J Korean Med Sci. 2018 May 28;33(22):e152
https://doi.org/10.3346/jkms.2018.33.e152
eISSN 1598-6357·pISSN 1011-8934

JKMS

Special Article
Editing, Writing & Publishing

Check for updates

Connecting Technological Innovation in Artificial Intelligence to Real-world Medical Practice through Rigorous Clinical Validation: What Peer-reviewed Medical Journals Could Do

Seong Ho Park [1] **and Herbert Y. Kressel** [2]

[1]Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Korea
[2]Department of Radiology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA

OPEN ACCESS

Received: Feb 9, 2018
Accepted: Mar 29, 2018

Address for Correspondence:
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

ORCID iDs
Seong Ho Park
https://orcid.org/0000-0002-1257-8315
Herbert Y. Kressel
https://orcid.org/0000-0002-1435-6329

Funding
This work was supported by the Industrial Strategic technology development program (10072064) funded by the Ministry of Trade Industry and Energy (MOTIE, Korea).

Disclosure
The authors have no potential conflicts of interest to disclose.

ABSTRACT

Artificial intelligence (AI) is projected to substantially influence clinical practice in the foreseeable future. However, despite the excitement around the technologies, it is yet rare to see examples of robust clinical validation of the technologies and, as a result, very few are currently in clinical use. A thorough, systematic validation of AI technologies using adequately designed clinical research studies before their integration into clinical practice is critical to ensure patient benefit and safety while avoiding any inadvertent harms. We would like to suggest several specific points regarding the role that peer-reviewed medical journals can play, in terms of study design, registration, and reporting, to help achieve proper and meaningful clinical validation of AI technologies designed to make medical diagnosis and prediction, focusing on the evaluation of diagnostic accuracy efficacy. Peer-reviewed medical journals can encourage investigators who wish to validate the performance of AI systems for medical diagnosis and prediction to pay closer attention to the factors listed in this article by emphasizing their importance. Thereby, peer-reviewed medical journals can ultimately facilitate translating the technological innovations into real-world practice while securing patient safety and benefit.

Keywords: Artificial Intelligence; Machine Learning; Decision Support Techniques; Peer Review; Journalism, Medical; Validation Studies

INTRODUCTION

Artificial intelligence (AI) is projected to substantially influence clinical practice in the foreseeable future, especially in areas of diagnosis, risk assessment and prognostication through predictive algorithms. Notably, promising results have recently been reported regarding the application of convolutional neural networks,[1-3] a deep learning technology used for analyzing images. In medicine, convolutional neural networks have been utilized in the diagnostic analysis of a variety of medical images such as those of the retinal fundus,[4,5] histopathology,[6] endoscopy,[7] and the full range of radiologic[8-11] images. However, despite the excitement around the technologies, it is yet rare to see examples of robust clinical validation

Author Contributions

Conceptualization: Park SH, Kressel HY. Supervision: Kressel HY. Writing - original draft: Park SH. Writing - review & editing: Kressel HY.

of these clinical applications and, as a result, very few are currently in clinical use.[12-14] Despite the potential of AI technologies, it cannot be denied that the application of AI in health care is overhyped and is at risk of commercial exploitation to a certain extent.[13] The ultimate purpose of introducing AI into medicine is to achieve better, safer care for our patients. A thorough, systematic validation of AI technologies using adequately designed clinical research studies before their integration into clinical practice is critical to ensure patient benefit and safety while avoiding any inadvertent harms. The importance of proper clinical validation of AI technologies used for medicine has recently been underscored by multiple premier peer-reviewed medical journals[13-16] and a comprehensive methodologic guide for the clinical validation[17] has also recently been published. Peer-reviewed medical journals play a crucial role in the pathway towards the clinical validation of AI technologies used for medicine as the peer-reviewed medical journals employ a fundamental mechanism that vets the scientific and clinical value, validity, and integrity of research studies. The importance of peer-reviewed medical journals as more authoritative and reliable sources for updates regarding clinical validation of AI technologies is further highlighted these days since many related research studies are also published without peer review, in repositories such as https://arxiv.org, a repository of electronic preprints, which is moderated but not peer-reviewed, and does rely on peer to peer expertise in clinical evaluation of these technologies, thereby potentially adding to the existing hype around AI.[18]

With these issues in mind, we would like to suggest several specific points regarding the role that peer-reviewed medical journals can play to help achieve proper and meaningful clinical validation of AI technologies designed for use in medicine, especially diagnostic and predictive software tools developed with deep learning technology and high-dimensional data. AI can be applied to medicine in various ways. Of those, in this article, we will consider AI technologies designed to make medical diagnosis and prediction, i.e., classification tasks (for example, a distinction between cancer and benign disease or between good and poor responders to a therapy), built with "big" clinical datasets, as such technologies underpin the data-driven precision medicine in the AI era. Appraisal of AI technologies for medical diagnosis and prediction can be performed at different levels of efficacy[19] such as diagnostic accuracy efficacy (for example, a study by Ting et al.[5]), patient outcome efficacy (for example, a study by the INFANT Collaborative Group[20]), and societal efficacy that considers cost-benefit and cost-effectiveness. This article will focus on the evaluation of diagnostic accuracy efficacy. Although this article deals with some fundamental methodologic principles, the purpose of this article is not to provide comprehensive explanations on related methodology. Further methodologic details can be found in a recent methodologic guide.[17] Also, for an exemplary paper which successfully addressed the points to be discussed in this article, a study by Ting et al.[5] could be referred.

POINTS FOR PEER-REVIEWED MEDICAL JOURNALS TO EMPHASIZE TO HELP ACHIEVE PROPER CLINICAL VALIDATION OF AI

First, clarification of the meaning of the word *validation* as used in AI and machine learning (ML) articles would be helpful since, unlike the commonly accepted definition of the term *validation* in medicine/health literature,[21] this term is also used in AI/ML literature as technical jargon with a somewhat different meaning. According to the convention in the field of AI/ML,

https://doi.org/10.3346/jkms.2018.33.e152

validation as the technical jargon also refers to a particular step in the sequence of training, validation, and test steps for algorithm development (**Fig. 1**), where the validation step is to fine-tune the algorithm after training (see studies by Lakhani and Sundaram[8] and Larson et al.[9] for example).[17,22] Journals should try to avoid confusion with the use of the term, for example, by referring to the fine-tuning step and clinical validation as "internal validation" and "external validation," respectively, or explicitly naming them "fine-tuning step" and "clinical validation," respectively.

Second, the use of adequately sized datasets that are collected in newly recruited patients or at different sites than the dataset used for algorithm development and training which effectively represent the target patients undergoing a given diagnostic/predictive procedure in a "real-world" clinical practice setting are essential for "external" validation of the clinical performance of AI systems (for example, a study by Ting et al.[5]) to achieve an unbiased assessment (**Fig. 1**). The importance of using proper external datasets in validating the performance of AI systems built with deep learning cannot be overstated because mechanistic interrogations of

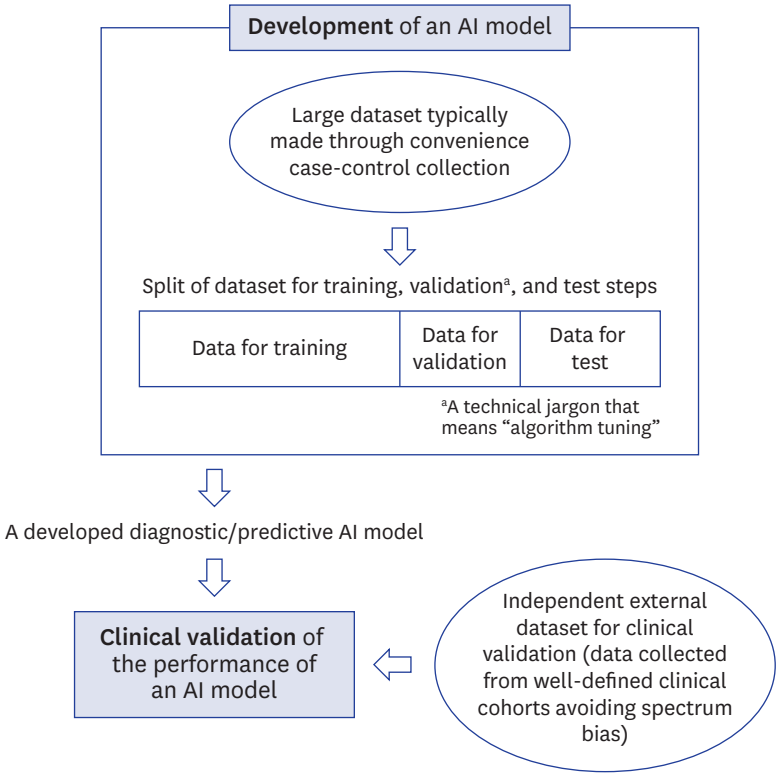

Fig. 1. Typical processes for development and clinical validation of an artificial intelligence model such as a deep learning algorithm for medical diagnosis and prediction.
The dataset used to develop a deep learning algorithm is typically convenience case-control data, which is prone to spectrum bias.[17] The algorithm development goes through training, validation, and test steps, for which the entire dataset is then split, for example, 50% for the training step and 25% each for the validation and test steps.[22] The term *validation* here is a technical jargon that means tuning of the algorithm under development, unlike the commonly accepted definition in medicine/health literature as in clinical validation. The test step, if performed using the typical split-sample "internal" validation method, should be distinguished from the true external validation as the former falls short of validating the clinical performance or generalizability of the developed algorithm. The use of a dataset that is collected in a manner that minimizes spectrum bias in newly recruited patients or at different sites than the dataset used for algorithm development, which effectively represents the target patients in a real-world clinical practice, is essential for external validation of the clinical performance of an AI algorithm.
AI = artificial intelligence.

https://doi.org/10.3346/jkms.2018.33.e152

the results created by a deep learning network are difficult due to the "black-box" nature of the technology, i.e., one cannot simply look inside a deep neural network to understand how it works to give a particular output due to the complex web of multiple interconnected layers and innumerable individual weights calculated with back-propagation for myriad artificial neuronal connections.[1,13,15,16] In addition, split-sample "internal" validation, i.e., validation of the performance using a fraction of data that is randomly split from the entire dataset and is kept unused for algorithm training (for example, studies by Ehteshami Bejnordi et al.[6], Lakhani and Sundaram[8], and Yasaka et al.[10]) should be distinguished from the true external validation mentioned above (**Fig. 1**). In contrast with the true external validation, split-sample validation, which has on occasion been termed as external validation in published papers, is statistically inadequate to account for overfitting and cannot generally avoid spectrum bias.[17] Therefore, although split-sample validation may demonstrate the internal technical validity of an AI algorithm, it falls short of validating its clinical performance or generalizability.[17] More in-depth explanations can be found elsewhere.[17] *Overfitting* and *spectrum bias* are significant pitfalls that can substantially exaggerate the performance of an AI system.[16,17,23] *Overfitting* refers to a situation in which a learning algorithm customizes itself too much to the training data, including idiosyncratic spurious statistical associations, to the extent that it negatively impacts the algorithm's ability to generalize to new data while exaggerating its performance on the training dataset.[1,17] It can be particularly problematic in overparameterized classification models built with high-dimensional data.[24,25] An *overparameterized* model is a mathematical model that contains too many "*x*" parameters (called *high-dimensional*) relative to the number of data to feed the model for training.[24,25] An example is an AI algorithm using convolutional neural network to analyze medical images as each pixel of an image is considered a separate *x* parameter in the mathematical model. *Spectrum bias* indicates a situation in which the spectrum of patient manifestations (e.g., severity, stage, or duration of the disease; presence and severity of comorbidities; demographic characteristics, etc.) in the data used for algorithm training does not adequately reflect the spectrum in those to whom the algorithm will be applied in clinical practice. This can be another source of data overfitting. Use of proper external datasets as explained earlier is crucial to avoid these pitfalls, for which prospectively collected data are better than those obtained in retrospective cohorts.

Third, use of large datasets obtained from multiple institutions for validation of the clinical performance of an AI system should be encouraged. Performance of an AI system may vary according to the selection of validation datasets due to differences in the degree of overfitting and patient manifestation spectrum between the datasets. Some types of data such as radiologic images may also be subject to additional sources of variability as different scanners/vendors and scan parameters/techniques may also influence the performance of AI systems.[16,26] Therefore, using datasets obtained from multiple institutions would be advantageous in achieving more robust validation of the performance. One good example is a study by Ting et al.[5] in which authors used ten multiethnic cohorts obtained from multiple institutions for external validation of the performance of their AI algorithm.

Fourth, prospective registration of studies to validate the performance of AI systems, like registration of clinical trials of interventions (for example, at clinicaltrials.gov), can be proposed to increase transparency in the validation. With varying performance results obtained with different datasets from multiple institutions as previously mentioned, some researchers or sponsors might be inclined to selectively report favorable results, which would create a problem of under-reporting unfavorable results. Such under-reporting was a significant reason why the policy of prospectively registering clinical trials was first

https://doi.org/10.3346/jkms.2018.33.e152

introduced in 2005 by the International Committee of Medical Journal Editors (ICMJE). In compliance with the ICMJE policy, numerous medical journals consider reports of trials for publication only if they had been registered in any of the publicly accessible trial registries accepted by the ICMJE before enrollment of the first study participant. Similar requirements have also been implemented by regulatory governmental organizations and funders. Likewise, prospective registration of diagnostic test accuracy studies, which include studies to validate the performance of AI systems, has already been proposed.[27] Adoption of this policy by medical journals as well as by governmental agencies and funders will enhance transparency in the validation of the performance of AI systems.

Fifth, in addition to the points mentioned above to improve the quality and transparency in validating the performance of AI systems, for AI systems developed with supervised learning (i.e., outcome statuses for an algorithm to predict are provided as labeled data for algorithm training) using data labeled by human interpreters, it would be important to advise the investigators to clarify the experience and training of the individuals doing the labeling and the variability between those doing the labeling. The ultimate performance of an AI system is profoundly influenced by the quality of the data used for training the system, and the quality of labeling is an important factor for evaluating the performance of an AI system.

Lastly, encouraging authors to refer to Standards for Reporting Diagnostic Accuracy (STARD)[28] and Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD),[21] the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) Network guidelines for reporting diagnostic accuracy studies and multivariable prediction model for individual prognosis or diagnosis, respectively, would be helpful for improving the completeness and consistency in reporting studies to validate the performance of AI systems although these guidelines are not customized for AI. Guidelines specific to reporting AI/ML predictive models[29] are also available albeit not as widely implemented as STARD or TRIPOD and would facilitate better reporting of the research results.

SUMMARY

Peer-reviewed medical journals can encourage investigators who wish to validate the performance of AI systems for medical diagnosis and prediction to pay closer attention to the factors listed in this article by emphasizing their importance. Thereby, peer-reviewed medical journals can promote execution and reporting of more robust clinical validation of AI systems and can ultimately facilitate translating the technological innovations into real-world practice while securing patient safety and benefit.

REFERENCES

1. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, et al. Deep learning: a primer for radiologists. *Radiographics* 2017;37(7):2113-31.
PUBMED | CROSSREF
2. Lee JG, Jun S, Cho YW, Lee H, Kim GB, Seo JB, et al. Deep learning in medical imaging: general overview. *Korean J Radiol* 2017;18(4):570-84.
PUBMED | CROSSREF
3. An intuitive explanation of convolutional neural networks. https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets. Updated August 11, 2016. Accessed March 22, 2018.

4. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316(22):2402-10.
PUBMED | CROSSREF
5. Ting DS, Cheung CY, Lim G, Tan GS, Quang ND, Gan A, 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(22):2211-23.
PUBMED | CROSSREF
6. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017;318(22):2199-210.
PUBMED | CROSSREF
7. Chen PJ, Lin MC, Lai MJ, Lin JC, Lu HH, Tseng VS. Accurate classification of diminutive colorectal polyps using computer-aided analysis. *Gastroenterology* 2018;154(3):568-75.
PUBMED | CROSSREF
8. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* 2017;284(2):574-82.
PUBMED | CROSSREF
9. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. *Radiology* 2018;287(1):313-22.
PUBMED | CROSSREF
10. Yasaka K, Akai H. Liver fibrosis: deep convolutional neural network for staging by using gadoxetic acid-enhanced hepatobiliary phase MR images. *Radiology* 2018;287(1):146-55.
PUBMED | CROSSREF
11. Yasaka K, Akai H. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. *Radiology* 2018;286(3):887-96.
PUBMED | CROSSREF
12. AI diagnostics need attention. *Nature* 2018;555(7696):285-6.
CROSSREF
13. The Lancet. Artificial intelligence in health care: within touching distance. *Lancet* 2018;390(10114):2739.
PUBMED | CROSSREF
14. Interview with Dr. Ziad Obermeyer on how collaboration between doctors and computers will help improve medical care--Supplement to the N Engl J Med 2017; 377:1209-11. http://www.nejm.org/action/showMediaPlayer?doi=10.1056%2FNEJMp1705348&aid=NEJMp1705348_attach_1&area=. Updated September 28, 2017. Accessed March 22, 2018.
15. Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. *JAMA* 2018;319(1):19-20.
PUBMED | CROSSREF
16. Kahn CE Jr. From images to actions: opportunities for artificial intelligence in radiology. *Radiology* 2017;285(3):719-20.
PUBMED | CROSSREF
17. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 2018;286(3):800-9.
PUBMED | CROSSREF
18. Video from RSNA 2017: how will AI change radiology? http://www.auntminnie.com/index.aspx?sec=sup&sub=aic&pag=dis&itemId=119197. Updated December 6, 2017. Accessed March 22, 2018.
19. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making* 1991;11(2):88-94.
PUBMED | CROSSREF
20. INFANT Collaborative Group. Computerised interpretation of fetal heart rate during labour (INFANT): a randomised controlled trial. *Lancet* 2017;389(10080):1719-29.
PUBMED | CROSSREF
21. 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.
PUBMED | CROSSREF
22. Hastie TJ, Tibshirani RJ, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York, NY: Springer; 2009.
23. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. *N Engl J Med* 2016;375(13):1216-9.
PUBMED | CROSSREF

24. Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat Rev Cancer* 2008;8(1):37-49.
PUBMED | CROSSREF

25. The curse of dimensionality in classification. http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification. Updated April 16, 2014. Accessed March 22, 2018.

26. Thrall JH, Li X, Li Q, Cruz C, Do S, Dreyer K, et al. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. *J Am Coll Radiol* 2018;15(3 Pt B):504-8.
PUBMED | CROSSREF

27. Korevaar DA, Hooft L, Askie LM, Barbour V, Faure H, Gatsonis CA, et al. Facilitating prospective registration of diagnostic accuracy studies: a STARD initiative. *Clin Chem* 2017;63(8):1331-41.
PUBMED | CROSSREF

28. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Radiology* 2015;277(3):826-32.
PUBMED | CROSSREF

29. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res* 2016;18(12):e323.
PUBMED | CROSSREF