Exploration of Medicine

Open Exploration

Open Access Perspective

Bridging the validation gap in artificial intelligence in radiology

Antonio Navarro-Ballester*

Department of Radiology, Hospital General Universitario de Castellón, 12004 Castellón de la Plana, Spain

***Correspondence:** Antonio Navarro-Ballester, Department of Radiology, Hospital General Universitario de Castellón, 12004 Castellón de la Plana, Spain. antonio.navarroball@gmail.com

Academic Editor: Michaela Cellina, ASST Fatebenefratelli Sacco, Italy

Received: March 28, 2026 **Accepted:** April 26, 2026 **Published:** June 11, 2026

Cite this article: Navarro-Ballester A. Bridging the validation gap in artificial intelligence in radiology. Explor Med. 2026;7:1001412. https://doi.org/10.37349/emed.2026.1001412

Abstract

Artificial intelligence (AI) has rapidly advanced in radiology, demonstrating high performance across a wide range of diagnostic tasks. However, clinical adoption remains slower and more uneven than anticipated. This discrepancy reflects a fundamental gap between algorithm validation and clinical implementation. Current validation strategies primarily rely on controlled datasets and performance metrics such as accuracy and area under the curve, which often fail to capture the complexity of clinical environments. This article examines the nature of this "validation gap" and argues that it reflects a broader structural mismatch between how AI systems are evaluated and how clinical care operates. We propose a conceptual framework comprising three levels of validation: technical validity, workflow validity, and clinical validity. While most studies focus on technical performance, limited attention is given to integration into clinical workflows and impact on patient outcomes. Key factors contributing to this gap include limited generalizability across diverse populations and imaging protocols, poor alignment with clinical workflows, and the underrepresentation of uncertainty in model outputs. These limitations hinder effective implementation and may reduce trust in AI systems. Bridging this gap requires a shift toward more comprehensive validation strategies, including multicenter and prospective studies, improved workflow integration, and explicit incorporation of uncertainty and human–AI interaction. Ultimately, the clinical value of AI in radiology should be assessed not only by its performance in controlled settings but also by its ability to support decision-making and improve patient outcomes in real-world practice.

Keywords

artificial intelligence, radiology, validation, generalizability, clinical implementation, uncertainty

Artificial intelligence (AI) has rapidly expanded within radiology, with a growing number of publications, commercially available tools, and regulatory approvals [1–6]. Despite this momentum, clinical adoption in routine practice remains slower and more uneven than anticipated. This discrepancy points to a fundamental issue: the gap between algorithm validation and clinical implementation. While many AI models demonstrate high performance under controlled conditions, their impact in routine practice is far less consistent [7]. This raises a critical question: why does strong algorithmic performance so often fail to translate into meaningful clinical benefit?

Even when evaluated beyond retrospective datasets, many AI systems fail to demonstrate meaningful clinical benefit, suggesting that current validation paradigms capture only a limited dimension of clinical effectiveness in practice. This limitation does not merely reflect model performance, but the inadequacy of prevailing evaluation frameworks, which remain largely centered on technical metrics rather than clinical implementation and impact [8, 9].

Such evaluations are typically conducted under controlled conditions, often with selected populations and standardized imaging protocols [10, 11]. However, these settings rarely capture the variability and uncertainty inherent to clinical environments. Differences in scanner types, acquisition parameters, disease prevalence, and patient characteristics are frequently underrepresented. More importantly, they are usually disconnected from clinical context, focusing on isolated detection tasks rather than integrated decision-making. Validation, in this sense, reflects performance under idealized conditions rather than clinical complexity.

A related and often overlooked dimension of this gap is the role of uncertainty. Clinical decision-making is inherently shaped by multiple layers of uncertainty, including variability in patient data, incomplete information, and ambiguity in diagnosis and treatment pathways [12, 13]. In contrast, most AI models are designed to produce deterministic outputs or point predictions, often without adequately quantifying or communicating their uncertainty.

This mismatch has important implications. Conventional performance metrics do not capture the confidence or reliability of individual predictions, and models may exhibit overconfidence even when incorrect, particularly in unfamiliar or out-of-distribution scenarios [14, 15]. Moreover, different sources of uncertainty, such as data-related variability and limitations in model knowledge, are rarely distinguished in clinical applications. As a result, AI systems may appear more certain than the clinical reality they are intended to support.

Incorporating uncertainty into AI predictions has the potential to improve decision-making, guide clinician attention, and mitigate automation bias. However, current approaches remain limited, both in their technical implementation and in how uncertainty is communicated to users. In practice, these approaches are not yet widely implemented in clinically deployed or regulatory-approved AI systems, where uncertainty is often communicated only through simplified outputs such as confidence scores. This further underscores that validation frameworks focused solely on accuracy fail to capture a critical component of clinical reasoning.

From a clinical perspective, uncertainty outputs should be presented in a way that is directly interpretable within the clinical workflow. For example, probability calibration can be used to indicate the confidence of a prediction, allowing radiologists to distinguish between high- and low-certainty outputs. In addition, systems may implement predefined uncertainty thresholds to flag or defer cases with low confidence, prompting closer human review. Out-of-distribution detection mechanisms can also be incorporated to identify cases that differ significantly from the training data, signaling that model predictions may be less reliable. Such approaches can help align algorithmic outputs with the inherent uncertainty of clinical decision-making.

Taken together, these observations suggest that the so-called “validation gap” reflects a broader structural mismatch between how AI systems are evaluated and how clinical care actually operates. We propose a structured conceptual framework comprising three levels of validation: technical validity, workflow validity, and clinical validity. Unlike prior frameworks that primarily address technical evaluation or implementation in isolation, this approach integrates technical performance, workflow integration, and clinical impact into a single translational continuum, explicitly focused on the gap between algorithm validation and routine clinical application. Technical validity refers to performance on predefined datasets using standard metrics. Workflow validity reflects how seamlessly a tool integrates into clinical processes and whether it enhances efficiency or usability. Clinical validity, ultimately, concerns whether the tool improves patient outcomes or influences management decisions (Figure 1). Most current AI studies remain confined to the first level, with limited evidence addressing the latter two. The conceptual framework is illustrated in Figure 1, and its practical implications are summarized in Table 1.

Table 1. Operational framework linking validation levels with study design, metrics, and clinical application.

Validation level	Typical study design	Key metrics	Example in radiology workflows
Technical validity	Retrospective dataset evaluation	Area under the curve (AUC), sensitivity, specificity	Detection of intracranial hemorrhage on CT
Workflow validity	Prospective or observational implementation studies	Reporting time, user interaction, and adoption rate	AI-based triage integrated into picture archiving and communication system (PACS), prioritizing urgent cases
Clinical validity	Prospective clinical studies or randomized trials	Change in management, diagnostic accuracy in practice, and patient outcomes	AI-assisted selection of patients for stroke thrombectomy

Several factors contribute to this gap. Generalizability remains a central challenge [16, 17]. Models trained in specific institutional settings may perform inconsistently when applied to different populations or imaging protocols. For example, AI systems for intracranial hemorrhage detection or pulmonary embolism triage have demonstrated high performance in controlled datasets, yet their impact may vary when deployed across institutions with different patient populations and imaging workflows. Even subtle variations in acquisition can alter performance. Evidence from recent systematic evaluations shows that models with strong internal performance frequently exhibit measurable declines when tested on external datasets, with specificity often being particularly affected [16, 17]. These performance drops are largely driven by domain shifts, including differences in patient demographics, disease prevalence, scanner hardware, and acquisition protocols.

This highlights that generalizability is not merely a technical limitation but a reflection of the gap between controlled validation environments and the heterogeneity of clinical data in practice. These domain shifts also contribute to increased uncertainty, which is rarely captured by conventional performance metrics.

In parallel, workflow misalignment limits usability [18]. Many AI tools function as external applications rather than integrated components of radiology workstations, creating friction in already time-constrained environments [19].

From an implementation perspective, minimizing workflow disruption is essential for effective adoption. AI tools should be seamlessly integrated into existing picture archiving and communication systems (PACS) and reporting systems, avoiding the need for external platforms or additional logins. Outputs should be presented within the standard reading interface, using intuitive visualizations and structured reporting elements that do not increase cognitive load. In addition, prioritization and triage functionalities should be aligned with existing clinical pathways, ensuring that AI-generated alerts are clinically meaningful and do not contribute to alert fatigue. These considerations highlight that successful integration depends not only on algorithm performance, but also on usability and workflow compatibility. In addition, the integration of AI systems into clinical workflows should be understood as a continuous and iterative process rather than a one-time deployment. Model performance must be regularly reassessed as new data become available, often requiring retraining and revalidation to maintain reliability. This cyclical process adds further complexity to implementation, highlighting the need for sustained technical and organizational support. This also implies the need for dedicated multidisciplinary teams to support ongoing model development, monitoring, and maintenance within clinical environments.

Cognitive and human factors further complicate adoption. Radiologists must calibrate trust in AI outputs, balancing the risks of overreliance and underutilization. Automation bias, alert fatigue, and time pressure all influence how these tools are used in practice. In addition, most AI tools are developed as static models, yet clinical environments are dynamic. Disease patterns evolve, imaging technologies change, and local practices adapt over time. Without longitudinal validation, model performance may degrade in ways that are not immediately apparent. An algorithm that performs well in isolation may fail when embedded in clinical reality.

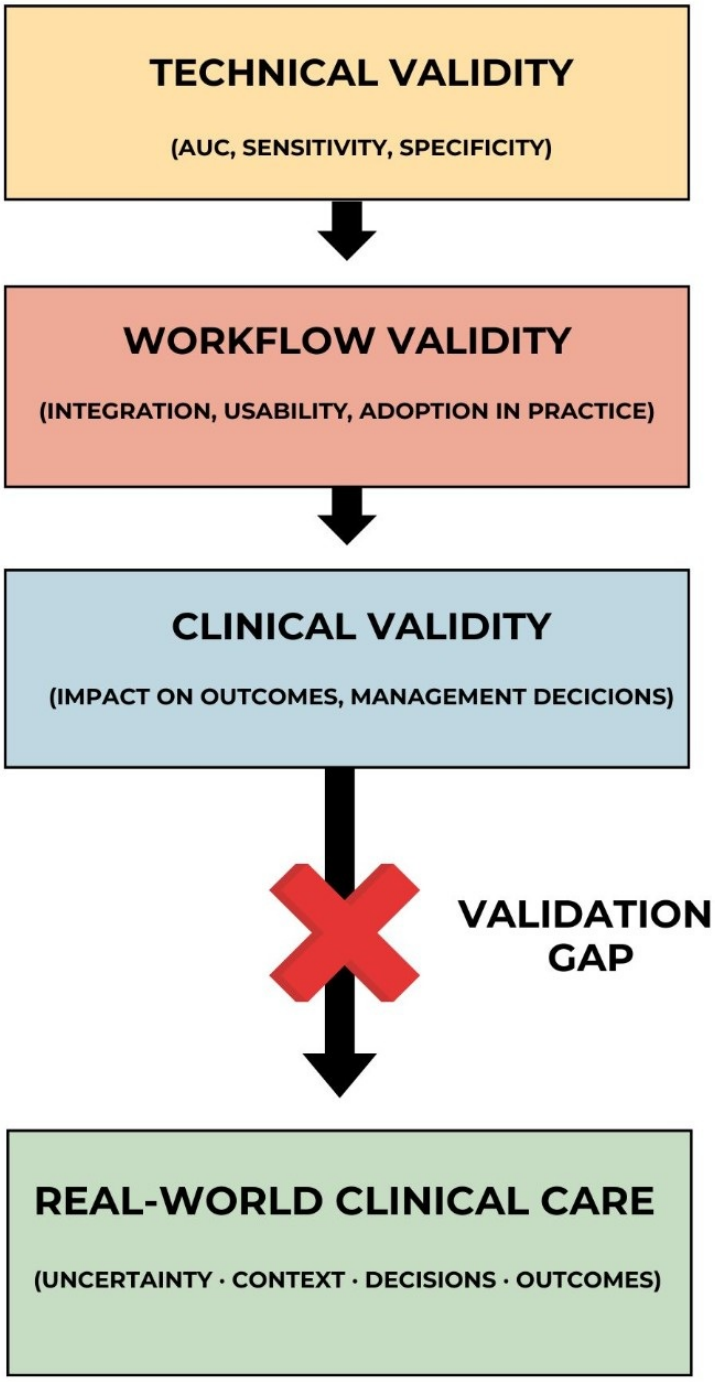

Figure 1. Conceptual framework of the validation gap in artificial intelligence for radiology. The diagram illustrates three hierarchical levels of validation: technical validity, workflow validity, and clinical validity. Current evaluation strategies are predominantly focused on technical performance, while workflow integration and clinical impact remain underexplored. The "validation gap," represented as a disruption between validated performance and routine clinical care, reflects the mismatch between algorithm-centered evaluation and the complexity of clinical practice, including uncertainty, context, and decision-making. AUC: area under the curve.

A critical limitation of the current literature is the relative scarcity of evidence demonstrating true clinical impact. While many studies report improved detection or classification accuracy, far fewer assess whether these improvements lead to better patient outcomes, reduced diagnostic errors in practice, or meaningful changes in therapeutic decisions. This distinction between detection and decision-making is essential. Enhanced sensitivity does not inherently translate into better care if it does not alter management. Emerging evidence suggests that even highly accurate models may have limited downstream impact when introduced into complex clinical pathways [20]. The ultimate benchmark is not accuracy, but

clinical consequence. In addition, the proposed framework has not been empirically validated, may be subject to selection bias in the referenced literature, and its applicability across different AI task types has not been specifically examined.

Although this validation gap is particularly evident in radiology, it reflects a broader challenge across medical AI applications. Similar discrepancies between technical performance and downstream clinical impact have been reported in other domains, including pathology, cardiology, and critical care, where implementation complexity and workflow integration remain limiting factors. However, radiology presents specific characteristics, such as its data-driven nature, high imaging volume, and central role in diagnostic pathways, which make it both especially suitable for AI development and particularly sensitive to failures in clinical translation.

Within this context, the role of the radiologist remains central. AI should be understood as an assistive tool rather than a replacement. Radiologists integrate imaging findings with clinical information, prior studies, and multidisciplinary input. They also act as supervisors, identifying potential errors and contextualizing algorithm outputs. This evolving role requires not only technical familiarity but also a critical understanding of model limitations. AI literacy is becoming an essential component of radiological expertise. Responsibility does not shift to the algorithm. The radiologist remains accountable for how AI outputs are interpreted and applied in patient care. In practice, this collaboration may involve shared decision thresholds, where AI outputs inform but do not determine reporting conclusions, as well as feedback loops in which radiologists' corrections contribute to ongoing model refinement. These interactions also have implications for training, requiring radiologists to develop skills in interpreting algorithmic outputs, uncertainty, and potential failure modes.

Bridging the validation gap requires a shift in both evaluation strategies and implementation approaches. Validation must extend beyond retrospective, single-center studies toward multicenter designs that incorporate real-world variability. Increasing attention should be given to workflow integration, ensuring that AI tools are embedded within existing systems rather than functioning as external add-ons. Prospective and longitudinal studies are needed to evaluate sustained performance and clinical impact over time. Regulatory frameworks are also evolving, with growing emphasis on post-deployment monitoring and adaptive systems[5]. Importantly, future research should focus not only on algorithm performance, but on human-AI collaboration, recognizing that clinical value emerges from their interaction rather than from either component alone.

In practice, the proposed framework may be used to structure the evaluation of AI systems across different stages of deployment, ensuring that technical performance, workflow integration, and clinical impact are assessed in a coordinated manner. This approach may help align validation strategies with clinically meaningful implementation.

This work has several limitations. As a narrative perspective, it does not provide a systematic or quantitative assessment of the literature, and the proposed framework is intended as a conceptual model rather than a validated evaluation tool. In addition, although increasing evidence supports the technical performance of AI systems, robust data demonstrating consistent downstream impact on clinical outcomes remains limited across many domains. These considerations should be taken into account when interpreting the proposed framework and its potential applications.

The challenge facing AI in radiology is not a lack of technological capability, but a difficulty in translating validated performance into meaningful clinical impact. The validation gap reflects a broader misalignment between algorithm-centered evaluation and patient-centered care. AI will undoubtedly shape the future of radiology, but its success will depend not on how well it performs in controlled environments, but on how effectively it integrates into the complexity of routine clinical practice.

Abbreviations

AI: artificial intelligence

AUC: area under the curve

PACS: picture archiving and communication system

Declarations

Author contributions

ANB: Conceptualization, Investigation, Writing—original draft, Writing—review & editing. The author read and approved the submitted version.

Conflicts of interest

The author declares that there are no conflicts of interest.

Ethical approval

Not applicable.

Consent to participate

Not applicable.

Consent to publication

Not applicable.

Availability of data and materials

Not applicable.

Funding

Not applicable.

Publisher's note

Open Exploration maintains a neutral stance on jurisdictional claims in published institutional affiliations and maps. All opinions expressed in this article are the personal views of the author(s) and do not represent the stance of the editorial team or the publisher.

References

1. Bhandari A. Revolutionizing Radiology With Artificial Intelligence. Cureus. 2024;16:e72646. [DOI] [PubMed] [PMC]
2. Warraich HJ, Tazbaz T, Califf RM. FDA Perspective on the Regulation of Artificial Intelligence in Health Care and Biomedicine. JAMA. 2025;333:241–7. [DOI] [PubMed]
3. Díaz O, Rodríguez-Ruíz A, Sechopoulos I. Artificial Intelligence for breast cancer detection: Technology, challenges, and prospects. Eur J Radiol. 2024;175:111457. [DOI] [PubMed]
4. Navarro-Ballester A. Artificial intelligence-driven radiological biomarkers: A narrative review of artificial intelligence in meningioma diagnosis. NeuroMarkers. 2025;2:100033. [DOI]
5. Ajmal CS, Yerram S, Abishek V, Nizam VPM, Aglave G, Patnam JD, et al. Innovative Approaches in Regulatory Affairs: Leveraging Artificial Intelligence and Machine Learning for Efficient Compliance and Decision-Making. AAPS J. 2025;27:22. [DOI] [PubMed]

6. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol. 2021;31: 3797–804. [DOI] [PubMed] [PMC]
7. Wu K, Wu E, Theodorou B, Liang W, Mack C, Glass L, et al. Characterizing the Clinical Adoption of Medical AI Devices through U.S. Insurance Claims. NEJM AI. 2024;1:1–13. [DOI]
8. van de Sande D, Chung EFF, Oosterhoff J, van Bommel J, Gommers D, van Genderen ME. To warrant clinical adoption AI models require a multi-faceted implementation evaluation. NPJ Digit Med. 2024;7: 58. [DOI] [PubMed] [PMC]
9. Gorenstein L, Soffer S, Apter S, Konen E, Klang E. AI in radiology: is it the time for randomized controlled trials? Eur Radiol. 2023;33:4223–5. [DOI] [PubMed]
10. Tanguay W, Acar P, Fine B, Abdolell M, Gong B, Cadrin-Chênevert A, et al. Assessment of Radiology Artificial Intelligence Software: A Validation and Evaluation Framework. Can Assoc Radiol J. 2022;74: 326–33. [DOI] [PubMed]
11. Yu AC, Mohajer B, Eng J. External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review. Radiol Artif Intell. 2022;4:e210064. [DOI] [PubMed] [PMC]
12. Abdar M, Khosravi A, Islam SMS, Acharya UR, Vasilakos AV. The need for quantification of uncertainty in artificial intelligence for clinical data analysis: increasing the level of trust in the decision-making process. IEEE Syst Man Cybern Mag. 2022;8:28–40. [DOI]
13. Marconi L, Cabitza F. Show and tell: A critical review on robustness and uncertainty for a more responsible medical AI. Int J Med Inform. 2025;202:105970. [DOI] [PubMed]
14. Ojha J, Presacan O, G. Lind P, Monteiro E, Yazidi A. Navigating Uncertainty: A User-Perspective Survey of Trustworthiness of AI in Healthcare. ACM Trans Comput Healthc. 2025;6:1–32. [DOI]
15. Berkowitz JS, Patock JR, Nawaz A, Gonzalez-Hernandez G, Tatonetti NP. A crisis of overconfidence: Why confidence, not accuracy, is the real risk in clinical AI. BioData Min. 2026;19:10. [DOI] [PubMed] [PMC]
16. Windecker D, Baj G, Shiri I, Kazaj PM, Kaesmacher J, Gräni C, et al. Generalizability of FDA-Approved AI-Enabled Medical Devices for Clinical Use. JAMA Netw Open. 2025;8:e258052. [DOI] [PubMed] [PMC]
17. Suleman MU, Mursaleen M, Khalil U, Saboor A, Bilal M, Khan SA, et al. Assessing the generalizability of artificial intelligence in radiology: a systematic review of performance across different clinical settings. Ann Med Surg. 2025;87:8803–11. [DOI] [PubMed] [PMC]
18. Hassoon A, Lin C, Woo HYJ, Irimia R, Marsteller JA, Li A, et al. Guiding artificial intelligence in public health and medicine with epidemiology: A lifecycle framework for mitigating AI misalignment. Ann Epidemiol. 2025;112:119–26. [DOI] [PubMed]
19. Blezek DJ, Olson-Williams L, Missert A, Korfiatis P. AI Integration in the Clinical Workflow. J Digit Imaging. 2021;34:1435–46. [DOI] [PubMed] [PMC]
20. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: Implications for clinical decision-making. PLOS Digit Health. 2024;3:e0000651. [DOI] [PubMed] [PMC]