AJNR

Critical Appraisal of Artificial Intelligence–Enabled Imaging Tools Using the Levels of Evidence System

N. Pham, V. Hill, A. Rauschecker, Y. Lui, S. Niogi, C.G. Fillipi, P. Chang, G. Zaharchuk and M. Wintermark

AJNR Am J Neuroradiol 2023, 44 (5) E21-E28
doi: https://doi.org/10.3174/ajnr.A7850
http://www.ajnr.org/content/44/5/E21

Critical Appraisal of Artificial Intelligence–Enabled Imaging Tools Using the Levels of Evidence System

N. Pham, V. Hill, A. Rauschecker, Y. Lui, S. Niogi, C.G. Fillipi, P. Chang, G. Zaharchuk, and M. Wintermark

ABSTRACT

SUMMARY: Clinical adoption of an artificial intelligence–enabled imaging tool requires critical appraisal of its life cycle from development to implementation by using a systematic, standardized, and objective approach that can verify both its technical and clinical efficacy. Toward this concerted effort, the ASFNR/ASNR Artificial Intelligence Workshop Technology Working Group is proposing a hierarchal evaluation system based on the quality, type, and amount of scientific evidence that the artificial intelligence–enabled tool can demonstrate for each component of its life cycle. The current proposal is modeled after the levels of evidence in medicine, with the uppermost level of the hierarchy showing the strongest evidence for potential impact on patient care and health care outcomes. The intended goal of establishing an evidence-based evaluation system is to encourage transparency, foster an understanding of the creation of artificial intelligence tools and the artificial intelligence decision-making process, and to report the relevant data on the efficacy of artificial intelligence tools that are developed. The proposed system is an essential step in working toward a more formalized, clinically validated, and regulated framework for the safe and effective deployment of artificial intelligence imaging applications that will be used in clinical practice.

ABBREVIATIONS: AI = artificial intelligence; HIPPA = Health Insurance Portability and Accountability Act

As artificial intelligence (AI) reimagines many facets of health care, radiology will be a leading force for developing and leveraging AI-based imaging technologies.[1-3] This past decade saw a dramatic rise in the number of commercially available AI products receiving US FDA approval for clinical use in imaging.[4] As of October 2022, there are 521 FDA-authorized AI-enabled medical devices, of which 75.2% are for radiology use.[5] Of these, neuroimaging applications comprise a large share, with estimates of up to 40% of products on the market.[6] With the increasing availability of AI software, a systematic method of integrating these tools into a clinically validated and regulated framework is necessary for the safe and effective deployment of medical imaging AI applications in routine clinical patient care. Unlike AI in other industries, such as entertainment and advertising, which can afford to be tolerant of errors, errors in medicine can be fatal.

Adoption of an AI-enabled tool requires critical appraisal of its life cycle from development to implementation, with careful consideration of the existing scientific evidence supporting its clinical utility. However, standardized objective metrics to quantitate AI quality and clinical utility are currently lacking, limiting the fair and accurate evaluation and comparison of different AI-enabled tools, especially when multiple products exist for the same clinical task.[7]

These are not new issues as they also affect other medical imaging software products, but the number and diversity of AI-enabled tools suddenly now hitting the market makes it a timely moment to consider practical and unbiased ways of assessing such tools. Thus, the ASFNR/ASNR has created an AI workshop technology working group with the goal of providing a practical approach for evaluating the potential effectiveness of AI technology in clinical practice.

Toward this goal, here we introduce an evaluation system using hierarchal levels of evidence that reflect the rigor of scientific data (Figure). Demonstration of clinical efficacy and value, at the pinnacle of this evaluation system, is the most important factor for clinical adoption.

Different points in the imaging workflow can be augmented by AI-enabled tools, with a range of clinical applications including

Received December 16, 2022; accepted after revision March 16, 2023.

From the Department of Radiology (N.P., G.Z.), Stanford School of Medicine, Palo Alto, California; Department of Radiology (V.H.), Northwestern University Feinberg School of Medicine, Chicago, Illinois; Department of Radiology (A.R.), University of California, San Francisco, San Francisco, California; Department of Radiology (Y.L.), NYU Grossman School of Medicine, New York, New York; Department of Radiology (S.N.), Weill Cornell Medicine, New York, New York; Department of Radiology (C.G.F.), Tufts University School of Medicine, Boston, Massachusetts; Department of Radiology (P.C.), University of California, Irvine, Irvine, California; and Department of Neuroradiology (M.W.), The University of Texas MD Anderson Cancer Center, Houston, Texas.

Please address correspondence to Nancy Pham, MD, Stanford School of Medicine, Department of Radiology, 453 Quarry Rd, 322.11, Palo Alto, CA 94304; e-mail: pham.nancy@gmail.com; @stanfordneuroil

http://dx.doi.org/10.3174/ajnr.A7850

but not limited to administrative, operational, patient, and image centered tasks.[8-10] For the purposes of this white paper, the hierarchal levels of evidence system is most useful for imaging and patient-related AI applications. However, the main principles can be generalized to other applications.

Finally, the radiologist continues to be an instrumental gatekeeper of patient care quality and safety, particularly now as we enter the era of AI. As clinical domain experts, radiologists provide important oversight on the effective use of AI software in the clinical setting.[11] To better position the radiologist in this role, this white paper presents a structured method of guidance on the critical appraisal of AI software using the levels of evidence system.

Levels of Evidence

To date, there are no agreed upon levels of evidence needed for the evaluation of AI-enabled tools; thus, the already established medicine model provides a practical starting point for the development of such a systematic process.[12] We propose a hierarchy of levels of evidence reflecting the critical elements of an AI product's life cycle from development to the clinical implementation phase (Figure).

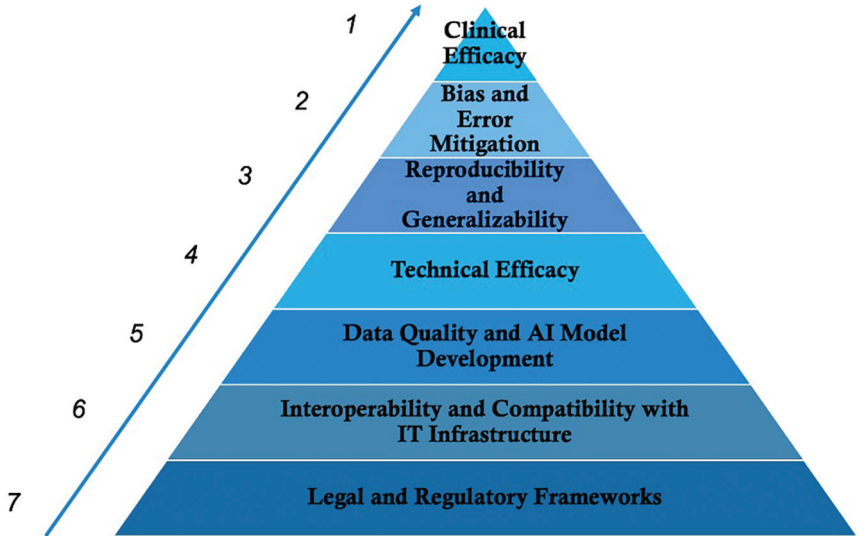

FIGURE. Levels of evidence. Proposed 7 levels of evidence for the systematic evaluation of an AI product's quality and effectiveness in the clinical setting.

The two levels at the base of the hierarchy, levels 6 and 7, are considered fundamental requirements that an AI product must meet before further consideration for implementation in the clinical workflow. For example, an AI product must comply with current legal and regulatory requirements (level 7) such as Health Insurance Portability and Accountability Act (HIPPA) and FDA clearance. Thereafter, it must be compatible with the information technology infrastructure (level 6) at the site where it will be deployed, before proceeding with other requirements listed in the hierarchy.

Further description of the levels of evidence from 1 to 7 is detailed below, with level 1 denoting the highest quality and strongest evidence for potential impact on patient care and health care outcomes. In addition, Table 1 provides an abbreviated summary, while Table 2 provides an expanded summary of each component of the evaluation system.

Data Quality and AI Model Development

AI models should be developed from data that are large, diverse, and reflective of the intended population. However, in practice, access to comprehensive and "big" data is challenging, and training is often performed on limited data.[13] This introduces bias that can affect reproducibility, generalizability, and performance outside the data range on which the model was trained. Thus, peer-reviewed publications including information on the source and characteristics of the data used to train, validate, and test the AI model can help end-users determine overall compatibility with the target patient population of interest.[14-16]

AI companies and developers do not typically publicly report detailed information on data used to develop or validate algorithms, despite having undergone the necessary FDA clearance process, limiting the ability of end-users to make informed decisions

Table 1: Summary of levels of evidence

Levels of Evidence	Element	Types of Evidence
Level 1	Clinical efficacy	One prospective or randomized clinical trial or meta-analysis
Level 2	Bias and error mitigation	At least 2 independent retrospective studies separate from original institution
Level 3	Reproducibility and generalizability	At least 2 retrospective studies with at least 1 from an institution independent of the original institution
Level 4	Technical efficacy	Two retrospective studies from the same institution
Level 5A	Data quality and AI model development with external testing	One retrospective study with internal and external data used for final performance reporting
Level 5B	Data quality and AI model development with internal testing	One retrospective study with only internal data used for final performance reporting
Level 6	Interoperability and integration into the IT infrastructure	AI company can provide a plan including interoperability standards for integration into the existing radiology and hospital digital information systems
Level 7	Legal and regulatory frameworks	AI-enabled tool is compliant with current patient data protection, security, privacy, HIPAA, and government regulations

Note:—IT indicates information technology.

Table 2: Detailed summary of levels of evidence[a]

Levels of Evidence	Element	Description	Types of Evidence	Significance
Level 1	Clinical efficacy	Clinical efficacy is the assessment of how the AI tool impacts patient care and health care outcomes	AI tool has been used in at least 1 prospective study, randomized clinical trial, or meta-analysis demonstrating potential for improved patient care and health care outcomes, including improved mortality, quality of life, morbidity, or reduced health care cost	Although measuring clinical efficacy and added value for an early AI technology is challenging, it remains the single most important feature for clinical success and adoption
Level 2	Bias and error mitigation	Biases of different types invariably exist in all data and can lead to AI modeling errors when applied to patients in different clinical settings	AI model can adapt to at least 2 separate institutions different from where it was initially developed (2 independent retrospective studies) AND Peer-reviewed results from those retrospective studies from above are available describing the various populations used to test the AI model, including age ranges, sex, and types of scanners AND	AI related errors in clinical practice may be harmful to patients; thus, the AI tool should be tested at multiple different sites, with differing patient demographics, disease prevalence, and imaging vendors to determine operational characteristics, generalizability, and potential pitfalls
			AI company has a process to continuously incorporate feedback and improve their model (postdeployment monitoring)	As clinical and patient care standards are constantly evolving, AI models may need routine surveillance and updates for performance and data drift
Level 3	Reproducibility and generalizability	AI-enabled tool can be applied to different clinical settings, while demonstrating consistent high-quality results	At least 2 retrospective studies showing that the AI tool has performance characteristics similar to or alternative methods in the literature AND At least 1 independent study different from where the AI tool was developed to demonstrate that the AI tool can adapt to at least 1 different institution	A multi-institution approach supports reproducibility and generalizability of the AI tool
Level 4	Technical efficacy	Technical efficacy is the assessment that the AI model correctly performs the task that it was trained to do	AI model performance has been shown in 2 retrospective studies to have potential clinical impact compared with similar or alternative methods in the literature AND These retrospective studies can be from the same institution	Recent investigations suggest that less than 40% of AI models have peer-reviewed evidence available on their efficacy[6]
Level 5A	Data quality and AI model development with external testing	AI models are prone to overfitting and an external test data set during development should be used to report final performance metrics	Level 5B evidence as described with an external test set used for performance validation	Inclusion of an external data set during the development phase supports the generalizability of the AI tool
Level 5B	Data quality and AI model development with internal testing	AI company should provide peer-reviewed information about the characteristics of the data used to develop the AI model AI company can explain how the AI model	One retrospective study showing the following: Peer-reviewed results detailing the inclusion/exclusion criteria, source, and type of data used to train, validate, and test the AI model AND	Data characteristics will influence the suitability and applicability of the AI model to the target patient population of interest

Continued on next page

Table 2: Continued

Levels of Evidence	Element	Description	Types of Evidence	Significance
		makes decisions that are relevant to patient care	Peer-reviewed results describing how the AI model was developed, including use of a standard of reference that is widely accepted for the intended clinical task AND No final external test set was used for final performance reporting	Selection of a high-quality standard of reference is important for accurately comparing the peak performance of the AI model to that of current clinical practice
Level 6	Interoperability and integration into the IT infrastructure	AI software should integrate seamlessly into the hospital information system, radiology information system, and PACS to be clinically useful	AI company can provide a plan including interoperability standards for integration into the existing radiology and hospital digital information systems	Successful clinical implementation of an AI tool requires close collaboration between the AI company and site experts, including radiologists, referring physicians, data scientists, and information technologists
			AI company can provide on-site demonstration of clinical integration and potential impact on workflow before full deployment	Real time demonstration is an important mechanism for identifying potential site-specific problems
Level 7	Legal and regulatory frameworks	Patient consent, privacy, and confidentiality laws will vary depending on state, local, and institutional regulations	AI-enabled tool is compliant with current patient data protection, security, privacy, HIPAA, and government regulations	AI companies, health care systems, and radiologists are key gatekeepers of patient autonomy, privacy, and safety

[a] Appropriate reporting of AI model performance will depend on the task; however, examples of relevant statistical measures include ROC, sensitivity, specificity, and positive and negative predictive values, among others.

about these products. Thus, the emphasis on more than 1 peer-reviewed publication in this white paper encourages some level of independent, critical, and structured analysis to provide scientific evidence for verifying the intended use and clinical impact of the AI product.

At the very least, even if a product does not meet this level of evidence expectation, it is most responsible for a company to provide information about their patient population, including demographic characteristics, model development and validation methods, and indicators of statistical efficacy. Purchasers and end-users should expect and require statistical evidence and, preferably, consider these levels of evidence as indicators for the strength of a tool's methodological quality of design, validity, and applicability to patient care.

Barriers to improving AI transparency include competing financial incentives among developers, data privacy and sharing restrictions, and some degree of acceptance of the "black box" nature of AI-based solutions. To overcome these limitations, initiatives have been proposed to establish minimum data reporting standards for AI in health care including but not limited to MINIMAR (MINimum Information for Medical AI Reporting), CONSORT-AI (Consolidated Standard of Reporting Trials-Artificial Intelligence), and CLAIM (Checklist for Artificial Intelligence in Medical Imaging).[17-19] Others have also introduced checklists, recommendations, and guidelines toward assessing the suitability of AI-based tools in the health care environment.[11,20-22] Our proposal utilizing the levels of evidence builds on these ongoing initiatives, with a greater focus on the availability of peer-reviewed evidence and publications, to improve confidence and trust for all stakeholders using AI-based tools.

Selection of a quality standard of reference during the development phase is critical for an accurate and fair comparison of the AI model's performance against the current standard of practice.[23,24] After all, the adoption of any clinical tool relies on scientific evidence that it imparts some advantage over an already existing approach to the problem. Using subpar proxies for the intended clinical task may overestimate the actual performance of the AI model in the clinical setting. For example, assessment of an AI-enabled tool for the detection of intracranial hemorrhage might utilize turnaround time in outpatients with unexpected bleeds as a metric rather than reporting the overall accuracy of the tool.[15,25]

To evaluate potential real-world clinical efficacy and generalizability, it is important to gauge an AI tool's performance on an external data set. Selection bias and reliance on retrospective data can lead to an AI model that too closely aligns with the original data and lacks the ability to generalize to new and unseen data. A recent study of deep learning algorithms for image-based radiologic diagnosis suggests that most will demonstrate diminished algorithm performance on the external data set, with some reporting a substantial performance decrease.[25]

External validation is increasingly recognized as a critical step for evaluating model performance but has been employed in relatively few published studies.[26] The latter may be attributed to the challenges of obtaining an appropriate external data set. However, nonetheless, it remains important to use an external testing data set, separate from the original data used to develop the model, to calculate final performance metrics.[15,25] This criterion is used to differentiate level 5A and level 5B. Potential sources of external

data includes information from a different institution or public data bases. Further rigorous external verification of performance, generalization, and reproducibility can be tested through a multi-institution approach.

To provide appropriate oversight on how AI decisions will impact patients, radiologists must encourage AI vendors to explain steps in the AI product's life cycle, in a manner that would allow for greater understandability and interpretability of its results. Of particular interest are details of the steps taken to reduce bias and ensure quality during the development process.[27] Detecting and mitigating bias in a machine learning model can be one of the most effort-intensive steps in the development process, as bias may be introduced at any point in the product's life cycle. Various approaches to reducing bias include emphasis on data transparency, mathematical approaches to de-biasing, interpretability/explainability of the decision-making process, and post-deployment surveillance strategies.[28]

Technical Efficacy versus Clinical Efficacy

There is a need to verify both the technical and clinical efficacy of any AI-enabled tool before clinical implementation.[29,30] Interestingly, a study in 2020 found that fewer than 40% of commercially available AI products had published, peer-reviewed evidence available demonstrating their efficacy.[4] Receiving FDA clearance for clinical use in radiology in no way guarantees clinical utility or clinical efficacy of the product.

Technical efficacy is defined by the ability of the AI model to correctly perform the task for which it was trained (level 4).[31] Scientific evidence that supports technical efficacy is often in the form of retrospective studies and includes peer-reviewed information about the AI model's data quality, development, and performance metrics, benchmarked against similar or alternatively accepted methods in the literature. For example, an automated brain tumor segmentation task may require initial published results on the Dice coefficient or Jaccard index score to demonstrate technical efficacy. Subsequently, it would be important to provide scientific evidence that performance is reproducible and generalizable across different clinical institutions, patient populations, MR imaging field strengths, and imaging vendors.[25]

Clinical efficacy is defined by the ability of the AI model to change patient care and health care outcomes (level 1). Therefore, this requires a higher level of evidence, often in the form of prospective and randomized clinical trials to prove that the AI-enabled tool can lead to results that are better than standard level of care. It is important to note that technical efficacy does not equate to clinical efficacy.[29-32] For example, performance metrics such as reproducibility, sensitivity, specificity, positive and negative predictive values, and area under the curve are able to summarize AI model performance well but provide little information on how it could change patient outcome. Thus, despite impressive and exciting AI research, we continue to see relatively slow adoption of this technology to the health care setting. This is partly attributed to the paucity of scientific evidence supporting clinical efficacy.[33]

Bias and Error Mitigation

AI clinical errors often reflect the interplay of different types of biases introduced by the imperfect process of collecting, training, and applying data (level 2).[16,34,35] Additionally AI-enabled tools can project societal and historical biases that may further exacerbate existing inequities related to sex, age, and socioeconomic differences, among others. Thus, it is important to have a systematic approach for monitoring performance variances in different patient populations.[36,37] Other mechanisms that can be used to mitigate errors include ensuring data quality, as described above; verifying generalizability and reproducibility across different clinical sites (level 3); and careful consideration of epidemiological and statistical factors, such as disease prevalence, that can impact AI performance on a specific population.[25,31] A major goal of this white paper is to emphasize the importance of peer-reviewed publications, including robust internal and external validation during model development and subsequent validation at other sites. Differing feature distribution among clinical sites and patient populations such as sex, ethnicity, age, socioeconomic condition, geographic distribution, disease risk factors, imaging equipment, and image quality can lead to unexpected model performance errors.

Health care is a fluid and dynamic landscape, with new and evolving clinical practice standards that will require routine re-evaluation of the performance of the AI-enabled tool. This is further compounded by the yet to be defined process of how AI models continuously learn and evolve over time with new data. Thus, defining a practical mechanism for postdeployment monitoring including incorporating an iterative feedback loop between the radiologist, AI-enabled tool, and AI company during the implementation phase will be critical for adapting to these changes and achieving long-term consistent effectiveness.[11,29,30,32]

Legal and Regulatory Frameworks

Policies pertaining to patient consent, data collection, and data usage will vary on a state, local, and institutional level. However, AI companies and health care systems should have standard operating procedures to maintain HIPPA compliance, patient data safety, confidentiality, and privacy (level 7).[36,38,39]

AI-enabled tools can be subjected to different regulatory requirements, depending on the proposed clinical setting and intended use. For example, for medically oriented AI-based tools, the FDA has 3 levels of clearance: the 510(k), premarket approval, and de novo pathways, each with its own specific criteria, which have been thoroughly explained elsewhere.[40]

Additionally, many other innovative and experimental AI research tools are being developed in-house under institutional internal review board approval outside the purview of government oversight.

Of the AI-enabled tools that have gone through FDA review, most have received FDA 510(k) clearance, which does not require safety or effectiveness data from clinical trials. Instead, the manufacturer can demonstrate that it is substantively equivalent to a predicate (another FDA-cleared or approved product). Thus, the emphasis on AI-enabled tools having more than 1 peer-reviewed publication is necessary in this white paper to encourage an independent, critical, and structured analysis of the AI-product. In contrast, substantially fewer products have gone through the FDA's more rigorous premarket approval or, alternatively, the de novo pathway, which is designed for AI-enabled medical devices that are not deemed high risk but do not have a predicate.

Currently, any major changes to an AI-enabled tool will require resubmission for FDA approval; thus, most AI algorithms may remain "static" or "locked" after they are introduced into the market. However, periodic surveillance and refinement of AI algorithms may be needed to adapt to the evolving health care environment,[41] without going through the full FDA review process again. This has prompted the FDA to consider more efficient and streamlined regulatory pathways to evaluate continuously learning AI through proposals such as the digital health precertification program and predetermined change control plan, which are currently under discussions. Unfortunately, as of now, no official process exists for major amendments to an existing AI algorithm.

The proposed hierarchy levels of evidence can be used to support an AI product's life cycle in both the static and continuously learning environment. For continuous learning AI, there is mobility between the levels of the hierarchy. As an example, once an AI-enabled tool has established its baseline technical and clinical efficacy, modifications to the AI algorithm requiring FDA approval may allow it to move between level 7 and any other upper levels by providing additional scientific data, since the other levels have been supported by scientific evidence during its development phase.

Interoperability and Integration into the IT Infrastructure

AI software should integrate seamlessly into the hospital information system, radiology information system, and PACS to be clinically and functionally useful.[30,32] A recent white paper on AI interoperability in imaging has explored the problems and challenges that must be addressed to achieve an ecosystem of interoperable AI products.[42] Until such harmonized standards are adopted, AI companies will need to provide a clear plan with defined interoperability standards for integration into the existing digital infrastructure (level 6).[43] The AI vendor should also be able to provide an on-site demonstration of the clinical tool in action in real time before full deployment. This will be an important opportunity to observe the AI model's performance on the target population, impact on workflow, and potential errors in clinical practice.

Added Clinical Value

It can take decades for health care innovations to become fully implemented into clinical practice.[44] Thus, the full clinical impact of AI on the health care system is likely to still mature and may not be completely apparent at the present time. Although challenging, defining and measuring the added value of an early technology remains the single most important factor for achieving clinical success and adoption.[2] No current consensus exists on how to measure the added value of an AI-enabled tool in clinical practice. However, one approach is to consider the tool's potential to improve patient outcomes compared with the cost of achieving that improvement in a value-based health care system:[45-47] Value = Patient Outcome/Cost. As emphasized previously, AI performance accuracy alone does not necessarily lead to improved patient outcomes; future prospective investigations, clinical trials, or meta-analyses (level 1 evidence) are needed to establish such a link. Similarly, AI-enabled tools may reduce cost to the patient and health care system by guiding clinical decision-making through a much more evidenced-based approach (ie, early detection of cerebral ischemia); however, more long-term investigations are still needed to understand the cost-benefit ratio. Randomized clinical trials are considered the gold standard for determining an intervention's impact on clinical care. Several recent failures to implement AI-based tools in the clinical setting have suggested their relevance for selecting AI products with meaningful clinical benefit, especially given some inherent opacity and incomplete understanding of the mechanistic basis for how AI models actually make predictions.[48,49] Toward establishing scientific evidence for clinical efficacy, several AI-enabled tools have successfully demonstrated a positive impact on patient-centered related outcomes in clinical trials (level 1 evidence).[50] The proposed hierarchy levels of evidence can be used to support an AI product's potential effectiveness and added value in the context of its available scientific data.

User Cases

To understand how the levels of evidence can be utilized, the following user cases derive from selected real-world applications of AI-enabled tools in the literature. Employing the levels of evidence can facilitate communication and understanding among stakeholders regarding the strength of peer-reviewed evidence available to support that tool's reported goal and potential clinical impact.

Level 1 Evidence. Strong scientific evidence exists for the positive clinical impact of AI-based tools used to guide clinical decision-making in stroke care.[51] Specifically, AI-based ischemic stroke triage and management have been shown to decrease patient morbidity and mortality while improving patient functionality through multiple practice-defining clinical trials.[52,53] There is also emerging evidence that these tools have the potential to reduce overall health care costs.[54]

Level 3 Evidence. AI-based tools can be used to augment aneurysm detection and analysis. In this example, the AI-based tool has at least 2 retrospective peer-reviewed publications inclusive of 2 or more different institutions.[55,56] However, there are currently no prospective data to assess the clinical impact of such a tool.

Level 5B Evidence. An AI-based tool designed to segment brain tumors with 1 retrospective study describing model development and performance without use of an external data set.

In summary, the levels of evidence are an important component of evidence-based medicine, and the adoption of such a classification system can help end-users prioritize information on the quality of AI products. Most importantly, AI-enabled tools exist on a spectrum with regard to their scientific rigor, with some products lacking peer-reviewed publications altogether to those that have been well-validated through multiple randomized clinical trials. The level of evidence that an AI-enabled tool will need, of course, will depend on its intended task, as illustrated above. As with all classification systems, level 1 evidence does not necessarily mean that these data should be accepted as fact while level 5B data should be disregarded. Our goal is to introduce a method

of scientific scrutiny to address the disconnect between expectations and reality.

CONCLUSIONS

Barriers to the clinical implementation of AI-enabled tools include factors related to the lack of understandability of the AI development and decision-making process, standardized criteria for comparing product quality and effectiveness, and rigorous scientific evidence supporting meaningful impact on patient care and health care outcomes. To overcome some of these challenges, the ASFNR/ASNR AI Workshop Technology Working Group has proposed hierarchal levels of evidence to objectively evaluate the scientific merit and potential effectiveness of AI technologies in clinical practice.

Disclosure forms provided by the authors are available with the full text and PDF of this article at www.ajnr.org.

REFERENCES

1. Zaharchuk G, Gong E, Wintermark M, et al. **Deep learning in neuroradiology.** *AJNR Am J Neuroradiol* 2018;39:1776–84 CrossRef Medline
2. Lui YW, Chang PD, Zaharchuk G, et al. **Artificial intelligence in neuroradiology: current status and future directions.** *AJNR Am J Neuroradiol* 2020;41:E52–59 CrossRef Medline
3. Bohr A, Memarzadeh K. **The rise of artificial intelligence in healthcare applications**. *Artificial Intelligence in Healthcare*. Cambridge, Massachusetts; Academic Press; 2020:25–60
4. van Leeuwen KG, de Rooij M, Schalekamp S, et al. **How does artificial intelligence in radiology improve efficiency and health outcomes?** *Pediatr Radiol* 2022;52:2087–93 CrossRef Medline
5. **Artificial intelligence and machine learning (AI/ML)-enabled medical devices.** October 5, 2022. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices. Accessed March 5, 2023
6. van Leeuwen KG, Schalekamp S, Rutten MJ, et al. **Artificial intelligence in radiology: 100 commercially available products and their scientific evidence.** *Eur Radiol* 2021;31:3 797–804 CrossRef Medline
7. Goergen SK, Frazer HM, Reddy S. **Quality use of artificial intelligence in medical imaging: what do radiologists need to know?** *J Med Imaging Radiat Oncol* 2022;66:225–32 CrossRef Medline
8. Letourneau-Guillon L, Camirand D, Guilbert F, et al. **Artificial intelligence applications for workflow, process optimization and predictive analytics.** *Neuroimaging Clin N Am* 2020;30:e1–15 CrossRef Medline
9. Kitamura FC, Pan I, Ferraciolli SF, et al. **Clinical artificial intelligence applications in radiology: Neuro.** *Radiol Clin North Am* 2021;59:1003–12 CrossRef Medline
10. Kaka H, Zhang E, Khan N. **Artificial intelligence and deep learning in neuroradiology: exploring the new frontier.** *Can Assoc Radiol J* 2021;72:35–44 CrossRef Medline
11. Scott I, Carter S, Coiera E. **Clinician checklist for assessing suitability of machine learning applications in healthcare.** *BMJ Health Care Inform* 2021;28:e100251 CrossRef Medline
12. Burns PB, Rohrich RJ, Chung KC. **The levels of evidence and their role in evidence-based medicine.** *Plast Reconstr Surg* 2011;128:305–10 CrossRef Medline
13. Willemink MJ, Koszek WA, Hardell C, et al. **Preparing medical imaging data for machine learning.** *Radiology* 2020;295:4–15 CrossRef Medline
14. Mongan J, Moy L, Kahn CE. **Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers.** *Radiol Artif Intell* 2020;2:e200029 CrossRef Medline
15. Bluemke DA, Moy L, Bredella MA, et al. **Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers-from the radiology editorial board.** *Radiology* 2020;294:487–89 CrossRef Medline
16. Yu AC, Eng J. **One algorithm may not fit all: how selection bias affects machine learning performance.** *Radiographics* 2020;40:1932–39 CrossRef Medline
17. Hernandez-Boussard T, Bozkurt S, Ioannidis JP, et al. **MINIMAR (MINimum information for medical AI reporting): developing reporting standards for artificial intelligence in health care.** *J Am Med Inform Assoc* 2020;27:2011–15 CrossRef Medline
18. Radiology: Artificial intelligence. **Checklist for Artificial Intelligence in Medical Imaging (CLAIM).** https://pubs.rsna.org/page/ai/claim?doi=10.1148%2Fryai&publicationCode=ai. Accessed March 2, 2023
19. Liu X, Rivera SC, Moher D, et al. **Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension.** *BMJ (Online)* 2020;370:m3164 CrossRef
20. Park Y, Jackson GP, Foreman MA, et al. **Evaluating artificial intelligence in medicine: phases of clinical research.** *JAMIA Open* 2020;3:326–31 CrossRef Medline
21. Jha A, Bradshaw T, Buvat I, et al. **Best practices for evaluation of artificial intelligence-based algorithms for nuclear medicine: the RELIANCE guidelines.** *J Nucl Med* 2022;63(Suppl 2):1725
22. Filice RW, Mongan J, Kohli MD. **Evaluating artificial intelligence systems to guide purchasing decisions.** *J Am Coll Radiol* 2020;17:1405–09 CrossRef Medline
23. Krause J, Gulshan V, Rahimy E, et al. **Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy.** *Ophthalmology* 2018;125:1264–72 CrossRef Medline
24. Duggan GE, Reicher JJ, Liu Y, et al. **Improving reference standards for validation of AI-based radiography.** *Br J Radiol* 2021;94:20210435 CrossRef Medline
25. Yu AC, Mohajer B, Eng J. **External validation of deep learning algorithms for radiologic diagnosis: a systematic review.** *Radiol Artif Intell* 2022;4:e210064 CrossRef Medline
26. Kim DW, Jang HY, Kim KW, et al. **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 CrossRef Medline
27. Dunnmon J. **Separating hope from hype: artificial intelligence pitfalls and challenges in radiology.** *Radiol Clin North Am* 2021;59:1063–74 CrossRef Medline
28. Vokinger KN, Feuerriegel S, Kesselheim AS. **Mitigating bias in machine learning for medicine.** *Commun Med (Lond)* 2021;1:25 CrossRef Medline
29. Kelly CJ, Karthikesalingam A, Suleyman M, et al. **Key challenges for delivering clinical impact with artificial intelligence.** *BMC Med* 2019;17:195 CrossRef Medline
30. He J, Baxter SL, Xu J, et al. **The practical implementation of artificial intelligence technologies in medicine.** *Nat Med* 2019;25:30–36 CrossRef Medline
31. Park SH, Han K. **Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction.** *Radiology* 2018;286:800–09 CrossRef Medline
32. Wolff J, Pauling J, Keck A, et al. **Success factors of artificial intelligence implementation in healthcare.** *Front Digit Health* 2021;3:594971 CrossRef Medline
33. Omoumi P, Ducarouge A, Tournier A, et al. **To buy or not to buy: evaluating commercial AI solutions in radiology (the ECLAIR guidelines).** *Eur Radiol* 2021;31:3786–96 CrossRef Medline
34. DeCamp M, Lindvall C. **Latent bias and the implementation of artificial intelligence in medicine.** *J Am Med Inform Asso.* 2020;27:2020–23 CrossRef Medline
35. Finlayson SG, Subbaswamy A, Singh K, et al. **The clinician and dataset shift in artificial intelligence.** *N Engl J Med* 2021;385:283–86 CrossRef Medline

36. Geis JR, Brady AP, Wu CC, et al. **Ethics of artificial intelligence in radiology: summary of the Joint European and North American Multisociety Statement.** *Radiology* 2019;293:436–40 CrossRef
37. Liu X, Glocker B, McCradden MM, et al. **The medical algorithmic audit.** *Lancet Digit Health* .2022;4:e384–97 CrossRef Medline
38. Vollmer S, Mateen BA, Bohner G, et al. **Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness.** *BMJ* 2020;368:l6927 CrossRef Medline
39. Spilseth B, McKnight CD, Li MD, et al. **AUR-RRA review: logistics of academic-industry partnerships in artificial intelligence.** *Acad Radiol* 2022;29:119–28 CrossRef Medline
40. **FDA-regulated AI algorithms: trends, strengths, and gaps of validation studies.** *Acad Radiol* 2022;29:559–66 CrossRef Medline
41. Pianykh OS, Langs G, Dewey M, et al. **Continuous learning AI in radiology: implementation principles and early applications.** *Radiology* 2020;297:6–14 CrossRef Medline
42. Genereaux B, O'Donnell K, Bialecki B, et al. **IHE radiology white paper: AI interoperability in imaging**. *Integrating the Healthcare Enterprise* 2021;1:????? https://www.ihe.net/uploadedFiles/Documents/Radiology/IHE_RAD_White_Paper_AI_Interoperability_in_Imaging.pdf. Accessed March 3, 2023
43. Wiggins WF, Magudia K, Schmidt TMS, et al. **Imaging AI in practice: a demonstration of future workflow using integration standards.** *Radiol Artif Intell* 2021;3:e210152 CrossRef Medline
44. Kirchner JE, Smith JL, Powell BJ, et al. **Getting a clinical innovation into practice: an introduction to implementation strategies.** *Psychiatry Res* 2020;283:112467 CrossRef Medline
45. Porter ME. **What is value in health care?** *N Engl J Med* 2010;363:2477–81 CrossRef
46. Brady AP, Visser J, Frija G, et al. **Value-based radiology: what is the ESR doing, and what should we do in the future?** *Insights Imaging* 2021;12:108 CrossRef
47. Teisberg E, Wallace S, O'Hara S. **Defining and implementing value-based health care: a strategic framework**. *Acad Med* 2020;95:682–85 CrossRef Medline
48. Plana D, Shung DL, Grimshaw AA, et al. **Randomized clinical trials of machine learning interventions in health care: a systematic review.** *JAMA Netw Open* 2022;5:e2233946 CrossRef Medline
49. Wilkinson J, Arnold KF, Murray EJ, et al. **Time to reality check the promises of machine learning-powered precision medicine.** *Lancet Digit Health* 2020;2:e677–80 CrossRef Medline
50. Campbell BC, Mitchell PJ, Kleinig TJ, et al; EXTEND-IA Investigators. **Endovascular therapy for ischemic stroke with perfusion-imaging selection.** *N Engl J Med* 2015;372:1009–18 CrossRef Medline
51. Soun JE, Chow DS, Nagamine M, et al. **Artificial intelligence and acute stroke imaging.** *AJNR Am J Neuroradiol* 2021;42:2–11 CrossRef Medline
52. Albers GW, Marks MP, Kemp S, et al; DEFUSE 3 Investigators. **Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging.** *N Engl J Med* 2018;378:708–18 CrossRef Medline
53. Ma H, Campbell BC, Parsons MW, et al; EXTEND Investigators. **Thrombolysis guided by perfusion imaging up to 9 hours after onset of stroke.** *N Engl J Med* 2019;380:1795–803 CrossRef Medline
54. van Leeuwen KG, Meijer FJ, Schalekamp S, et al. **Cost-effectiveness of artificial intelligence aided vessel occlusion detection in acute stroke: an early health technology assessment.** *Insights Imaging* 2021;12:133 CrossRef Medline
55. Heit JJ, Honce JM, Yedavalli VS, et al. **RAPID aneurysm: artificial intelligence for unruptured cerebral aneurysm detection on CT angiography.** *J Stroke Cerebrovasc Dis* 2022;31:106690 CrossRef Medline
56. Sahlein DH, Gibson D, Scott JA, et al. **Artificial intelligence aneurysm measurement tool finds growth in all aneurysms that ruptured during conservative management.** *J Neurointerv Surg* 2022 Sep 30. [Epub ahead of print] CrossRef Medline