EDITORIAL

Randomized Clinical Trials of Artificial Intelligence

Derek C. Angus, MD, MPH

As patient data are increasingly captured digitally, the opportunities to deploy artificial intelligence (AI), especially machine learning, are increasing rapidly. Machine learning is automated learning by computers using tools such as artificial neural networks to search data iteratively for optimal solutions.[1] Typical applications include searching for novel patterns (eg, latent cancer subtypes[2]), making a diagnosis or outcome prediction (eg, diabetic retinopathy[3]), and optimizing treatment decisions (eg, fluid and vasopressor titration for septic shock[4]). Although many express excitement regarding the promise of AI, others express concern about adverse consequences, such as loss of physician and patient autonomy or unintended bias, and still others claim that the entire endeavor is largely hype, with virtually no data that actual patient outcomes have improved.[5,6]

Related article

One issue complicating this debate is that the classic measure of clinical benefit, the randomized clinical trial (RCT), is rare in this field, if not entirely absent. For example, the US Food and Drug Administration recently approved AI-enabled decision support tools (also called software as medical devices or SaMDs) for diagnosis of diabetic retinopathy on digital fundoscopy and early warning of stroke on computed tomography scans.[7,8] In neither instance was approval based on any RCT evidence that the information provided by the SaMD improved care. Of course, diagnostic tools entered clinical practice for decades without the requirement for RCT data. However, these tools were traditionally framed as providing data, leaving judgment to physicians. In contrast, AI can provide data and judgment, thus altering clinician actions much more substantially.[9] Against this backdrop, in this issue of *JAMA*, the report by Wijnberge et al[10] of a clinical trial of an AI-derived clinical decision support tool provides important insight into the way RCTs might inform the debate on AI in health care.

The study, somewhat ironically entitled the Hypotension Prediction During Surgery (HYPE) trial, randomized 68 patients who were undergoing elective noncardiac surgery to intraoperative management guided by an AI-based early warning system (intervention group) or standard care (control group). The primary objective was to test whether the intervention reduced the depth and duration of intraoperative hypotension (calculated as a time-weighted average). Both groups were managed with continuous radial artery pressure monitoring. In the intervention group, the arterial pressure was monitored by a SaMD that used 23 waveform variables, extracted continuously, to provide an updated prediction every 20 seconds of the likelihood of intraoperative hypotension (defined as mean arterial pressure <65 mm Hg) in the subsequent 15 minutes. The device issued an alarm when the risk exceeded 85% and encouraged the anesthesiologist to take preemptive action. The device displayed the risk score and a read-out of key variables (eg, stroke volume) used by the algorithm. The investigators also educated the anesthesiologists about the features of the device and provided a protocol to aid interpretation together with training on suggested actions to take (eg, intravenous fluid bolus or infusion of vasoactive agent) when the algorithm generated an alarm. In the control group, arterial wave data still flowed to the AI algorithm to allow it to run in silent mode, but only routine pulse and blood pressure data were displayed, as per standard care.

The trial demonstrated that the intervention successfully reduced patients' exposure to hypotension, as evidenced by the primary outcome of a lower time-weighted average of intraoperative hypotension (0.1 vs 0.44 mm Hg; median difference, 0.38 mm Hg; 95% CI, 0.14-0.43 mm Hg; $P < .001$), fewer episodes of hypotension per patient (3 vs 8 episodes per patient; median difference, 4 episodes; 95% CI, 1-7 episodes; $P = 004$), and fewer mean minutes with hypotension (8.0 vs 32.7 minutes; median difference, 16.7 minutes; 95% CI, 7.7-31 minutes; $P < .001$). Intraoperative management was also different between groups. The AI algorithm provided warnings frequently, including 377 alarms lasting longer than 1 minute (12 alarms per patient), and prompted anesthesiologists to initiate treatment within 2 minutes on 304 occasions (81%). Across several prespecified and post hoc analyses, anesthesiologists' behaviors differed between groups. Broadly speaking, anesthesiologists in the intervention group acted more often, acted sooner, and selected different treatments. There are several notable features about this trial.

First, the problem the authors addressed is ideal for machine learning. Millions of patients require anesthesia every year, during which hypotension is both common and associated with adverse sequelae.[11,12] High-fidelity continuous blood pressure monitoring is routinely collected but potentially underused because of the limits of human cognition. The algorithm, generated and published previously, was built from an automated search across arterial waveform data from more than 1000 patients, exploring the potential contribution of more than 3000 waveform features in more than 2.6 million feature combinations.[13] The final model predicts the likelihood of future hypotension via measurement of multiple variables characterizing dynamic interactions between left ventricular contractility, preload, and afterload. Although clinicians can look at arterial pulse pressure waveforms and, in combination with other patient features, make

educated guesses about the possibility of upcoming episodes of hypotension, the likelihood is high that an AI algorithm could make more accurate predictions.

Nevertheless, in the study by Wijnberge et al, all learning was performed before the first patient was enrolled in the trial (a so-called locked algorithm), precluding 2 possible opportunities for further improvement. First, when data are organized in multiple dimensions, such as the case in this study, even very large data sets can be too sparse (the so-called curse of dimensionality[14]), so the accrual of more patients could permit improved calibration. Second, the algorithm may perform better for each individual patient if it can learn during deployment, akin to the voice recognition algorithms that train and calibrate to individual users. However, the problem with an algorithm that continues to learn is that the RCT quickly becomes much more complicated to design and interpret. Although Food and Drug Administration approvals have largely been limited to locked algorithms, the agency is aware of the need to establish an approval path for continuously learning SaMDs.[15]

A second feature of this study is the effort involved. Much work already had been invested in the development of the algorithm and construction of the hardware and software to provide an adequately reliable system for data acquisition, processing, and display. The investigators then further nested this warning system into a care setting in which anesthesiologists were instructed in the use of the device and given guidance on actions to take when prompted. Rather than hope that the anesthesiologists would know what to do when presented with an alarm of pending hypotension and information about left ventricular hemodynamics, the authors provided guidelines to encourage specific actions. Providing this education and treatment guideline may well have been critical in engaging physicians and in directly converting the information provided by the SaMD into actions that mitigated hypotension. However, the magnitude of this contribution is unclear, and scaling this intensity of deployment may be challenging.

Third, the experiment, in part because of the issues discussed above, was constrained, providing an estimate of efficacy on an intermediate end point, a process outcome, in a highly controlled setting. The sample size was too small to evaluate the effects on subsequent patient-centered outcomes or overall safety. Because the study was limited to a single hospital, it is unknown if the willingness of anesthesiologists to follow the advice of the SaMD is generalizable and how much training and education would be required to achieve a similar effect. The representativeness of care in the control group, which affects the magnitude of incremental benefit, is also unknown. By linking the SaMD alarms to specific training on interpretation and recommended actions, it is unknown if similar benefits could be achieved if other actions were taken and if some actions were more effective than others. In addition, by studying so few patients, there is no information on whether the SaMD or the recommended actions generate consistent treatment effects across patients or whether there is important heterogeneity in the fidelity of the predictions or efficacy of the recommended interventions.

In other words, there is good news and challenging news. The study by Wijnberge et al provides an excellent and arguably unique example of a proof-of-concept trial demonstrating the causal effects of an AI-enabled care strategy. In many respects, this investigation resembles a positive phase 2b drug trial, providing valuable information regarding the intervention's ability to influence physician actions and change proximate patient outcomes. To fully evaluate the importance of machine learning, more studies like this should be conducted. However, positive phase 2b drug trials are usually followed by a phase 3 trial in which the major change is just a larger sample size to determine effects on patient-centered outcomes. Investigators typically take care to keep as many other features of the trial identical to increase the odds of success. But, as can be seen from the study by Wijnberge et al, the questions that arise will not be addressed by simply increasing sample size. Rather, there remain multiple dimensions of uncertainty, more akin to the types of questions more typical in evaluation of highly complex health care delivery interventions.

Specifically, at least 6 questions could routinely arise after initial proof-of-concept demonstration of a SaMD intended to augment clinical decision-making: (1) Has the machine learned enough? Or should the intervention be allowed to continue learning (and improving)? (2) Is the supporting suite of implementation strategies (eg, hardware configuration and reliability, information display, or user education) optimized? Or are some elements redundant or missing? (3) Is the information provided by the SaMD of homogeneous accuracy and utility? Or do some instructions "work" better than others? (4) What characteristics of the health care delivery environment (eg, clinician knowledge and attitudes, existing care patterns) influence the incremental benefit? (5) What characteristics of the patient population influence the incremental benefit? (6) How do these features interact to influence the effect of the SaMD on the proximate (eg, intraoperative hypotension) and more important distal (eg, postoperative recovery) patient outcomes?

These questions are multidimensional and interrelated. Just as computer scientists wrestle with the curse of dimensionality when generating an AI algorithm, clinical investigators will wrestle with a second curse of dimensionality during evaluation of the utility of the algorithm. Almost certainly, such evaluation will stretch and possibly overwhelm standard RCT designs. However, randomization remains an invaluable tool in the causal inference tool set. The near total lack of RCTs of AI to date has been a glaring deficit. As the number of potential SaMD applications grows, studies like that of Wijnberge et al will be crucial in guiding the deployment of AI in health care.

ARTICLE INFORMATION

Author Affiliations: University of Pittsburgh, Pittsburgh, Pennsylvania; Associate Editor, *JAMA*.

Corresponding Author: Derek C. Angus, MD, MPH, University of Pittsburgh, 3550 Terrace St, 614 Scaife Hall, Pittsburgh, PA 15261 (angusdc@upmc.edu).

Published Online: February 17, 2020. doi:10.1001/jama.2020.1039

Conflict of Interest Disclosures: None reported.

REFERENCES

1. Hinton G. Deep learning—a technology with the potential to transform health care. *JAMA*. 2018;320 (11):1101-1102. doi:10.1001/jama.2018.11100

2. Li A, Walling J, Ahn S, et al. Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes. *Cancer Res*. 2009;69(5):2091-2099. doi:10.1158/0008-5472.CAN-08-2100

3. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402-2410. doi: 10.1001/jama.2016.17216

4. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med*. 2018;24(11):1716-1720. doi:10. 1038/s41591-018-0213-5

5. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA*. 2019; 322(24):2377-2378. doi:10.1001/jama.2019.18058

6. Emanuel EJ, Wachter RM. artificial intelligence in health care: will the value match the hype? *JAMA*. 2019;321(23):2281-2282. doi:10.1001/jama.2019.4914

7. 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;1:39. doi: 10.1038/s41746-018-0040-6

8. US Food and Drug Administration. FDA permits marketing of clinical decision support software for alerting providers of a potential stroke in patients. https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-clinical-decision-support-software-alerting-providers-potential-stroke. Published February 13, 2018. Accessed January 2020.

9. Maddox TM, Rumsfeld JS, Payne PRO. Questions for artificial intelligence in health care. *JAMA*. 2019; 321(1):31-32. doi:10.1001/jama.2018.18932

10. Wijnberge M, Geerts BF, Hol L, et al. Effect of a machine learning–derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial [published online February 17, 2020]. *JAMA*. doi:10.1001/jama.2020.0592

11. Holmer H, Bekele A, Hagander L, et al. Evaluating the collection, comparability and findings of six global surgery indicators. *Br J Surg*. 2019;106(2):e138-e150. doi:10.1002/bjs.11061

12. Davies SJ, Vistisen ST, Jian Z, Hatib F, Scheeren TWL. Ability of an arterial waveform analysis–derived hypotension prediction index to predict future hypotensive events in surgical patients. *Anesth Analg*. 2020;130(2):352-359. doi: 10.1213/ANE.0000000000004121

13. Hatib F, Jian Z, Buddi S, et al. Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis. *Anesthesiology*. 2018;129(4):663-674. doi:10.1097/ ALN.0000000000002300

14. Bellman R. *Dynamic Programming*. Princeton University Press; 1957.

15. US Food and Drug Administration. Digital health. https://www.fda.gov/medical-devices/digital-health. Published 2019. Accessed January 2020.