science editing
pISSN 2288-8063
eISSN 2288-7474

Sci Ed 2019;6(2):91-98
https://doi.org/10.6087/kcse.164

Review

Ethical challenges regarding artificial intelligence in medicine from the perspective of scientific editing and peer review

Seong Ho Park[1], Young-Hak Kim[2], Jun Young Lee[3], Soyoung Yoo[4], Chong Jai Kim[5]

[1]Department of Radiology and Research Institute of Radiology, [2]Cardiology Division, Asan Medical Center, University of Ulsan College of Medicine, Seoul; [3]National IT Industry Promotion Agency, Jincheon; [4]Health Innovation Big Data Center, Asan Medical Center, Seoul; [5]Department of Pathology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea

Abstract

This review article aims to highlight several areas in research studies on artificial intelligence (AI) in medicine that currently require additional transparency and explain why additional transparency is needed. Transparency regarding training data, test data and results, interpretation of study results, and the sharing of algorithms and data are major areas for guaranteeing ethical standards in AI research. For transparency in training data, clarifying the biases and errors in training data and the AI algorithms based on these training data prior to their implementation is critical. Furthermore, biases about institutions and socioeconomic groups should be considered. For transparency in test data and test results, authors should state if the test data were collected externally or internally and prospectively or retrospectively at first. It is necessary to distinguish whether datasets were convenience samples consisting of some positive and some negative cases or clinical cohorts. When datasets from multiple institutions were used, authors should report results from each individual institution. Full publication of the results of AI research is also important. For transparency in interpreting study results, authors should interpret the results explicitly and avoid over-interpretation. For transparency by sharing algorithms and data, sharing is required for replication and reproducibility of the research by other researchers. All of the above mentioned high standards regarding transparency of AI research in healthcare should be considered to facilitate the ethical conduct of AI research.

Keywords

Artificial intelligence; Ethics; Research; Publishing; Bias

Received: May 16, 2019
Accepted: May 28, 2019

Correspondence to Seong Ho Park
seongho@amc.seoul.kr

ORCID
Seong Ho Park
https://orcid.org/0000-0002-1257-8315
Young-Hak Kim
https://orcid.org/0000-0002-3610-486X
Jun Young Lee
https://orcid.org/0000-0003-0698-2036
Soyoung Yoo
https://orcid.org/0000-0002-2953-508X
Chong Jai Kim
https://orcid.org/0000-0002-2844-9446

Introduction

Artificial intelligence (AI), which makes use of big data based on advanced machine learning techniques involving multiple layers of artificial neural networks (i.e., deep learning), has the potential to substantially improve many aspects of healthcare [1]. With new technological developments, new ethical issues are also introduced. Many international authorities, including some in the medical field, are attempting to establish ethical guidelines regarding the use of AI [2-5]. Healthcare is a field in which the implementation of AI involves multiple ethical challenges. Notable ethical issues related to AI in healthcare are listed in Table 1 (not exhaustive) [1,2,6-22]. AI is unlikely to earn trust from patients and healthcare professionals without addressing these ethical issues adequately.

Some of these ethical issues are relevant to the scientific editing and peer review processes of academic journals. Transparency is one of the key ethical challenges surrounding AI in healthcare [23]. The scientific editing and peer review processes of medical journals are well positioned to ensure that studies on AI in healthcare are held to a high standard of transparency, thereby facilitating the ethical conduct of research studies and the ethical spread of knowledge. The role of peer-reviewed medical journals in this field is particularly important because many research studies on AI in healthcare are published without peer review through preprint servers, such as arXiv.org, most of which are not accepted by the medical field [1,24]. This article highlights several specific areas in research studies on AI in healthcare that currently require additional transparency, explains why additional transparency is needed, and discusses how to achieve it from the perspective of scientific editing and peer review. This article can serve as a guide for authors and reviewers to ensure that research reports on AI in healthcare are held to a high standard of transparency. However, it is not intended to serve as an all-inclusive guide for writing and reviewing research articles on AI in healthcare, nor is it intended to provide general ethical guidelines for AI in healthcare. More general guides can be found elsewhere [2-5,25].

Transparency in Training Data

Reports of research studies on AI in healthcare should explain the details of how authors collected, processed, and organized the data used in studies thoroughly (with specific mention of dates and medical institutions), in addition to describing the baseline demographic characteristics, clinical characteristics (such as the distribution of severity of the target condition, distribution of alternative diagnoses, and comorbidities), and technical characteristics (such as techniques for image acquisition) of the collected data thoroughly to help readers understand biases and errors in the data [13,23,25,26]. In both academia and industry, researchers are praised for training increasingly sophisticated algorithms. However, relatively little attention is paid to how data are collected, processed, and organized [13]. Therefore, improvements in this area are needed. Several related guidelines are available to assist in transparent reporting [25,27-29].

Modern AI algorithms built using big data and multiple layers of artificial neural networks have achieved superior accuracy compared to past algorithms. However, current AI algorithms are strongly dependent on their training data. The accuracy of these algorithms cannot go beyond the information inherent to the datasets on which they are trained, meaning they cannot avoid the biases and errors in the training data. Because the datasets used to train AI algorithms for medical diagnosis/prediction are prone to selection biases and may not adequately represent a target population in real-world scenarios for various reasons (explained below), this strong dependency on training data is particularly concerning. Clarifying the biases and errors in training data and AI algorithms based on these training data prior to their implementation is critical, especially given the black box nature of AI and the fact that cryptic biases and errors can harm nu-

Table 1. Notable ethical issues related to AI in healthcare (not exhaustive)

Ethical issue
Privacy and data protection, consent for data use, and data ownership.
Fairness and bias in data and AI algorithms. If data underrepresent any particular groups of patients (e.g., ethnicity, gender, and economic status), then the resulting AI algorithms will have biases against these groups.
Evidence to ensure the greatest benefit to patients while avoiding any harm (i.e., rigorous clinical validation of AI).
Equitable access (e.g., if resource-poor hospitals and patients have limited access to AI, disparities in healthcare may be exacerbated).
Conflicts of interest (e.g., if healthcare professionals involved in patient care hold positions in AI startups or other commercial entities, it may increase the risk that professional judgment or actions regarding a primary interest will be unduly influenced by a secondary interest).
Accountability (i.e., who should be liable for adverse events related to the use of AI?).
The exploitation of AI for unethical purposes (e.g., manipulating AI outputs with malicious intent by covertly modifying data to perform an adversarial attack [20]).

AI, artificial intelligence.

merous patients simultaneously and negatively affect health disparities at a large scale [10].

Complex mathematical AI models for medical diagnosis/prediction require a large quantity of data for training. Producing and annotating this magnitude of medical data is resource intensive and difficult [13,22,30,31]. Additionally, the medical data accumulated in clinical practice are generally heterogeneous across institutions and practice settings based on variations in patient composition, physician preference, equipment and facilities, and health policies. Many data are also unstructured and unstandardized in terms of both their final form and process of acquisition. Missing data are also relatively common. As a result, most clinical data, whether from electronic health records or medical billing claims, are poorly defined and largely insufficient for effective exploitation by AI techniques [16,32]. In other words, they are "not AI ready" [16,22,31,32], which makes the data collection and curation process even more difficult. Therefore, researchers who collect big medical data to develop AI algorithms might rely on whatever data are available, even if these data are prone to various selection biases [13,30,33]. Existing large public medical datasets are also used for developing AI. However, few such databases are currently available, and most are small and lack real-world variation [2,25,32]. Additionally, any assumptions or hidden biases within such data may not be explicitly known [2,25,32].

Dataset shifting in medicine poses another challenge. In disciplines where medical equipment for generating data evolves rapidly (such as various radiologic scanners), dataset shifting occurs relatively frequently [2,9]. For example, if an AI algorithm is trained only on images from a 1.5-Tesla magnetic resonance imaging scanner, it may or may not output the same results for examinations performed using a 3-Tesla magnetic resonance imaging scanner.

Biases in medical data are sometimes macroscopic [2,10-12,16]. For example, electronic health records and insurance claim datasets are records of patient's clinical courses, but they also serve as a tool for healthcare providers to justify specific levels of reimbursement. Consequently, data may reflect reimbursement strategies and payment mechanisms more than providing an objective clinical assessment. As another example, health record data may contain biases for or against a particular race, gender, or socioeconomic group. However, in many cases, the biases are complex and difficult to anticipate [2,12]. Such biases may manifest as inadvertent discrimination against under-represented subsets of a population, limited interoperability (algorithms trained on patients from a single institution may not be generalizable across different institutions and populations), and the frame problem [2,10,17]. The frame problem is exemplified by a recent accident caused by an experimental autonomous driving car from Tesla that crashed into the trailer of a truck turning left, killing the driver, because it failed to recognize the white side of the trailer as a hazard [10]. Simply put, AI cannot classify what it is not trained on. This raises significant concerns in medicine because unexpected situations can occur in real-world clinical practice at any time, and such situations are not infrequent. Any anticipated biases in data, as well as any unintended consequences and pitfalls that can occur based on these biases, should be transparently disclosed in research reports.

Transparency in Test Data and Test Results

In addition to the points raised in the previous section regarding transparency of training data, there are several other points worth noting regarding the transparency of datasets for testing the performance of AI algorithms. Firstly, for the same reasons mentioned above, external validation (i.e., assessing the performance of an AI algorithm using datasets collected independently from the training dataset) is essential when testing the performance of an AI algorithm for medical diagnosis/prediction [10,13,17,18,25,26,33,34]. Computer scientists evaluate algorithms on test datasets, but these are typically subsamples (such as random-split samples) of the original dataset from which the training data were also drawn, meaning they are likely to contain the same biases [13,35]. Research reports should clearly distinguish preliminary performance evaluations using split subsamples from genuine external validations. The lack of adequate external validation for AI algorithms designed for medical diagnosis/prediction is a pressing concern [35]. According to a recent systematic review of the research studies published between January 1, 2018, and August 17, 2018, that investigated the performance of AI algorithms for analyzing medical images to provide diagnostic decisions, only 6% performed some type of external validation [35]. A clear editorial guide regarding external validation will promote adequate external validation.

Secondly, when describing the process of collecting test datasets, it is necessary to distinguish whether datasets were convenience samples consisting of some positive and some negative cases or clinical cohorts that adequately reflect the epidemiological characteristics and disease manifestation spectrum of clinically-defined target patients in real-world practice [36]. The former is referred to as diagnostic case-control design, while the latter is referred to as diagnostic cohort design [36-38]. For example, when testing an AI algorithm that detects lung cancer on chest radiographs, testing its performance on a dataset consisting of some cases with lung cancer and some cases without lung cancer is a diagnostic case-control design [36]. By contrast, a diagnostic cohort de-

sign defines the clinical setting and patients first by establishing eligibility criteria. For example, a study might consider asymptomatic adults aged X–Y years with Z-packs-per-year smoking history. Then, all (or a random selection) of those who fulfilled the criteria within a certain period are recruited and examined by the AI algorithm. It is recommended to perform a diagnostic cohort study in a prospective manner.

A diagnostic cohort is a better representation of real-world practice than a convenience case-control sample because it has a more natural prevalence of disease, more natural demographic characteristics, and a more natural disease manifestation spectrum including patients with disease-simulating conditions, comorbidities that may pose diagnostic difficulty, and findings for which the concrete distinction of disease versus non-disease is inappropriate [36]. Case-control design is prone to spectrum bias, which can potentially lead to an inflated estimation of diagnostic performance [33,39]. A diagnostic cohort design not only results in a less biased estimation of the clinical performance of an AI algorithm, but it also allows for the assessment of higher-level endpoints that are more clinically relevant, such as positive predictive value (or post-test probability), diagnostic yield, and the rate of false referrals [17,35,38].

A diagnostic cohort study using AI should describe patient eligibility criteria explicitly; it should also clarify the reasons and subject numbers for any incidents of individuals who were eligible but unenrolled, or those who were enrolled but were not included in the analysis of study outcomes [27,29]. Typical reasons for such incidents include technical failure, drop-out/follow-up loss, and missing reference standard information.

Finally, for the same reasons mentioned above, the performance of an AI algorithm may vary across different institutions [40-43]. Therefore, it is essential to use test datasets from multiple institutions and report all individual institutional results to assess the interoperability of an AI algorithm and generalizability of study results accurately. Underreporting of negative or unfavorable study results is a well-known pitfall in medical research in general; similarly, some researchers or sponsors of AI research studies may be inclined to report favorable results selectively. Underreporting of negative or unfavorable study results was a significant reason why the policy of prospectively registering clinical trials was first introduced in 2005 by the International Committee of Medical Journal Editors. Currently, numerous medical journals consider reports of clinical trials for publication only if they have been registered *a priori* in publicly accessible trial registries (e.g., clinicaltrials.gov) with key study plans.

Transparency through the full publication of the results of AI research is equally important [15,25,26,44]. A similar requirement for the prospective registration of studies for clinical validation of AI algorithms will help increase confidence in study results among patients and healthcare professionals, as well as in the process of regulatory approval. In fact, the requirement for prospective registration of diagnostic test accuracy studies has already been proposed by some medical journals [45]. Studies to validate the clinical performance of AI algorithms belong to the broader category of diagnostic test accuracy studies. Therefore, the adoption of this policy would have an instant effect.

Transparency in Interpreting Study Results

The interpretation of results in research reports on AI in healthcare should be explicit and avoid over-interpretation (also referred to as "spin") [46]. Because AI in healthcare is a topic in which not only related professionals but also the public have considerable interest, the reporting of research studies should consider laypeople as potential readers. An explicit interpretation of study results without spin is critical to prevent misinforming the public or lay media. Spinning study results may make a study "look better." However, excessive hype [1] that is generated inadvertently or exacerbated through misinformation will ultimately erode faith in AI for both the public and healthcare professionals. The scientific editing and peer review processes play an important role in building trust in AI by publishing clearer, more accurate information.

Typical examples of spinning study results include describing the results from split samples as external validation, claiming proof of clinical validity or utility for a limited external validation using diagnostic case-control design, and claiming evidence regarding the impact on healthcare outcomes based on accuracy results alone. Accuracy results obtained from test datasets split from an original dataset do not represent external validation and may only show technical feasibility at best [13,25,26,34,35]. High-accuracy results from external datasets in a case-control design may further support technical/analytical validity, but they are still not sufficient to prove clinical validity [25,36]. High-accuracy results from external datasets collected in a diagnostic cohort design without strong selection biases can support clinical validity more strongly [25,26,33]. However, such high-accuracy results cannot directly determine the impact of AI on healthcare and clinical utility [1,33,47]. One would need clinical trials focusing on health outcomes or observational research studies with appropriate analytical methods to account for confounders, preferably in the form of prospective design, to address the impact of AI on healthcare and clinical utility [1,33,48-53].

Additional Transparency by Sharing Algorithms and Data

Addressing the issues mentioned above would help to enhance the transparency of research studies on AI. However, the effects would be indirect. By contrast, sharing AI algorithms and data from a research study with other researchers or practitioners so they can independently validate the algorithms and compare them to similar algorithms is a more direct means of ensuring the reproducibility and generalizability of AI algorithms for greater transparency. Lack of sharing appears to be an important reason why innovative medical software solutions with clinical potential in most software research, including AI research, have largely been discarded and failed in the transition from academic use cases to widely applicable clinical tools [24,54].

Based on this phenomenon, one prominent medical journal in the field of AI in medicine has recently adopted a policy to strongly encourage making the computer algorithms reported in the journal available to other researchers [24]. Additionally, a body of researchers has recently published the FAIR Guiding Principles for scientific data management and stewardship to provide guidelines to improve the findability, accessibility, interoperability, and reuse of digital assets [55]. Scientific editing and peer review can facilitate such movements by embracing them. However, the proprietary nature of AI algorithms, as well as data protection and ownership, are issues that must be resolved carefully.

Conclusion

Healthcare is a field in which the implementation of AI involves multiple ethical challenges. AI in healthcare is unlikely to earn trust from patients and healthcare professionals without addressing these ethical issues adequately. Transparency is one of the key ethical issues surrounding AI in healthcare. A list of specific questions to ask to make studies evaluating the performance of AI algorithms more ethically transparent is provided in Table 2. Note that Table 2 is not a comprehensive checklist for reporting research studies on AI in healthcare. Further information and relevant checklists can be found elsewhere [56] and should also be referred to appropriately. The scientific editing and peer review processes of medical journals are well positioned to ensure that studies of AI in healthcare are held to a high standard regarding transparency, thereby facilitating the ethical conduct of research studies and spread of knowledge.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44-56.

Table 2. Questions to ask to improve ethical transparency in studies evaluating the performance of artificial intelligence algorithms

Question to ask
Regarding training data
Do the authors thoroughly explain how they collected, processed, and organized the data?
Do the authors thoroughly describe the characteristics of the data/patients including demographic characteristics, clinical characteristics, and technical characteristics?
Do the authors explicitly disclose anticipated biases in the data as well as unintended consequences and pitfalls that could result from the biases?
Regarding test data and results
In addition to the above questions, the following questions should also be asked.
Do the authors clearly state if the test data were collected prospectively or retrospectively?
Do the authors clearly state whether the test data were a subsample of the initial dataset from which the training data were also drawn or independent external data?
For external data, do the authors clearly state whether the test data represent a convenience series or a clinical cohort?
For a clinical cohort, do the authors clearly explain patient eligibility criteria and which specific clinical settings they represent?
If test datasets from multiple institutions were used, do the authors report results from each individual institution?
Do the authors clarify (by providing the name of the registry and a study identifier) if they prospectively registered the study in a publicly accessible registry?
Regarding interpretation of study results
Do the authors interpret the results explicitly and avoid over-interpretation?
Regarding sharing of algorithms and data
Do the authors explain how to access their algorithms and data in the report (e.g., placing a link to a web page for download) if they are willing to share them?
For shared data, do the authors explain how they have ensured patient privacy and data protection?

https://doi.org/10.1038/s41591-018-0300-7
2. American College of Radiology; European Society of Radiology; Radiology Society of North America, et al. Ethics of AI in radiology: European and North American multisociety statement [Internet]. Reston, VA: American College of Radiology [cited 2019 May 9]. Available from: https://www.acrdsi.org/News-and-Events/Call-for-Comments
3. Chakchouk M. Information meeting on UNESCO's role in artificial intelligence [Internet]. Paris: UNESCO; 2018 [cited 2019 May 9]. Available from: https://en.unesco.org/sites/default/files/unesco_ai_info_meeting_ppt_220119_en.pdf
4. European Commission's High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI [Internet]. Brussels: European Commission; 2019 [cited 2019 May 9]. Available from: https://ec.europa.eu/futurium/en/ai-alliance-consultation
5. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically aligned design [Internet]. Piscataway, NJ: IEEE [cited 2019 May 9]. Available from: https://ethicsinaction.ieee.org
6. American Medical Association. Augmented intelligence in health care [Internet]. Chicago, IL: American Medical Association; 2018 [cited 2019 May 9]. Available from: https://www.ama-assn.org/system/files/2019-01/augmented-intelligence-policy-report.pdf
7. SFR-IA Group; CERF; French Radiology Community. Artificial intelligence and medical imaging 2018: French Radiology Community white paper. Diagn Interv Imaging 2018;99:727-42. https://doi.org/10.1016/j.diii.2018.10.003
8. Gostin LO, Halabi SF, Wilson K. Health data and privacy in the digital era. JAMA 2018;320:233-4. https://doi.org/10.1001/jama.2018.8374
9. Tang A, Tam R, Cadrin-Chenevert A, et al. Canadian Association of Radiologists white paper on artificial intelligence in radiology. Can Assoc Radiol J 2018;69:120-35. https://doi.org/10.1016/j.carj.2018.02.002
10. Yu KH, Kohane IS. Framing the challenges of artificial intelligence in medicine. BMJ Qual Saf 2019;28:238-41. https://doi.org/10.1136/bmjqs-2018-008551
11. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018;178:1544-7. https://doi.org/10.1001/jamainternmed.2018.3763
12. Char DS, Shah NH, Magnus D. Implementing machine learning in health care: addressing ethical challenges. N Engl J Med 2018;378:981-3. https://doi.org/10.1056/NEJMp1714229
13. Zou J, Schiebinger L. AI can be sexist and racist: it's time to make it fair. Nature 2018;559:324-6. https://doi.org/10.1038/d41586-018-05707-8
14. AI diagnostics need attention. Nature 2018;555:285. https://doi.org/10.1038/d41586-018-03067-x
15. Greaves F, Joshi I, Campbell M, Roberts S, Patel N, Powell J. What is an appropriate level of evidence for a digital health intervention? Lancet 2019;392:2665-7. https://doi.org/10.1016/S0140-6736(18)33129-5
16. Maddox TM, Rumsfeld JS, Payne PO. Questions for artificial intelligence in health care. JAMA 2019;321:31-2. https://doi.org/10.1001/jama.2018.18932
17. Parikh RB, Obermeyer Z, Navathe AS. Regulation of predictive analytics in medicine. Science 2019;363:810-2. https://doi.org/10.1126/science.aaw0029
18. Park SH, Do KH, Choi JI, et al. Principles for evaluating the clinical implementation of novel digital healthcare devices. J Korean Med Assoc 2018;61:765-75. https://doi.org/10.5124/jkma.2018.61.12.765
19. The Lancet. Is digital medicine different? Lancet 2018;392:95. https://doi.org/10.1016/S0140-6736(18)31562-9
20. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science 2019;363:1287-9. https://doi.org/10.1126/science.aaw4399
21. Prabhu SP. Ethical challenges of machine learning and deep learning algorithms. Lancet Oncol 2019;20:621-2. https://doi.org/10.1016/S1470-2045(19)30230-X
22. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol 2019;20:e262-73. https://doi.org/10.1016/s1470-2045(19)30149-4
23. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. PLoS Med 2018;15:e1002689. https://doi.org/10.1371/journal.pmed.1002689
24. Bluemke DA. Editor's note: publication of AI research in radiology. Radiology 2018;289:579-80. https://doi.org/10.1148/radiol.2018184021
25. England JR, Cheng PM. Artificial intelligence for medical image analysis: a guide for authors and reviewers. AJR Am J Roentgenol 2019;212:513-9. https://doi.org/10.2214/AJR.18.20490
26. Park SH, Kressel HY. Connecting technological innovation in artificial intelligence to real-world medical practice through rigorous clinical validation: what peer-reviewed medical journals could do. J Korean Med Sci 2018;33:e152. https://doi.org/10.3346/jkms.2018.33.e152
27. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015;350:g7594. https://doi.org/10.1136/bmj.g7594
28. Lambin P, Leijenaar RT, Deist TM, et al. Radiomics: the

bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017;14:749-62. https://doi.org/10.1038/nrclinonc.2017.141
29. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351:h5527. https://doi.org/10.1136/bmj.h5527
30. Nsoesie EO. Evaluating artificial intelligence applications in clinical settings. JAMA Netw Open 2018;1:e182658. https://doi.org/10.1001/jamanetworkopen.2018.2658
31. Vigilante K, Escaravage S, McConnell M. Big data and the intelligence community: lessons for health care. N Engl J Med 2019;380:1888-90. https://doi.org/10.1056/NEJMp1815418
32. Langlotz CP, Allen B, Erickson BJ, et al. A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 NIH/RSNA/ACR/The Academy Workshop. Radiology 2019;291:781-91. https://doi.org/10.1148/radiol.2019190613
33. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology 2018;286:800-9. https://doi.org/10.1148/radiol.2017171920
34. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016;18:e323. https://doi.org/10.2196/jmir.5870
35. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol 2019;20:405-10. https://doi.org/10.3348/kjr.2019.0025
36. Park SH. Diagnostic case-control versus diagnostic cohort studies for clinical validation of artificial intelligence algorithm performance. Radiology 2019;290:272-3. https://doi.org/10.1148/radiol.2018182294
37. Pepe MS. Study design and hypothesis testing. In: Pepe MS, editor. The statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press; 2003. p. 214-7.
38. Newman TB, Browner WS, Cummings SR, Hulley SB. Designing studies of medical tests. In: Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB, editors. Designing clinical research. Philadelphia, PA: Lippincott Williams & Wilkins; 2013. p. 171-87.
39. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem 2005;51:1335-41. https://doi.org/10.1373/clinchem.2005.048595
40. Ting DS, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 2017;318:2211-23. https://doi.org/10.1001/jama.2017.18152
41. Li X, Zhang S, Zhang Q, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. Lancet Oncol 2019;20:193-201. https://doi.org/10.1016/S1470-2045(18)30762-9
42. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med 2018;15:e1002683. https://doi.org/10.1371/journal.pmed.1002683
43. Hwang EJ, Park S, Jin KN, et al. Development and validation of a deep learning-based automated detection algorithm for major thoracic diseases on chest radiographs. JAMA Netw Open 2019;2:e191095. https://doi.org/10.1001/jamanetworkopen.2019.1095
44. Gill J, Prasad V. Improving observational studies in the era of big data. Lancet 2018;392:716-7. https://doi.org/10.1016/S0140-6736(18)31619-2
45. Korevaar DA, Hooft L, Askie LM, et al. Facilitating prospective registration of diagnostic accuracy studies: a STARD Initiative. Clin Chem 2017;63:1331-41. https://doi.org/10.1373/clinchem.2017.272765
46. Ochodo EA, de Haan MC, Reitsma JB, Hooft L, Bossuyt PM, Leeflang MM. Overinterpretation and misreporting of diagnostic accuracy studies: evidence of "spin". Radiology 2013;267:581-8. https://doi.org/10.1148/radiol.12120527
47. Burke W. Genetic tests: clinical validity and clinical utility. Curr Protoc Hum Genet 2014;81:9.15.11-8. https://doi.org/10.1002/0471142905.hg0915s81
48. Wang P, Berzin TM, Glissen Brown JR, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut 2019 Feb 29 [Epub]. https://doi.org/10.1136/gutjnl-2018-317500
49. INFANT Collaborative Group. Computerised interpretation of fetal heart rate during labour (INFANT): a randomised controlled trial. Lancet 2017;389:1719-29. https://doi.org/10.1016/S0140-6736(17)30568-8
50. Steinhubl SR, Waalen J, Edwards AM, et al. Effect of a home-based wearable continuous ECG monitoring patch on detection of undiagnosed atrial fibrillation: the mSToPS Randomized Clinical Trial. JAMA 2018;320:146-55. https://doi.org/10.1001/jama.2018.8102
51. Denis F, Basch E, Septans AL, et al. Two-year survival comparing web-based symptom monitoring vs routine surveillance following treatment for lung cancer. JAMA

2019;321:306-7. https://doi.org/10.1001/jama.2018.18085
52. Story A, Aldridge RW, Smith CM, et al. Smartphone-enabled video-observed versus directly observed treatment for tuberculosis: a multicentre, analyst-blinded, randomised, controlled superiority trial. Lancet 2019;393:1216-24. https://doi.org/10.1016/S0140-6736(18)32993-3
53. Park HJ, Jang JK, Park SH, et al. Restaging abdominopelvic computed tomography before surgery after preoperative chemoradiotherapy in patients with locally advanced rectal cancer. JAMA Oncol 2018;4:259-62. https://doi.org/10.1001/jamaoncol.2017.4596
54. Elhalawani H, Fuller CD, Thompson RF. The potential and pitfalls of crowdsourced algorithm development in radiation oncology. JAMA Oncol 2019 Apr 19 [Epub]. https://doi.org/10.1001/jamaoncol.2019.0157
55. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018. https://doi.org/10.1038/sdata.2016.18
56. EQUATOR Network. Reporting guidelines for main study types [Internet]. Oxford: EQUATOR Network [cited 2019 May 9]. Available from: http://www.equator-network.org