenn Cohen 2019; Racine et al. 2019; SFR-IA Group 2018; Vayena and Blasimme 2018). For diagnostic ML-HCAs, training data will likely be based on data collected from individual patients obtained during routine clinical care (such as laboratory test values, biopsy findings, or diagnostic images) or from individual enrollees in health insurance plans (such as medical diagnoses from medical encounters or health care utilization patterns), along with personal demographic information. Other ML-HCAs may be based on data from non-clinical sources (such as personal devices, social media, financial, or legal sources), which may contain potentially controversial data elements or have been collected via novel means that we cannot foresee. While privacy laws and regulations are currently in place, open questions need to be addressed regarding who owns this data, the traceability of specific data elements from each individual patient into the "big" datasets, and whether patient rights to privacy should be extended or curtailed.

To focus on one example: how should we adjudicate claims regarding the value of the data—and the value of each individual's contribution of their data to the aggregate dataset on which a ML-HCA is constructed—and the pricing of the ML-HCA itself? Most likely, large health systems will have generated and compiled much of this "big data," which in turn was paid for by insurance premiums and co-pays. Many data sets, particularly those involving image or biopsy interpretations, may also reflect the significant intellectual contributions of interpreting clinicians. The subsequent effort to curate the data and then develop the ML-HCA adds value to the raw data, but certainly not all of the value. Just as there are debates regarding drug pricing, when the initial development of a drug was supported by federal or nonprofit funding prior to acquisition and further development by a pharmaceutical company, similar debates are already emerging with ML-HCAs (Ornstein and Thomas 2018). There has also been ongoing patient activism for inclusion in recognition for specimen contribution to scientific advances (Bledsoe and Grizzle 2013).

Calibration: Accuracy, Trading off Test Characteristics, and Calibrated Risk of Harm

In order for a ML-HCA to maximize clinical benefits and minimize harm, the application must perform in accordance with the cardinal design features of safety (to prevent injuries and hazards), efficiency (that the application effectively solves the problem it was designed for and does so at a reasonable cost, in particular regarding the costs of incorrect classifications, such as false negative or false positive diagnoses), and equity (that the advantages of the application are shared fairly by all). In concrete terms, this means at a minimum that the application will need to provide

accurate diagnostic or predictive information on the vast majority of patients for whom the ML-HCA is intended to be used, irrespective of subgroup such as age or race.

Determining the accuracy of a ML-HCA is, however, not straightforward. Unlike ML designed for other contexts, such as to play games of skill (e.g. chess, go), many medical decisions and diagnoses cannot be perfectly labeled as correct or incorrect and down-stream outcomes cannot always be anticipated (Fenton et al. 2007). This is a known challenge with reference "gold standards" in healthcare (Frieden 2017). While ML accuracy can be higher than that of individual experts in interpretation of clinical images such as radiologic scans, pathology slides, and photographs of skin lesions (Ching et al. 2018), the estimated accuracy of a ML-HCA is dependent on the clinical context in which the application is being assessed. Validation studies therefore need to be done not only in the context of rigorously managed research trials, but also in general populations of patients. In these settings, endpoints should address patient safety (measured as sensitivity, assuring that patients with the disease or in a designated risk category are not missed), efficiency of the application to provide an accurate diagnosis (measured as specificity, assuring that patients without the disease are not over-diagnosed, along with corresponding positive and negative predictive values.) An equitable ML-HCA will provide equivalent levels of accuracy within the intended-use population across multiple patient subgroups or characteristics, and also achieve equivalent levels of "determinability," or the ability of the ML-HCA to provide a clinically relevant output based on the clinically available inputs (and not simply declare that the inputted information is not sufficient).

The notion of accuracy in an ML-HCA, inherently involves tradeoffs between test characteristics, guided by designer value judgments with consequent ethical implications. For any diagnostic or predictive test, whether the test uses ML or not, the performance is calibrated to trade off a higher level of one test characteristic (such as more people with the condition being correctly classified as having the condition) with a corresponding lower level of another test characteristic (such as more people who do not have the condition being misclassified as having the condition). Both of these test characteristics will also be influenced by the determinability characteristics of the test (that is, whether the test can use the clinically available information, or whether the test cannot make a determination of disease status or determine a predicted probability), and the determinability test characteristic itself is also a calibrated tradeoff between returning a result or declaring that the inputted information is insufficient.

Even if a specific ML-HCA is found to be superior to an established clinical practice with regard to all test characteristics, that specific ML-HCA will have calibrated not only greater accuracy, but also specific forms of inaccuracy: the design will predictably generate false positives and false negatives, or indeterminate results, as must be the case with any method of classification, whether based on human judgment or machine learning. The key ethical consideration would be whether these inaccuracies (and any consequent harms) are outweighed by potential benefits and distributed among patients in an equitable manner.

Implementation, Evaluation, and Oversight: Adverse Events, Ongoing Assessment of Accuracy and Usage

During development, when ML systems may be validated on idealized data, their accuracy may be measured to be "perfect" (in other words, not statistically different from a perfect algorithm or observer who always outputs the true state of disease). But in real-world settings—where there is the potential for human operator error, data inputs of lower quality and nearly infinite variance, and additional potentially relevant data captured in a modality not accessible to the ML-HCA—the true accuracy is typically lower, even when the underlying ML-HCA has been locked and unchanged (Abràmoff et al. 2018). As the measured sensitivity, specificity, and determinability change, so too will the potential benefits and potential harms, and the resulting benefit-to-harm ratio. For example, earlier computer-aided diagnostic tools such as EKG interpretation and mammography appeared in preliminary studies to offer value-adding diagnostic accuracy, yet in subsequent evaluations of their actual intended use (specifically, to assist front-line clinicians in making medical decisions) have failed to demonstrate benefit and raised the possibility of some degree of harm (Fenton et al. 2007; Schläpfer and Wellens 2017). In a similar manner, as a ML-HCA moves beyond the initial implementation setting and into a wider-ranging clinical use setting, assessing whether patients continue to benefit will need to be ongoing.

An evaluation and oversight process (Figure 1, row 2) will have to address questions of whether, across sites and populations (including across races,

ethnicities, sex and ages), and over time, use of the ML-HCA continues to provide benefit. More prosaically, just like every other health care device, every particular ML-HCA in clinical use should undergo inspection from time-to-time to determine if the accuracy of the output deviates from the application's previous performance standard. In addition to addressing pragmatic concerns of making sure the ML-HCA continues to perform as intended, such evaluation and oversight can uncover additional values-based issues, which raise ethical considerations (Figure 1).

ML-HCA's interpretion of patient data, even if superior to human interpretation, will certainly not be perfect. Interpretation errors may result in patient harm. In such instances, there is a tendency to judge machine-based error more severely than human error (Cathy O'Neil 2017). This tendency warrants scrutiny. If, comparing the machine-based and the human-based scenarios, the nature and probability of the error and the magnitude of the ensuing harm are equivalent, this tendency does not appear to legitimate, instead reflecting a pro-human or anti-machine bias. Determining the appropriate degree of privilege to accord an established practice presents a tradeoff between the prospect of more accurate interpretation of data via the novel ML-HCA and appropriate caution in the face of heightened uncertainty.

In addition, ML-HCAs will create new information flows and consequently need resource allocation, including the important resource of clinical attention. Accordingly, evaluations of the impact of ML-HCA output on clinical workflow will be warranted. ML-HCAs may simply add information 'noise' to an already crowded clinical environment, becoming something followed either blindly or poorly. Some have speculated that users may feel that ML-HCAs may remove their own liability in clinical decision making (O'Sullivan et al. 2019). The output from a ML-HCA—even one that is billed as being only advisory, to offer guidance—may take on an authority never intended. This has been the case in non-healthcare contexts, where individuals who have challenged a ML-based recommendation have frequently been required to provide significantly more robust evidence to refute the ML recommendation than the evidence or data which the ML recommendation was actually based on (Cathy O'Neil 2017).

Unintended uses of a ML-HCA, with new potential harms as well as any hoped-for benefits, will also need to be monitored. Some potential unintended uses may be predictable before implementation (such as a ML system for mortality prediction being co-opted to limit hospital mortality statistics or costs). Assuring that a ML system is not being inadvertently yet inappropriately re-purposed will also require ongoing monitoring. For example, a system intended for diagnosis of diabetic retinopathy might be co-opted (or unintentionally interpreted by patients or health providers) as an ophthalmic screening exam for broader conditions than just diabetic retinopathy.

Lastly, based on experiences with the implementation of electronic medical record platforms, monitoring will also be warranted to assess the equity of access to ML-HCA, which may be more readily available in larger or better financed health systems than in small systems or practices, which in turn could result in poorer outcomes in these smaller sites.

USING THE PIPELINE FRAMEWORK

Now that we have laid out the framework of a pipeline model of ML-HCAs, let us outline how the framework can be used for the purpose of ethical analysis.

As the model makes clear, there are many potential points in the ML-HCA pipeline where an individual or a group might want to identify and think through ethical considerations that arise specifically at that point in the overall pipeline. The questions posed in the framework for a given stage of the pipeline may help in identifying other, novel considerations.

The framework also should be used, even when focused on a particular point in the pipeline, to identify and examine ethical considerations in previous steps. ML-HCA developers and users poised at a particular point in the pipeline inheret the ethical operating characteristics that arise from previous decisions about how the ML-HCA has been constructed.

Heading in the other direction, the framework can also be used to look ahead, anticipating future development and implementation (or implementations in other settings). Identification of potential future consequences can aid ethical evaluation and decisions regarding design, development, implementation, and evaluation.

As mentioned above, these activities can be done by individuals or groups, in particular multi-stakeholder groups. Given the protracted sequence of steps in ML-HCA development and implementation, the potentially illuminating (and obfuscating) technical details of the inner ML workings of the application, and the complicated and rather expansive set of ethical considerations, the pipeline framework provides a

guide to help these individuals and groups with the task of identifying and evaluating present, past, and future ethical issues.

The pipeline framework also offers groups of diverse stakeholders a "bigger picture" of ML-HCAs that can, with dialogue, help to forge a shared mental model of the range of revelant questions and ethical considerations that should guide design and evaluation decisions. The broadness of the framework will help combat any tendency to focus narrowly on one ethical consideration while potentially neglecting other relevant considerations and thus sidestepping grappling with tradeoffs. Lastly, the common basic elements of the pipeline—an application is conceived of, developed, calibrated, implemented, and evaluated, with various forms of oversight—allows for ready comparison of the ML-HCA pipeline to the pipelines of other medical technologies, and to see that while ML-HCAs do raise some novel issues, they also raise many issues common to existing diagnostic or therapeutic technologies. This can put a check on unwarranted ML-HCA exceptionalism in our thinking about the ethics of this emerging technology.

CONCLUSION

Machine learning in healthcare has arrived. Along with many potential benefits to healthcare delivery, ML-HCA is likely to raise complex and as yet only partially considered ethical considerations with implementation. The pipeline framework, starting with a map of the conception, development, implementation, and the parallel evaluation and oversight tasks of ML-HCAs, and then layering over this map key questions, value-based issues, and ethical considerations, is an approach for systematically identifying these ethical considerations and for facilitating inter-disciplinary dialogue and collaboration to better understand and subsequently manage the ethical implications of ML-HCAs.

DISCLOSURE STATEMENT

Michael Abràmoff is founder and CEO of Digital Diagnostics, Inc., and has patents, patent applications, ownership, employment, and consultancy related to the subject of this article.

FUNDING

This work was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number [grant number K01HG008498 to DC]. MDA was supported by P30 EY025580, and an unrestricted grant from Research to Prevent Blindness, New York, NY.

REFERENCES

Aboueid, S., R. H. Liu, B. N. Desta, A. Chaurasia, and S. Ebrahim. 2019. The use of artificially intelligent self-diagnosing digital platforms by the general public: Scoping review. *JMIR Medical Informatics* 7 (2):e13445. doi: 10.2196/13445.

Abràmoff, M. D., P. T. Lavin, M. Birch, N. Shah, and J. C. Folk. 2018. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Medicine* 1 (1):39. doi: 10.1038/s41746-018-0040-6.

Amarasingham, R., A.-M. J. Audet, D. W. Bates, et al. 2016. Consensus statement on electronic health predictive analytics: A guiding framework to address challenges. *EGEMS* 4 (1):1163. doi: 10.13063/2327-9214.1163.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Begoli, E., T. Bhattacharya, and D. Kusnezov. 2019. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence* 1 (1):20–23. doi: 10.1038/s42256-018-0004-1.

Bledsoe, M. J., and W. E. Grizzle. 2013. Use of human specimens in research: The evolving United States regulatory, policy, and scientific landscape. *Diagnostic Histopathology* 19 (9):322–330. doi: 10.1016/j.mpdhp.2013.06.015.

Bostrom, N., and E. Yudkowski. 2011. The ethics of artificial intelligence. In *The Cambridge handbook of artificial intelligence*, 316–334. Cambridge, UK: Cambridge University Press.

Brey, P. A. E. 2012. Anticipatory ethics for emerging technologies. *NanoEthics* 6 (1):1–13. doi: 10.1007/s11569-012-0141-7.

Cathy O'Neil. 2017. The ivory tower can't keep ignoring tech. *The New York Times*, November 14, 2017. https://www.nytimes.com/2017/11/14/opinion/academia-tech-algorithms.html

Challen, R., J. Denny, M. Pitt, L. Gompels, T. Edwards, and K. Tsaneva-Atanasova. 2019. Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety* 28 (3):231–237. doi: 10.1136/bmjqs-2018-008370.

Char, D. S., N. H. Shah, and D. Magnus. 2018. Implementing machine learning in health care – Addressing ethical challenges. *The New England Journal of Medicine* 378 (11):981–983. doi: 10.1056/NEJMp1714229.

Ching, T., D. S. Himmelstein, B. K. Beaulieu-Jones, et al. 2018. Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface* 15 (141):20170387. doi: 10.1098/rsif.2017.0387.

Churchman, C. W., R. L. Ackoff, and E. L. Arnoff. 1957. *Introduction to operations research*. Oxford, England: Wiley. doi: 10.2307/3006881.

Cohen, I. G., R. Amarasingham, A. Shah, B. Xie, and B. Lo. 2014. The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Affairs*

(Project Hope) 33 (7):1139–1147. doi: 10.1377/hlthaff.2014.0048.

Comfort, N. 2016. The overhyping of precision medicine. *The Atlantic*, 2016. https://www.theatlantic.com/health/archive/2016/12/the-peril-of-overhyping-precision-medicine/510326/

Commissioner Office of the 2020. FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. *FDA*. February 20, 2020. http://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye.

Daniel, G. 2019. Current state and near-term priorities for ai-enabled diagnostic support software in health care. June, 51.

Fenton, J. J., S. H. Taplin, P. A. Carney, et al. 2007. Influence of computer-aided detection on performance of screening mammography. *The New England Journal of Medicine* 356 (14):1399–1409. doi: 10.1056/NEJMoa066099.

Feudtner, C., T. Schall, P. Nathanson, and J. Berry. 2018. Ethical framework for risk stratification and mitigation programs for children with medical complexity. *Pediatrics* 141 (Suppl 3):S250–S258. doi: 10.1542/peds.2017-1284J.

Finlayson, S. G., J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane. 2019. Adversarial attacks on medical machine learning. *Science* 363 (6433):1287–1289. doi: 10.1126/science.aaw4399.

Fischer, T., K. B. Brothers, P. Erdmann, and M. Langanke. 2016. Clinical decision-making and secondary findings in systems medicine. *BMC Medical Ethics* 17 (1):32. doi: 10.1186/s12910-016-0113-5.

Fiske, A., P. Henningsen, and A. Buyx. 2019. Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of Medical Internet Research* 21 (5):e13216. doi: 10.2196/13216.

Frieden, T. R. 2017. Evidence for health decision making - Beyond randomized, controlled trials. *The New England Journal of Medicine* 377 (5):465–475. doi: 10.1056/NEJMra1614394.

Gijsberts, C. M., K. A. Groenewegen, I. E. Hoefer, et al. 2015. Race/ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events. *PLoS One* 10 (7):e0132321. doi: 10.1371/journal.pone.0132321.

Goodman, S. N., S. Goel, and M. R. Cullen. 2018. Machine learning, health disparities, and causal reasoning. *Annals of Internal Medicine* 169 (12):883–884. doi: 10.7326/M18-3297.

Gruson, D., J. Petrelluzzi, J. Mehl, A. Burgun, and N. Garcelon. 2018. Ethical, legal and operational issues of artificial intelligence. *La Revue du praticien* 68 (10):1145–1148.

Henshall, C., L. Marzano, K. Smith, et al. 2017. A web-based clinical decision tool to support treatment decision-making in psychiatry: A pilot focus group study with clinicians, patients and carers. *BMC Psychiatry* 17 (1):265. doi: 10.1186/s12888-017-1406-z.

Hern, A. 2017. Royal free breached UK data law in 1.6m patient deal with Google's DeepMind. *The Guardian*, July 3, 2017. https://www.theguardian.com/technology/2017/jul/03/google-deepmind-16m-patient-royal-free-deal-data-protection-act.

Ho, A. 2019. Deep ethical learning: Taking the interplay of human and artificial intelligence seriously. *The Hastings Center Report* 49 (1):36–39. doi: 10.1002/hast.977.

Jaremko, J. L., M. Azar, R. Bromwich, et al. 2019. Canadian Association of Radiologists white paper on ethical and legal issues related to artificial intelligence in radiology. *Canadian Association of Radiologists Journal = Journal l'Association canadienne des radiologistes* 70 (2):107–118. doi: 10.1016/j.carj.2019.03.001.

Keeney, R. A. 1992. *Value-focused thinking: A path to creative decisionmaking*. Cambridge, MA: Harvard University Press.

Kohli, M., L. M. Prevedello, R. W. Filice, and J. R. Geis. 2017. Implementing machine learning in radiology practice and research. *American Journal of Roentgenology* 208 (4):754–760. doi: 10.2214/AJR.16.17224.

Maddox, T. M., J. S. Rumsfeld, and P. R. O. Payne. 2019. Questions for artificial intelligence in health care. *JAMA* 321 (1):31–32. doi: 10.1001/jama.2018.18932.

Manders-Huits, N. 2011. What values in design? The challenge of incorporating moral values into design. *Science and Engineering Ethics* 17 (2):271–287. doi: 10.1007/s11948-010-9198-2.

Matheny, M. E., D. Whicher, and S. Thadaney Israni. 2020. Artificial intelligence in health care: A report from the national academy of medicine. *JAMA* 323 (6):509. doi: 10.1001/jama.2019.21579.

Matheny, M., S. Thadaney Israni, M. Ahmed, and D. Whicher, eds. 2019. *Artificial intelligence in health care*: The *hope, the hype, the promise, the peril. The Learning Helath System Series*. Washington, DC: National Academy of Medicine.

Nabi, J. 2018. How bioethics can shape artificial intelligence and machine learning. *The Hastings Center Report* 48 (5):10–13. doi: 10.1002/hast.895.

Nicas, J., J. Glanz, and D. Gelles. 2019. In test of Boeing Jet, Pilots had 40 seconds to fix error. *The New York Times*, March 25, 2019, sec. Business. https://www.nytimes.com/2019/03/25/business/boeing-simulation-error.html (accessed 11 September 2020).

Nicholson Price, W., and I. Glenn Cohen. 2019. Privacy in the age of medical big data. *Nature Medicine* 25 (1):37–43. doi: 10.1038/s41591-018-0272-7.

Obermeyer, Z., and E. J. Emanuel. 2016. Predicting the future – Big data, machine learning, and clinical medicine. *The New England Journal of Medicine* 375 (13):1216–1219. doi: 10.1056/NEJMp1606181.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366 (6464):447–453. doi: 10.1126/science.aax2342.

Ornstein, C., and K. Thomas. 2018. Sloan Kettering's Cozy deal with start-up ignites a new uproar. *The New York Times*, September 20, 2018. https://www.nytimes.com/2018/09/20/health/memorial-sloan-kettering-cancer-paige-ai.html

O'Sullivan, S., S. Leonard, A. Holzinger, et al. 2019. Operational framework and training standard requirements for AI-empowered robotic surgery. *International*

Journal of Medical Robotics and Computer Assisted Surgery May 30:e2020.
Racine, E., W. Boehlen, and M. Sample. 2019. Healthcare uses of artificial intelligence: Challenges and opportunities for growth. *Healthcare Management Forum* 32 (5): 272–275. doi: 10.1177/0840470419843831.
Rajkomar, A., J. Dean, and I. Kohane. 2019. Machine learning in medicine. *The New England Journal of Medicine* 380 (14):1347–1358. doi: 10.1056/NEJMra1814259.
Rajkomar, A., M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin. 2018. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine* 169 (12):866–872. doi: 10.7326/M18-1990.
Rittel, H. W. J., and M. M. Webber. 1973. Dilemmas in a general theory of planning. *Policy Sciences* 4 (2):155–169. doi: 10.1007/BF01405730.
Rosenberg, M., and S. Frenkel. 2018. Facebook's role in data misuse sets off storms on two continents. *The New York Times*, March 1,82,018 sec. U.S. https://www.nytimes.com/2018/03/18/us/cambridge-analytica-facebook-privacy-data.html.
Rosenhead, J., and M. John, eds. 2008. *Rational analysis for a problematic world revisited: Problem structuring methods for complexity, uncertainty and conflict.* 2nd ed., repr. Chichester: Wiley.
Ross, C., and I. Swelitz. 2017. IBM pitched Watson as a revolution in cancer care. It's nowhere close. *Stat*, September 5, 2017. https://www.statnews.com/2017/09/05/watson-ibm-cancer/.
Schläpfer, J., and H. J. Wellens. 2017. Computer-interpreted electrocardiograms: Benefits and limitations. *Journal of the American College of Cardiology* 70 (9):1183–1192. doi: 10.1016/j.jacc.2017.07.723.
SFR-IA Group. 2018. Artificial intelligence and medical imaging 2018: French radiology community white paper. *Diagnostic and Interventional Imaging* 99 (11):727–742. 10.1016/j.diii.2018.10.003.
Shah, A., S. Lynch, M. Niemeijer, et al. 2018. Susceptibility to misdiagnosis of adversarial images by deep learning based retinal image analysis algorithms. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 1454–1457. Washington, DC: IEEE.
Shilton, K. 2018. Values and ethics in human-computer interaction. *Foundations and Trends® in Human–Computer Interaction* 12 (2):107–171. doi: 10.1561/1100000073.
Stenmark, C. K., A. L. Antes, C. E. Thiel, J. J. Caughron, X. Wang, and M. D. Mumford. 2011. Consequences identification in forecasting and ethical decision-making. *Journal of Empirical Research on Human Research Ethics* 6 (1): 25–32. doi: 10.1525/jer.2011.6.1.25.
Taljaard, M., M. Tuna, C. Bennett, et al. 2014. Cardiovascular Disease Population Risk Tool (CVDPoRT): Predictive algorithm for assessing CVD risk in the community setting. A study protocol. *BMJ Open* 4 (10):e006701. doi: 10.1136/bmjopen-2014-006701.
Vayena, E., and A. Blasimme. 2018. Health research with big data: Time for systemic oversight. *The Journal of Law, Medicine & Ethics: A Journal of the American Society of Law, Medicine & Ethics* 46 (1):119–129. doi: 10.1177/1073110518766026.

APPENDIX: LITERATURE REVIEW METHODS

A systematic search technique was used to identify relevant literature. Librarians from both the Lane Library at Stanford University School of Medicine and Robert Crown Library at Stanford University School of Law were consulted to define comprehensive search strategies in relevant databases.

References were identified by searching articles in PubMed from Jan 1, 1995, until July 25, 2019, using the search terms "artificial intelligence OR "decision making, computer-assisted" OR Artificial Intelligence OR "Machine Learning" OR "Deep Learning" OR "Algorithm" OR "Algorithms" OR "latent variable model" OR "latent variable models AND "delivery of health care" OR "Healthcare" OR "health care" AND "ethics, clinical" OR "ethics, medical" OR "bioethics OR "clinical ethics" OR "medical ethics" OR "bioethics" OR "ethics" OR "ethical." This search produced 306 articles. 61 of these articles discussed clinical implementation of AI technologies and were included in the final reference list. 37 of additional references were identified through backward and forward searching from selected texts.

To capture nontraditional literature surrounding the topic of AI additional searches were completed using MEDLINE, ISI, Google Scholar, Web of Science, ProQuest Congressional, The Federal Register, and Congress.gov. and additional references added from these databases.