Editorial

eISSN 2005-8330
https://doi.org/10.3348/kjr.2023.1016
Korean J Radiol 2024;25(2):123-125

Korean Journal of Radiology
KJR

Artificial Intelligence for Improved Patient Outcomes—The Pragmatic Randomized Controlled Trial Is the Secret Sauce

Daniel W. Byrne, Henry J. Domenico, Ryan P. Moore

The Advanced Vanderbilt Artificial Intelligence Laboratory (AVAIL), Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA

Keywords: Artificial intelligence; Outcome; Pragmatic trials; Study design and best practice; Clinical evaluation; Clinical research; Randomized clinical trial; Real-world evidence; Research method

Artificial Intelligence (AI) has exploded in the media for both its astonishing power and disturbing weaknesses. Of the potential applications of AI that are most likely to benefit society, most thought leaders point to medicine. Yet, to date, we have almost no rigorous evidence that AI improves patient health outcomes [1-3]. Why is there a dearth of evidence? What needs to change?

First, let's look at what has not been working. Most applications of AI in healthcare have had no outcome evaluation or had one with an inadequate study design that would not result in reproducible research [4]. Many AI evaluations are based on observational studies that are so profoundly biased that they provide no compelling evidence that the AI tool is safe or improves patient outcomes [4]. This is compounded by the problem that some junior AI researchers are new to clinical research and are unaware that their outcome evaluations are fatally flawed. Those in healthcare making the decisions about purchasing or implementing AI tools sometimes lack both the skills to fully understand the nuances of the AI model and to recognize the statistical flaws in the evaluation. Despite these issues, AI is often seen as a "shiny new object" with great potential and given a free pass. Many flawed AI evaluations are published with surprisingly little criticism—a sign that the self-correcting aspect of science is failing in this unique niche—at least in the short term during this honeymoon phase [2,4].

To understand why these studies are flawed, we need to understand the central issue: with weak, non-randomized evaluations, the two groups of patients—usual care vs. AI intervention groups—differ systematically (non-randomly) on characteristics affecting the outcome. To make cause-effect conclusions about the AI application improving patient outcomes, differences in all other factors need to be due to chance alone or adequately adjusted. This is rarely the case in observational studies or weak study designs, such as before-after or poorly planned stepped-wedge designs. A common misconception is that adjusting for potential confounding variables will solve this problem and eliminate treatment selection bias and residual confounding. Often, this is not true because the required variables are not recorded, available, or even known to the researchers. In most cases, rather than listing residual confounding as a limitation in the paper and then proceeding to use the model for patient care, the residual confounding should be treated as a fatal flaw, and the model should be tested with randomization before being used in patient care.

AI start-ups are increasingly creating tools that are being

Received: October 16, 2023 **Revised:** October 17, 2023
Accepted: October 23, 2023

Corresponding author: Daniel W. Byrne, MS, The Advanced Vanderbilt Artificial Intelligence Laboratory (AVAIL), Department of Biostatistics, Vanderbilt University Medical Center, 2525 West End Avenue, Suite 11000, Nashville 37203-1741, TN, USA
• E-mail: daniel.byrne.research@gmail.com

used to treat patients and making claims about the benefits that are not supported by rigorous evidence. They often fail to provide the transparency to understand how their "proprietary" black box models work. And yet, with hand-waving, slick marketing, and hype, they raise millions of dollars from venture capitalists. Some will say that this positive spin and exaggeration in Silicon Valley and the AI field are common and harmless. However, Elizabeth Holmes, the founder of Theranos, was recently sent to prison for 11 years for following this approach—and taking it to the extreme.

What needs to change? In one word: "Randomization." The excuses for why randomization is not an option are plentiful and roll off the tongue, yet in our experience, the excuses frequently turn out to be unjustified [2,5,6]. The reason that we have safe and effective prescription drugs in the US is that the US Food and Drug Administration (FDA) demands evidence from rigorous randomized controlled trials (RCTs) and would never consider the excuses used for not randomizing that are tolerated in the AI field.

A traditional RCT may not be an option in certain cases, primarily due to patient safety and ethical considerations, as well as cost and logistical constraints. Pragmatic trial designs, however, offer a solution to mitigate most of these impediments. Some will argue that the RCT will take many years to complete, and with the fast-moving developments in AI, informatics, and technology, this is an unacceptable pace. We have demonstrated, however, that pragmatic RCTs can be performed rapidly [6-9] and are a superior option to weaker study designs. Speed and rigor are not mutually exclusive; in fact, rigor nearly always increases the speed of introducing new methods to medicine in a responsible and ethical way.

Using AI algorithms to improve patient health outcomes can be viewed as progress on the "last mile" between a model creation and improved outcomes. Building an AI model that works is only the first step. The application must be accepted and seamlessly integrated into the healthcare workflow in a frictionless manner. This work requires a collaborative multidisciplinary team of AI experts, physicians, nurses, informaticians, biostatisticians, and others. AI tools should be evaluated in the healthcare workflow using pragmatic patient-level RCTs. Because successful implementation will likely require several iterations, an adaptive platform trial is optimal; an adaptive platform trial enables multiple interventions to be tested in parallel, dropping study arms for futility and adding new arms in a perpetual manner without long delays between studies [2]. For example, various forms of escalation can be assessed regarding changing clinician behavior.

Financial, career, and training incentives also need to change before we will see AI improving patient outcomes. Specifically, the incentives need to be modified to reward those who perform rigorous AI research. The current incentives reward creating new but redundant models and promoting them despite minimal evaluations. For example, there are more than 100 publications describing predictive models for hospital readmissions—but not one has been shown, with a rigorous study design, to reduce readmissions.

Generative AI, such as ChatGPT and GPT-4, has provided the medical community with impressive tools that can be used to assist with writing emails, patient summaries, and many other applications. Outside of medicine, these models are fairly stable (although imperfect) because they are trained on text that is, for the most part, factual. These models are unreliable for improving patient outcomes because they are trained on the medical literature, which is a scientific debate rather than a set of facts [2]. In some areas of medicine, there are many papers that report incorrect conclusions. The models do not critically interpret the medical literature the way a trained medical expert would. More alarmingly, these tools will create "facts" and support them by fabricating studies and medical publications that do not exist. For example, asking ChatGPT whether AI had been used to improve patient outcomes resulted in "hallucinations" or more accurately fabrications. Until these issues are resolved, other AI approaches, such as predictive and classification models, are more likely to improve patient outcomes.

By using pragmatic RCTs to know when AI has moved the needle regarding outcomes, healthcare can move more rapidly into a modern era of AI that benefits both patients and clinicians. Rather than comparing the performance of AI models vs. physician decisions, future trials should measure the performance of physicians alone versus physicians assisted by AI models.

In 2016, the "godfather of AI," Geoffrey Hinton, said, "People should stop training radiologists now. It is just completely obvious that within five years, deep learning is going to do better than radiologists." Perhaps this should be updated to "People should start training physicians to create, use, and evaluate AI tools in a modern and rigorous manner." This training needs to include AI evaluation skills, such as regression to the mean, reverse causation, issues with overall accuracy, residual confounding, pragmatic patient-level RCTs, and adaptive platform trials.

Once we embrace reproducible research in the form of pragmatic RCTs with prespecified study designs and endpoints, AI will begin to improve patient outcomes—overall and within important subgroups. AI also has great potential to reduce the workload and burnout among physicians. Despite the concerns, “AI won’t replace doctors, but doctors who use AI will replace doctors who don’t.” The pragmatic RCT is the secret sauce—but only if physician-scientists with modern training lead these projects.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

Author Contributions

Conceptualization: all authors. Writing—original draft: all authors. Writing—review & editing: all authors.

ORCID IDs

Daniel W. Byrne
https://orcid.org/0000-0001-9330-4334
Henry J. Domenico
https://orcid.org/0000-0001-7390-7947
Ryan P. Moore
https://orcid.org/0000-0003-2883-9580

Funding Statement

None

REFERENCES

1. Park SH, Choi JI, Fournier L, Vasey B. Randomized clinical trials of artificial intelligence in medicine: why, when, and how? *Korean J Radiol* 2022;23:1119-1125
2. Byrne DW. *Artificial intelligence for improved patient outcomes: principles for moving forward with rigorous science.* 1st ed. Baltimore: Wolters Kluwer Health, 2022
3. Topol E. *Deep medicine: how artificial intelligence can make healthcare human again.* 1st ed. New York: Basic Books, 2019:309
4. Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Systematic review finds “spin” practices and poor reporting standards in studies on machine learning-based prediction models. *J Clin Epidemiol* 2023;158:99-110
5. Walker SC, French B, Moore R, Domenico HJ, Wanderer JP, Balla S, et al. Use of a real-time risk-prediction model to identify pediatric patients at risk for thromboembolic events: study protocol for the children’s likelihood of thrombosis (CLOT) trial. *Trials* 2022;23:901
6. Walker SC, French B, Moore RP, Domenico HJ, Wanderer JP, Mixon AS, et al. Model-guided decision-making for thromboprophylaxis and hospital-acquired thromboembolic events among hospitalized children and adolescents: the CLOT randomized clinical trial. *JAMA Netw Open* 2023;6:e2337789
7. Semler MW, Self WH, Wanderer JP, Ehrenfeld JM, Wang L, Byrne DW, et al. Balanced crystalloids versus saline in critically Ill adults. *N Engl J Med* 2018;378:829-839
8. Yiadom MYAB, Domenico HJ, Byrne DW, Hasselblad M, Kripalani S, Choma N, et al. Impact of a follow-up telephone call program on 30-day readmissions (FUTR-30): a pragmatic randomized controlled real-world effectiveness trial. *Med Care* 2020;58:785-792
9. Noto MJ, Domenico HJ, Byrne DW, Talbot T, Rice TW, Bernard GR, et al. Chlorhexidine bathing and health care-associated infections: a randomized clinical trial. *JAMA* 2015;313:369-378

https://doi.org/10.3348/kjr.2023.1016