Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

The Joint Commission Journal on Quality and Patient Safety 2020; 46:596–599

Considering the Safety and Quality of Artificial Intelligence in Health Care

Patrick Ross, MPH; Kathryn Spates, JD, ACNP-BC

The role of machine learning and artificial intelligence (AI) in the health care sector is rapidly expanding. This expansion may be accelerated by the global spread of COVID-19, which has provided new opportunities for AI prediction, screening, and image processing capabilities.[1] Applications of AI can be as straightforward as using natural language processing to turn clinical notes into electronic data points or as complex as a deep learning neural network performing image analysis for diagnostic support. The goal of these tools is not to replace health care professionals, but to enable better patient experience and better inform the clinical decision-making process to improve the safety and reliability of clinicians.

Clinicians and health systems using these new tools should be aware of some of the key issues related to safety and quality in the development and use of machine learning and AI. The performance of a chatbot on a shopping website poses little harm to users, but AI used in health care, particularly clinical decision supports or diagnostic tools, can have significant impact on a patient's treatment. Inaccurate algorithm output or incorrect interpretation by clinicians could lead to significant adverse events for patients, such as inaccurate diagnoses, discriminatory clinical practices, private data leaks, or uses that generate profits for users at the expense of patient care.[2] We discuss key considerations for the safe development of AI tools, how to promote the safety and quality of care in AI implementation, and the transparency needs essential to building public trust.

The United States remains in the early stages of determining how AI will be regulated. The US Food and Drug Administration (FDA) has proposed a regulatory framework for the review and approval of AI devices, though much of this policy is still in a proposed or pilot stage.[3,4] The agency's intent is still clear: Manufacturers should have policies that ensure data are clinically relevant, acquired in a consistent manner, and sufficiently transparent to support the trust and understanding of data users.

SAFETY AND QUALITY ISSUES IN AI DEVELOPMENT

A primary limitation in the development of effective AI systems is the caliber of data.[5] Machine learning, the underlying technology for many AI tools, feeds data traits such as patient demographic data or disease state into an algorithm from large data sets to draw more accurate relationships between input traits and outcomes.[6] The limitation for any AI is that the program cannot exceed the performance level reflected in the training data. Accuracy and completeness in the training data contribute to the accuracy of the model, giving rise to the aphorism "garbage in, garbage out."

Ensuring an Accurate Ground Truth

Many AI models rely on training data that reflect the "ground truth," a best-case scenario in which researchers know the outcome in question, based on direct observation. Retroactively establishing the "ground truth" requires careful clinical review and annotation, which is a time- and resource-intensive process. However, the "ground truth" is not always easily determined. Clinicians may interpret cases differently or assign different labels for broadly defined conditions, leading to poor reproducibility.[7] Structured reporting in electronic health records (EHRs) can improve the availability of accurately annotated data.[8] Structured reporting avoids retrospective analysis by labeling information when it is collected, and the consistent language and notation format fosters the ability to convert structured text into data annotations.[9]

Relying on Health Records

Relying on EHRs is not always a viable option for a variety of reasons. Privacy protection statutes may limit information sharing, particularly for protected health information. This information is protected by the Health Insurance Portability and Accountability Act (HIPAA), which requires de-identification prior to sharing. Whether, and how, these data sharing agreements are disclosed will also affect the public's sense of trust and level of comfort with AI tools.[10] Balancing data security and business transparency practices can build patient trust in how health data are used in the development of AI tools.

In addition, records from health systems or plans may not be complete and may contain mistakes. As an example, a patient's primary diagnoses may be well-documented in his or her health record, while comorbidities or complications after a hospital stay are less accurately recorded.[11] Inaccurate data drawn from EHRs limit the accuracy of an algorithm's analysis.

1553-7250/$-see front matter

https://doi.org/10.1016/j.jcjq.2020.08.002

Threat of Selection and Implicit Biases

From data collection through testing and use, developers must carefully consider the threat of selection and implicit biases. The promise of machine learning lies in advanced pattern-finding capabilities, which also serve as a potential pitfall. Pattern recognition can lead a program to incorporate unintended human biases, such as racial, gender, or socioeconomic biases.[12–14] Biases incorporated into AI reflect current and historical disparities in clinical care and outcomes, and without careful review, stand to further entrench bias into the health care system, preserving worse outcomes for vulnerable populations. A widely deployed algorithm was recently shown to underselect Black patients for follow-up care. Instead of relying on clinical risk levels, the algorithm predicted needs based on projected levels of care spending, a historically disparate measure resulting from unequal access to care.[15] Other examples include natural language processing programs that recognized implicit associations between stereotypical male and female roles based on language contained in clinical records.[16] Preventing implicit bias in AI development requires a close understanding of what the training data is measuring. Testing algorithms for potential discriminatory outcomes regularly during development and across diverse health systems can reduce the risk of unintended bias in AI systems.[17]

Data Sourcing and Security

Data drawn from a single source may also present a risk for algorithms intended for application in large populations. Environmental exposure, social behavior, economic status, and racial and ethnic composition of the population all influence health outcomes and thus the associations learned by AI. Basing an algorithm on a single hospital or region may limit its generalizability.[18] In addition to geographic regions, developers should ensure that AI tools intended for widespread use are validated across multiple health care environments (for example, hospitals, long term care facilities, and so on).

Data silos in independent health systems discourage the exchange of health information among developers. Improved data sharing capabilities offer an opportunity to greatly expand the availability of data beyond a single site. In recent years, federal regulators have taken steps to require health information technology (IT) developers to facilitate data sharing through EHRs. As part of the Medicare payment system, the Promoting Interoperability Program has encouraged health information exchange capabilities through required reporting,[19] and the Centers for Medicare & Medicaid Services (CMS) and the Office of the National Coordinator for Health IT (ONC) recently issued rules aimed at standardizing data formats and curbing practices that would inhibit information sharing.[20,21]

Cloud-based computing platforms, which can store tremendous amounts of data, provide researchers an opportunity to store, exchange, and access data beyond their home organization. Such large, third-party repositories of annotated clinical data could provide a trustworthy, transparent source of data to developers, an idea supported by the FDA.[22] These platforms could mitigate the risks of single-site data, yet these data exchange goals necessitate additional security measures because electronic data—particularly health data—are a target for criminals. HIPAA requires that organizations take steps to ensure the security of protected health data, such as regular organizational audits and controlled access to protected information. Cloud-based platforms require additional encryption measures to prevent data loss, as data transmissions and storage are vulnerable to theft or unauthorized modification.[23] Encryption methods are continually adapting to meet the challenge of protecting against data breaches.

The fulfillment of interoperability efforts would provide researchers a secure means of health data exchange and foster rigorous AI development. High-quality output, and thus improved clinical care, requires high-quality input.

SAFETY AND QUALITY ISSUES IN AI USE

With more AI applications coming to market and health systems beginning to develop machine learning algorithms in-house, it is vital that providers take a thoughtful approach to implementing AI into the care process. Although the interplay between human work and digital information is not a new phenomenon, using AI–enabled clinical decision support magnifies the role software plays in the clinical planning process.

Automation Biases

"Deskilling" and "automation complacency" are two challenges to health care quality that can result from the increased emphasis on digital decision support. *Deskilling* is the loss of skills after a human task is automated.[24] This loss of skills can adversely affect clinician autonomy, reducing decision-making quality, diagnostic reasoning, and communication with patients.

Overreliance on automated decision support can also lead to automation complacency, as the human users or interpreters of AI output become overly reliant on AI support.[25] In decision making, confirmation bias leads clinicians to give inordinate weight to evidence that supports their prediction. Automation bias instead causes decision makers to stop looking for evidence after given a machine-generated output.[26] Complex tasks, such as diagnostic image interpretation, increase the likelihood of machine reliance, and studies have shown declines in human performance as a result of repeated use of computer-aided decision support.[27–29] Deskilling and automation biases can contribute to adverse events if complacent clinicians miss computer errors. The primacy of technological assessment may also affect examination skills as physicians lose the nu-

ance of unique patient histories and subjugate informed decision making to technology-dependent reasoning.[30]

Health care organizations should take steps to ensure that personnel who make decisions based on AI output have the training to recognize and prevent potential adverse safety events. Avoiding dependence on machine input is critical to preventing medical error. If an AI system is unable to make a prediction, relying on unready or untrained staff to make the same prediction is likely not an appropriate or safe work process. Work systems should be designed to empower personnel to speak up or flag unusual predictions for review. Clinicians should also participate in activities designed to maintain their skills and expertise to combat automation complacency.

Patient Relationships

Many clinicians hope that AI will free them to focus on patient interaction, but the popularization of computer-aided decision supports may just as well have negative effects on patient relationships. Research on the overreliance of technology in medicine has found that the increased use of EHRs has led to a prioritization of physician-technology interactions over physician-patient interactions, leading to decreased patient satisfaction, a scenario that could foreshadow the role of AI in patient care.[30] Patients also deserve to be informed of the use of AI in clinical care. The failure to disclose the use of AI tools to patients may diminish the trust between patients and clinicians.[31]

BUILDING PUBLIC TRUST BY INCREASING TRANSPARENCY AND UNDERSTANDING

Whether AI tools are readily adopted in health care settings depends on developing public trust of AI. Clinicians and patients must believe that AI applications are safe and effective, with a basic process that is explainable to users. Achieving this goal requires transparency from developers and clinical credibility for users.

An algorithm's clinical credibility demands reproducibility. Developers should be encouraged to provide ample transparency of methods, models, and data used in AI development to establish reproducibility. This transparency fosters the development of AI tools that are explainable. If an algorithm is bound by a set of rules, these conditions should be understandable to clinicians, and the results or predictions of an algorithm should make sense to providers using the tool.[32] Requiring manufacturers to publicly post understandable summaries of algorithms and updates as part of the FDA approval process has been proposed as one method of ensuring explainable AI systems.[33]

Transparent and explainable AI also helps combat the "black box" problem. In complex AI models, such as artificial neural networks, decision-making steps become opaque as multiple inputs are considered and weighted before arriving at an output. Between the input data and output prediction arises a "hidden layer" of computing, which makes it harder to spot errors or bias even as it expands the capabilities of machine learning. The "black box" of hidden layer connections can mask why a particular decision was made. Detecting architectural bias in a black box system is difficult and requires statistical analysis of the output, which can be aided with some technical solutions, such as "saliency maps" that identify the area of an image that was weighted the most in decision making.[34] Being able to determine what information was available for use by the software can help prevent errors. Enabling a learning opportunity for mistakes or near misses and real-world performance monitoring is critical. Additional research must be done on integrating developing technologies like saliency maps into learning and accountability programs. In the meantime, the use of AI tools in clinical practice should be controlled to scenarios in which tools behave predictably and are designed to fail safely, such as rejecting to make any predictions that do not meet a set threshold for confidence, and yield to human interpretation.[5]

Despite the potential for AI in health care to improve diagnosis or reduce human error, a failure in an AI program would affect large numbers of patients. Clinical uptake will ultimately be dependent on building a thorough evidence base to demonstrate safety and security.

CONCLUSION

Health care providers should be building their understanding of how machine learning algorithms are developed. Collecting large amounts of data has caused growth in new AI tools, but the success of these tools relies on ensuring that data are high quality: accurate, clinically relevant, and tested across multiple settings. While the regulatory structure for AI tools is being codified, health systems should examine how to safely introduce these new tools into their workflow. Carefully considering how to integrate AI tools can prevent adverse events and help deliver on the promise of machine learning.

Conflicts of Interest. The authors report no conflicts of interest.

Patrick Ross, MPH, is Policy and Regulatory Affairs Manager, The Joint Commission, Washington, DC. **Kathryn Spates, JD, ACNP-BC**, is Executive Director, Federal Relations, The Joint Commission, Washington, DC. Please address correspondence to Patrick Ross, pross@jointcommission.org.

References

1. McCall B. COVID-19 and artificial intelligence: protecting health-care workers and curbing the spread. Lancet Digit Health. 2020;2:e166–e167.

2. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. N Engl J Med. 2018 Mar 15;378:981–983.
3. US Food and Drug Administration. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)–Based Software as a Medical Device (SaMD): Discussion Paper and Request for Feedback, 2019. Accessed Aug 17, 2020. https://www.fda.gov/media/122535/download.
4. US Food and Drug Administration. Developing a Software Precertification Program: A Working Modelm, ver. 1, Jan 2019. Accessed Aug 17, 2020. https://www.fda.gov/media/119722/download.
5. Ellahham S, Ellahham N, Simsekler MCE. Application of artificial intelligence in the health care safety context: opportunities and challenges. Am J Med Qual. 2020;35:341–348.
6. Jiang F, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017 Jun 21;2:230–243.
7. Challen R, et al. Artificial intelligence, bias and clinical safety. BMJ Qual Saf.. 2019;28:231–237.
8. Willemink M, et al. Preparing medical imaging data for machine learning. Radiology. 2020;295:4–15.
9. European Society of Radiology (ESR) ESR paper on structured reporting in radiology. Insights Imaging. 2018;9:1–7.
10. Copeland R, Mattioli D, Evans M. Inside Google's quest for millions of medical records. Wall Street J. Epub. 2020 Jan 11.
11. Foley & Lardner LLP. AI Is Here to Stay: Are You Prepared?, Apr 2019. Accessed Aug 17, 2020. https://www.foley.com/-/media/files/insights/publications/2019/04/ai-is-here-to-stay–are-you-prepared–april-2019-r.pdf?la=en.
12. Zou J, Schiebinger L. AI can be sexist and racist—it's time to make it fair. Nature. 2018;559:324–326.
13. Larrazabal AJ, et al. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci U S A. 2020 Jun 9;117:12592–12594.
14. Koenecke A, et al. Racial disparities in automated speech recognition. Proc Natl Acad Sci U S A. 2020 Apr 7;117:7684–7689.
15. Obermeyer Z, et al. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019 Oct 25;366:447–453.
16. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. Science. 2017 Apr 14;356:183–186.
17. Gianfrancesco MA, et al. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med. 2018 Nov 1;178:1544–1547.
18. Nabi J. How bioethics can shape artificial intelligence and machine learning. Hastings Cent Rep. 2018;48(5):10–13.
19. Centers for Medicare & Medicaid Services. Promoting Interoperability Programs. (Updated: Jul 13, 2020.), 2020. Accessed Aug 17. https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms.
20. Centers for Medicare & Medicaid Services Medicare and Medicaid Programs; Patient Protection and Affordable Care Act; Interoperability and Patient Access for Medicare Advantage Organization and Medicaid Managed Care Plans, State Medicaid Agencies, CHIP Agencies and CHIP Managed Care Entities, Issuers of Qualified Health Plans on the Federally-Facilitated Exchanges, and Health Care Providers. Fed Regist. 2020 May 1;85:25510–25640.
21. Office of the National Coordinator for Health Information Technology 21st Century Cures Act: Interoperability, Information Blocking, and the ONC Health IT Certification Program. Fed Regist. 2020 May 1;85:25642–25961.
22. US Food and Drug Administration. FDA's Comprehensive Effort to Advance New Innovations: Initiatives to Modernize for Innovation. Gottlieb S, editor, 2020. Aug 29, 2018. Accessed Aug 17. https://www.fda.gov/news-events/fda-voices-perspectives-fda-leadership-and-experts/fdas-comprehensive-effort-advance-new-innovations-initiatives-modernize-innovation.
23. Al-Issa Y, Ottom MA, Tamrawi A. eHealth cloud security challenges: a survey. J Healthc Eng. 2019 Sep 3;2019 7516035.
24. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA. 2017 Aug 8;318:517–518.
25. Macrae C. Governing the safety of artificial intelligence in healthcare. BMJ Qual Saf. 2019;28:495–498.
26. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc. 2017 Mar 1;24:423–431.
27. Sujan M, et al. Human factors challenges for the safe use of artificial intelligence in patient care. BMJ Health Care Inform. 2019;26:e100081.
28. Alberdi E, et al. Effects of incorrect computer-aided detection (cad) output on human decision-making in mammography. Acad Radiol. 2004;11:909–918.
29. Lyell D, et al. Automation bias in electronic prescribing. BMC Med Inform Decis Mak. 2017 Mar 16;17:28.
30. Lu J. Will medical technology deskill doctors? International Education Studies.. 2016;9(7):130–134.
31. STAT. An Invisible Hand: Patients Aren't Being Told About the AI Systems Advising Their Care. Robbins R, Brodwin E, editors, Jul 15, 2020. Accessed Aug 17, 2020. https://www.statnews.com/2020/07/15/artificial-intelligence-patient-consent-hospitals/.
32. Jamieson T, Goldfarb A. Clinical considerations when applying machine learning to decision-support tasks versus automation. BMJ Qual Saf. 2019;28:778–781.
33. Hwang TJ, Kesselheim AS, Vokinger KN. Lifecycle regulation of artificial intelligence- and machine learning-based software devices in medicine. JAMA. 2019 Dec 17;322:2285–2286.
34. Guidotti R, et al. A survey of methods for explaining black box models. ACM Comput Surv. 2018;51:93.