Editorial

eISSN 2005-8330
https://doi.org/10.3348/kjr.2022.0834
Korean J Radiol 2022;23(12):1119-1125

Korean Journal of Radiology
KJR

Randomized Clinical Trials of Artificial Intelligence in Medicine: Why, When, and How?

Seong Ho Park[1], Joon-Il Choi[2], Laure Fournier[3], Baptiste Vasey[4]

[1]Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea; [2]Department of Radiology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Korea; [3]Department of Radiology, Université Paris Cité, AP-HP, Hôpital Européen Georges Pompidou, PARCC UMRS 970, INSERM, Paris, France; [4]Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK

Take-home points

- One of the ultimate purposes of using artificial intelligence (AI) in clinical practice is to improve patient care outcomes. Therefore, evaluating the effect of AI on clinical outcomes beyond diagnostic/predictive performance, ideally using randomized clinical trials (RCTs), is important because improvements in diagnostic/predictive performance alone do not guarantee improved clinical outcomes.
- However, RCTs may not always be a practical design, albeit methodologically ideal, for generating evidence to guide the use of AI in real-world practice.
- Understanding when RCTs are most appropriate and feasible for the clinical evaluation of AI systems is helpful.
- Reporting guidelines for early-stage clinical evaluation of AI systems and RCTs of AI interventions have recently been published, including DECIDE-AI, CONSORT-AI, and SPIRIT-AI, which also help to overview the necessary methodological elements.

The best way to determine the actual clinical value of an artificial intelligence (AI) system is to see whether the use of the AI improves clinical outcomes compared to when the system is not used, for which a randomized clinical trial (RCT) is the ideal design [1,2]. This approach is increasingly recognized for clinical AI [3-7] from a scientific perspective and because demonstrating improved health outcomes is desirable in value-based healthcare [8], echoing a similar practice in drug development. So far, multiple RCTs on AI have been published, although they remain uncommon compared to the number of preclinical AI studies [7,9-15].

Why?

One of the ultimate purposes of using AI in clinical practice is to improve patient care outcomes. Therefore, evaluating the effect of AI on clinical outcomes beyond diagnostic or predictive performance is crucial for assessing the actual clinical utility of an AI system. Indeed, diagnostic or predictive performance improvements do not guarantee improved clinical outcomes for various reasons, including but not limited to those listed below [1,15]. First, overlapping diagnostic/predictive information is often present in real-world patient care, for example, the physician's evaluation of symptoms and physical signs, various laboratory test results, radiologic images, or psychosocial factors. Therefore, the results from a

Received: October 29, 2022 **Accepted:** October 30, 2022
Corresponding author: Seong Ho Park, MD, PhD, Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Korea.
• E-mail: parksh.radiology@gmail.com

particular AI system may not influence subsequent patient management as much as when the method provides only available information. Second, small increases in diagnostic/predictive efficacy created by AI may be diluted or even vanish in the chain of multiple steps involved in patient care that follow a diagnostic/predictive decision. Third, because the final clinical outcomes are achieved through preventive or therapeutic actions taken based on diagnostic/predictive decisions, if no effective preventive or therapeutic actions are available or taken timely, no effects would be made on the clinical outcomes. Fourth, preventive or therapeutic interventions can also bring about adverse effects and added burden on the healthcare system, and improved accuracy may not always be defined the same from a technical and clinical perspective. For example, a higher recall rate in mammography screening can disproportionately increase the number of unnecessary biopsies compared with the number of years of life saved thanks to additionally detected cancers.

Practical Challenges

However, although methodologically ideal, RCTs may not always be practical for generating real-world evidence about medical devices [16]. RCTs are time- and resource-intensive. Given the number and diversity of AI systems proposed continuously, conducting full-fledged RCTs for each would overload healthcare centers' academic and clinical capacity. RCTs are not particularly suited to evaluating evolving interventions, which could be true for self-learning AI systems for example. RCTs conducted in a controlled environment also have the limitation of not always presenting real-world practice well [17,18]. Furthermore, RCTs might not be the best design to identify risks, which remains one of the main concerns in medical device regulation.

In some cases, non-randomized cohort studies that compare AI-assisted care and AI-unassisted conventional care or monitor adverse event occurrences in the long term or diagnostic/predictive performance studies, can be more reasonable designs for clinical evaluation. Indeed, demonstrating improved diagnostic/predictive performance or safety alone, yet without proof of improved patient outcomes, can be meaningful in some clinical scenarios, and the results provide higher confidence in the systems and clarity for patient care while making better use of the available resources. Such studies can be conducted more effectively using various other designs instead of RCTs [19]. Therefore, while the general importance of high-quality RCTs before the clinical adoption of AI should be appreciated, taking RCTs indiscriminately as "the Holy Grail" of clinical evaluation of AI would not be a practical approach.

When are RCTs More Appropriate for the Clinical Evaluation of AI Systems?

In this context, understanding when RCTs are more appropriate to create the necessary evidence for the clinical adoption of AI systems is helpful. Below is a non-exhaustive list of situations where RCTs should primarily be used for the clinical evaluation of AI systems.

When Patient Outcomes are More Relevant Parameters than Performance Metrics

For example, one area in which AI systems are expected to play a critical role is the assessment of acute ischemic stroke [20]. Acute ischemic stroke is a time-sensitive, high-stakes clinical scenario that requires a rapid approach to facilitate the hyperacute evaluation and management of patients. AI performing automated triage functions can be used to automatically identify suspected large vessel occlusion (LVO) strokes on head computed tomography angiography simultaneously with image acquisition and immediately trigger alerts to the neurovascular team [21]. In this regard, AI may improve clinical outcomes substantially as it enables more patients with LVO to be identified and receive endovascular therapy within a golden time compared with conventional care, resulting in better neurological recovery. Provided it does not have an unnecessary burden on the clinical workflow, even an AI with moderate accuracy may improve patient outcomes, as it would expedite patient management for at least some patients, even if it is a small number. Although it is true that the better the AI accuracy, the greater the positive effects expected on patient outcomes with fewer false emergency call-ups, the clinical benefit of the AI system is difficult to grasp based solely on the performance metrics (such as the area under the receiver operating characteristic curve, sensitivity, and specificity), unlike the direct assessment of patient outcomes.

When AI Systems Predict the Risk for Future Adverse Events Influenced by Clinical Care

For example, one study evaluated an AI system that

https://doi.org/10.3348/kjr.2022.0834

Korean Journal of Radiology
KJR

continuously analyzes arterial pressure waveform during surgery and warns if hypotensive events are expected within the next 15 minutes [22]. This RCT showed that using AI compared to standard care resulted in less intraoperative hypotension. The primary purpose of AI systems in this category is to determine if a patient is at high or low risk for developing adverse events in the future and, thereby, direct an intervention to prevent such events from occurring. In this context, the clinical evaluation of AI systems should primarily focus on whether AI reduces the occurrence of adverse events compared with conventional care, which ideally requires RCTs, rather than investigating the predictive performance of AI per se. Indeed, the clinical value of these systems depends on more factors than accuracy alone, such as timeliness, the existence of mitigation measures, and user trust in the system recommendation. Moreover, the accuracy of AI systems for time-to-event predictions of future events, as evaluated using relevant performance metrics [23], is often less impressive than that for static diagnosis because predicting future events is generally more difficult. However, the relatively low-performance results may not necessarily indicate a lack of clinical benefits. If the use of predictive AI can reduce the occurrence of adverse events without impeding workflows, AI is clinically valuable.

When Repeated Diagnostic/Predictive Assessment of the Same Patient is Not Possible

Another promising clinical application of AI is colorectal neoplasia detection assistance during colonoscopy. The adenoma detection rate is a key outcome parameter in this setting, which requires removal or biopsy of lesions detected during colonoscopy for pathological confirmation. Therefore, once a colonoscopy is performed either with or without AI and the detected lesions are removed, the patient is no longer eligible for examination with the other method. Therefore, comparing AI-assisted and conventional care generally requires two parallel groups for which RCTs are ideal. Consequently, multiple RCTs have been conducted to evaluate the effects of AI software tools to assist colonoscopy in detecting colorectal neoplasia [11-14].

When Authors Make Summative Claims About an AI Intervention's Effectiveness in Improving Patient Care

Inflated claims are a recurrent problem in evaluating clinical AI [24]. RCTs remain the gold standard for evaluating the effectiveness of an intervention to improve clinical outcomes, and authors making claims of improved effectiveness must have appropriate evidence to back them. Non-randomized controlled studies, of course, have a place in the early stage of clinical evaluation, but their results remain part of the formative evaluation and should be primarily used to inform the design of robust RCTs rather than making summative claims about effectiveness.

Reporting Guidelines for RCTs of AI Intervention

Reporting guidelines for RCTs of AI and their preparatory phase have recently been published, including CONSORT-AI, SPIRIT-AI, and DECIDE-AI [3,4,25]. These guidelines can be helpful beyond reporting as they provide an overview of the necessary methodological elements to consider. These guidelines represent a good source of information about the current expert consensus on AI scientific evaluation requirements. CONSORT-AI (for trial reports) and SPIRIT-AI (for trial protocols) are guidelines dedicated to RCTs that evaluate interventions with an AI component. DECIDE-AI can be applied to various study designs, including pilot RCTs, and focuses on the early, small-scale, and live clinical evaluation of AI systems as a stepping stone toward methodologically robust RCTs [25]. DECIDE-AI is unique because it is a stage-specific guideline highlighting the importance of a staged approach to complex intervention evaluation [26,27]. Indeed, many important parameters for designing a robust RCT need to be investigated before the start of the trial and should be determined during small-scale preparatory studies. These guidelines contain both general items that are universally applicable to all RCTs/pilot RCTs and AI-specific items. The AI-specific items in the checklists are summarized in Table 1, many of which are shared by the three guidelines.

CONCLUSION

It is important to evaluate the effect of AI on clinical outcomes, ideally using RCTs, beyond diagnostic/predictive performance. However, despite their superior methodological quality, RCTs may not be practical for the clinical evaluation of many AI systems. This article summarizes some key arguments as to why RCTs are needed in evaluating clinical AI and when they should be favored over other evaluation forms. Recent developments in reporting guidelines for RCTs of AI and their preparatory phase have also been

 https://doi.org/10.3348/kjr.2022.0834

Table 1. AI-Specific Items in the Checklists of CONSORT-AI, SPIRIT-AI, and DECIDE-AI [3,4,25]

	CONSORT-AI and SPIRIT-AI	DECIDE-AI
Title and Abstract*	• Indicate that the intervention involves AI/ML in the title and/or abstract and specify the type of model • State the intended use of the AI intervention within the trial in the title and/or abstract	Title: • Identify the study as early clinical evaluation of a decision support system based on AI or ML, specifying the problem addressed
Introduction	• Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (e.g., healthcare professionals, patients, public) • Describe any pre-existing evidence for the AI intervention†	Intended use: • Describe the targeted medical condition(s) and problem(s), including the current standard practice, and the intended patient population(s) • Describe the intended users of the AI system, its planned integration in the care pathway, and the potential impact, including patient outcomes, it is intended to have
Methods	Eligibility criteria, participants, and study setting: • Inclusion and exclusion criteria at the level of participants • Inclusion and exclusion criteria at the level of the input data • Onsite or offsite requirements needed to integrate the AI intervention into the trial setting Intervention: • Version of the AI algorithm used • Procedure for acquiring and selecting the input data for the AI intervention • Procedure for assessing and handling poor quality or unavailable input data • Any human-AI interaction in handling of the input data (e.g., selection of ROI) and level of expertise required for users • Output of the AI intervention • How the AI intervention's outputs contributed to decision-making or other elements of clinical practice, i.e., how humans interact with AI (e.g., how AI results are presented to, interpreted by, and acted on by human practitioners) Harms: • Specify any plans to identify and analyze performance errors, or if there are no plans for this, justify why not†	Participants: • Describe how patients were recruited, stating the inclusion and exclusion criteria at both patient and data level, and how the number of recruited patients was decided • Describe how users were recruited, stating the inclusion and exclusion criteria, and how the intended number of recruited users was decided • Describe steps taken to familiarize the users with the AI system, including any training received prior to the study Al system: • Briefly describe the AI system, specifying its version and type of underlying algorithm used. Describe, or provide a direct reference to, the characteristics of the patient population on which the algorithm was trained and its performance in preclinical development/validation studies • Identify the data used as inputs. Describe how the data were acquired, the process needed to enter the input data, the pre-processing applied, and how missing/low-quality data were handled • Describe the AI system outputs and how they were presented to the users Implementation: • Describe the settings in which the AI system was evaluated • Describe the clinical workflow/care pathway in which the AI system was evaluated, the timing of its use, and how the final supported decision was reached and by whom Safety and errors: • Provide a description of how significant errors/malfunctions were defined and identified • Describe how any risks to patient safety or instances of harm were identified, analyzed, and minimized Human factors: • Describe the human factors tools, methods or frameworks used, the use cases considered, and the users involved Ethics: • Describe whether specific methodologies were utilized to fulfil an ethics-related goal (such as algorithmic fairness) and their rationale

https://doi.org/10.3348/kjr.2022.0834

Korean Journal of Radiology
KJR

Table 1. AI-Specific Items in the Checklists of CONSORT-AI, SPIRIT-AI, and DECIDE-AI [3,4,25] **(Continued)**

	CONSORT-AI and SPIRIT-AI	DECIDE-AI
Results	Harms: • Describe results of any analysis of performance errors and how errors were identified, or if no such analysis was planned or done, justify why not*	Participants: • Describe the baseline characteristics of the patients included in the study, and report on input data missingness • Describe the baseline characteristics of the users included in the study Implementation: • Report on the user exposure to the AI system, on the number of instances the AI system was used, and on the users' adherence to the intended implementation • Report any significant changes to the clinical workflow or care pathway caused by the AI system Modifications: • Report any changes made to the AI system or its hardware platform during the study. Report the timing of these modifications, the rationale for each, and any changes in outcomes observed after each of them Human-computer agreement: • Report on the user agreement with the AI system. Describe any instances of and reasons for user variation from the AI system's recommendations and, if applicable, users changing their mind based on the AI system's recommendations Safety and errors: • List any significant errors/malfunctions related to: AI system recommendations, supporting software/hardware, or users. Include details of: (i) rate of occurrence, (ii) apparent causes, (iii) whether they could be corrected, and (iv) any significant potential impacts on patient care • Report on any risks to patient safety or observed instances of harm (including indirect harm) identified during the study Human factors: • Report on the usability evaluation, according to recognized standards or frameworks • Report on the user learning curves evaluation
Discussion		Support for intended use: • Discuss whether the results obtained support the intended use of the AI system in clinical settings Safety and errors: • Discuss what the results indicate about the safety profile of the AI system. Discuss any observed errors/malfunctions and instances of harm, their implications for patient care, and whether/how they can be mitigated
Data and code availability	• State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use	• Disclose if and how data and relevant code are available

*Included in CONSORT-AI alone, †Included in SPIRIT-AI alone. AI = artificial intelligence, ML = machine learning, ROI = region of interest

introduced, which help provide an overview of the necessary methodological elements.

Key words
Artificial intelligence; Machine learning; Deep learning; Clinical research; Clinical evaluation; Randomized clinical

 https://doi.org/10.3348/kjr.2022.0834

trial; Research method; Real-world evidence; Value-based healthcare

Availability of Data and Material

Data sharing does not apply to this article as no datasets were generated or analyzed during the current study.

Conflicts of Interest

Seong Ho Park is the Editor-in-Chief of the *Korean Journal of Radiology*.

Laure Fournier received speaker fees from Bayer, Novartis, Janssen, Sanofi, General Electric Healthcare, and Median technologies and had research collaboration with Philips, Evolucare, ArianaPharma, Siemens, General Electric Healthcare, and Dassault Systems regarding AI unrelated to the current article.

All other authors have no conflicts of interest to disclose.

Author Contributions

Conceptualization: all authors. Writing—original draft: Seong Ho Park. Writing—review & editing: Joon-Il Choi, Laure Fournier, Baptiste Vasey.

ORCID iDs

Seong Ho Park
https://orcid.org/0000-0002-1257-8315
Joon-Il Choi
https://orcid.org/0000-0003-0018-8712
Laure Fournier
https://orcid.org/0000-0002-1878-0290
Baptiste Vasey
https://orcid.org/0000-0002-0017-8891

Funding Statement

None

REFERENCES

1. Newman TB, Browner WS, Cummings SR, Hulley SB. *Designing studies of medical tests*. In: Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB, eds. *Designing clinical research*, 4th ed. Philadelphia, PA: Lippincott Williams & Wilkins, 2013:171-187
2. Rodger M, Ramsay T, Fergusson D. Diagnostic randomized controlled trials: the final frontier. *Trials* 2012;13:137
3. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ* 2020;370:m3164
4. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. *BMJ* 2020;370:m3210
5. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 2018;286:800-809
6. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med* 2022;28:31-38
7. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17:195
8. NEJM Group. What is value-based healthcare? NEJM Catalyst Web site. https://catalyst.nejm.org/doi/full/10.1056/CAT.17.0558. Published January 1, 2017. Accessed October 30, 2022
9. Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJY, Kann BH. Randomized clinical trials of machine learning interventions in health care: a systematic review. *JAMA Netw Open* 2022;5:e2233946
10. Zhou Q, Chen ZH, Cao YH, Peng S. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. *NPJ Digit Med* 2021;4:154
11. Spadaccini M, Iannone A, Maselli R, Badalamenti M, Desai M, Chandrasekar VT, et al. Computer-aided detection versus advanced imaging for detection of colorectal neoplasia: a systematic review and network meta-analysis. *Lancet Gastroenterol Hepatol* 2021;6:793-802
12. Zhang Y, Zhang X, Wu Q, Gu C, Wang Z. Artificial intelligence-aided colonoscopy for polyp detection: a systematic review and meta-analysis of randomized clinical trials. *J Laparoendosc Adv Surg Tech A* 2021;31:1143-1149
13. Deliwala SS, Hamid K, Barbarawi M, Lakshman H, Zayed Y, Kandel P, et al. Artificial intelligence (AI) real-time detection vs. routine colonoscopy for colorectal neoplasia: a meta-analysis and trial sequential analysis. *Int J Colorectal Dis* 2021;36:2291-2303
14. Hassan C, Spadaccini M, Iannone A, Maselli R, Jovani M, Chandrasekar VT, et al. Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis. *Gastrointest Endosc* 2021;93:77-85.e6
15. Park SH, Choi J, Byeon JS. Key principles of clinical validation, device approval, and insurance coverage decisions of artificial intelligence. *Korean J Radiol* 2021;22:442-453
16. Páez A, Rovers M, Hutchison K, Rogers W, Vasey B, McCulloch P; IDEAL Collaboration. Beyond the RCT: when are randomized trials unnecessary for new therapeutic devices, and what should we do instead? *Ann Surg* 2022;275:324-331
17. Frieden TR. Evidence for health decision making - beyond randomized, controlled trials. *N Engl J Med* 2017;377:465-475

18. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet* 2005;365:82-93
19. Park SH, Han K, Jang HY, Park JE, Lee JG, Kim DW, et al. Methods for clinical evaluation of artificial intelligence algorithms for medical diagnosis. *Radiology* 2022 Nov 8. [Epub]. https://doi.org/10.1148/radiol.220182
20. Vagal A, Saba L. Artificial intelligence in "code stroke"-a paradigm shift: do radiologists need to change their practice? *Radiol Artif Intell* 2022;4:e210204
21. Morey JR, Zhang X, Yaeger KA, Fiano E, Marayati NF, Kellner CP, et al. Real-world experience with artificial intelligence-based triage in transferred large vessel occlusion stroke patients. *Cerebrovasc Dis* 2021;50:450-455
22. Wijnberge M, Geerts BF, Hol L, Lemmers N, Mulder MP, Berge P, 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. *JAMA* 2020;323:1052-1060
23. Park SY, Park JE, Kim H, Park SH. Review of statistical methods for evaluating the performance of survival or other time-to-event prediction models (from conventional to deep learning approaches). *Korean J Radiol* 2021;22:1697-1707
24. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020;368:m689
25. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* 2022;377:e070904
26. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. *BMJ* 2021;374:n2061
27. Hirst A, Philippou Y, Blazeby J, Campbell B, Campbell M, Feinberg J, et al. No surgical innovation without evaluation: evolution and further development of the IDEAL framework and recommendations. *Ann Surg* 2019;269:211-220