

# Artificial Intelligence in Neuroradiology: Diagnostic Performance, Real-World Evidence, and Transferability — A Literature Review

Osinachi Okoye, MD<sup>1\*</sup>, Asia Filatov, MBA, MD, PhD<sup>2</sup>

<sup>1</sup>Northwest Health Family Medicine Residency program (PGY-2). Valparaiso, Indiana, USA.

<sup>2</sup>Northwest Health- Porter Neurology, Valparaiso, Indiana. University of Chicago, Chicago, Illinois, USA.

\*Corresponding Author: Asia Filatov, MBA, MD, PhD, Northwest Health- Porter Neurology, Valparaiso, Indiana. University of Chicago, Chicago, Illinois, USA.

DOI: <https://doi.org/10.58624/SVOANE.2026.07.044>

Received: October 06, 2026

Published: October 28, 2026

Citation: Okoye O, Filatov A. Artificial Intelligence in Neuroradiology: Diagnostic Performance, Real-World Evidence, and Transferability — A Literature Review. *SVOA Neurology* 2026, 7:5, 361-365. doi.org/10.58624/SVOANE.2026.07.044

## Abstract

In 2016, Geoffrey Hinton predicted that artificial intelligence (AI) would surpass radiologists within five years and that training new ones should stop. A decade later, this review tests that prediction against the accumulated evidence. To do so, studies were identified through a structured literature search of PubMed and EMBASE for English-language articles published between January 2021 and December 2026. Inclusion criteria were original research and reviews focused on AI application in neuroradiology, with exclusion of case reports, editorials, and non-peer-reviewed materials. This review integrates literature from 2021 to 2026 across four major application areas — acute stroke and intracranial hemorrhage (ICH) triage, brain tumor segmentation and classification, multiple sclerosis (MS) lesion detection, and large language models (LLMs) applied to radiology reporting — alongside two cross-cutting themes: real-world and registry-based evidence, and AI generalizability over sites and populations. Stroke and ICH triage represent the most mature application, with an actual deployment study of 101,944 CT scans across 17 facilities demonstrating 82.2% sensitivity and 97.6% specificity<sup>3</sup>, alongside documented real-world misclassification cases. LLMs perform impressively on narrow, structured reporting tasks but substantially less well on open-ended diagnostic reasoning. The single most consistent finding across the literature reports a systematic gap between internally and externally validated AI performance, with specificity declines of up to 24 percentage points on external datasets<sup>13</sup>, pointing to the need for mandatory external validation before broader clinical deployment.

**Keywords:** Artificial Intelligence; Deep Learning; Neuroradiology; Stroke; Intracranial Hemorrhage; Large Language Models; External Validation

## 1. Introduction

In 2016, computer scientist Geoffrey Hinton — now widely regarded as one of the founding figures of modern deep learning — told an audience of physicians that institutions should stop training radiologists, predicting that artificial intelligence would match or exceed human performance in image interpretation within five years. A decade later, that prediction offers a natural benchmark against which to test what AI in neuroradiology has actually delivered.

This review integrates literature from 2021 through 2026 spanning four major application areas: acute stroke and intracranial hemorrhage (ICH) triage, and large language models (LLMs) applied to radiology report generation and quality assurance, alongside two cross-cutting themes that determine whether these tools translate into genuine clinical benefit: real-world, registry-scale evidence, and the persistent problem of AI generalizability over sites, scanners, and populations. Measured against Hinton's forecast, the picture that emerges is neither the wholesale replacement he predicted nor the dismissal his critics initially offered, but something more specific and, in its own way, more interesting: a technology that has become a genuinely useful triage and workflow tool while remaining a poor fit for autonomous diagnostic authority, for reasons the literature itself explains. For radiology students and trainees, these findings highlight the need to develop competency not only in conventional image interpretation, but also in understanding, critically appraising, and efficiently integrating AI tools into clinical practice. Familiarity with the strengths and limitations of AI in neuroradiology will become a core aspect of future training and day-to-day decision-making.

## 2. Technical Background

Most neuroradiology AI applications reviewed here are built on convolutional neural network (CNN) architectures — U-Net and its variants for segmentation tasks, ResNet- and VGG-family classifiers for lesion detection and classification — trained on labeled imaging datasets to learn hierarchical spatial attributes associated with a target label (hemorrhage present/absent, tumor sub-region, lesion boundary). Since 2023, transformer-based models have been adapted from natural language processing. Since 2023, transformer-driven architectures adapted from natural language processing (Visual Transformers) and, separately, large language models developed with text have been increasingly incorporated, either as standalone classifiers or as components of multimodal pipelines that fuse imaging with clinical or textual data.

A handful of recent studies illustrate where this is heading. Zeineldin and colleagues built a hybrid Vision Transformer-CNN model that segments gliomas across multiple MRI sequences at once, pairing the CNN's strength at picking up fine local detail with the Transformer's ability to track longer-range patterns across the image.[16] Around the same time, Zakariah and colleagues took a similar dual-architecture approach with their Vision Transformer-DSUNET model, reaching a combined Dice score of 91.29% across all tumor sub-regions.[17] Transformers have also started pulling in information beyond the image itself. Usuzaki and colleagues fed a variable Vision Transformer patient demographics, radiomic features, and MRI together to predict IDH mutation status in diffuse glioma; the model was perfect on its own internal test set but dropped to 93.5% on an outside dataset — a gap that says a lot about why external validation matters before any of this reaches the clinic.[18] On the language side, Kanzawa and colleagues fine-tuned a BERT-based model that could sort brain MRI reports into categories with real accuracy, and Schramm and colleagues found that GPT-4V did noticeably better on hard brain MRI cases once it had the patient's clinical history and a written description of the imaging to work with, not just the pictures alone.[19] Taken together, these studies point in a consistent direction: transformer-based and multimodal approaches keep showing real promise in neuroradiology, but nearly every one of them comes with the same caveat — independent, external validation is still the piece most often missing before any of this is ready for routine clinical use.

These concrete examples show how next-generation models are being used to tackle complex neuroradiology tasks. A recurring methodological distinction throughout this literature is between internal validation (testing a model on data drawn from the same institution, scanner, or dataset used for training) and external validation (testing on genuinely independent data), a distinction central to the generalizability findings discussed below.

## 3. Stroke and Intracranial Hemorrhage Triage

Stroke and ICH triage represent the most clinically mature application of AI in neuroradiology, with the largest base of FDA-cleared tools and actual performance data. A diagnostic performance study of a commercially available deep learning tool for detecting acute ischemic lesions on brain MRI in 995 suspected-stroke patients found high sensitivity and specificity relative to an experienced neuroradiologist, with sensitivity varying by lesion characteristics and specificity varying with image quality.[1] A 2025 systematic review and hierarchical summary ROC meta-analysis of deep-learning-based non-contrast CT interpretation for acute ischemic stroke concluded that these tools can meaningfully support triage and worklist prioritization, while noting frequent methodological weaknesses across the underlying literature, including unclear reference standards and train/test leakage.[2]

A large-scale real-world evaluation of a widely deployed FDA-cleared ICH triage model analyzed 101,944 non-contrast head CT examinations from 74,142 patients across a 17-facility academic health system, demonstrating 82.2% sensitivity, 97.6% specificity, and 96.6% overall accuracy, with sensitivity highest for large, multi-compartment hemorrhages and notably lower for smaller or more subtle bleeds.[3] A case report published in *Radiology* documents a real-world failure mode in which an FDA-cleared algorithm misclassified an ischemic stroke as ICH, prompting the authors to review the FDA's regulatory process for AI/ML-based software as a medical device.[4] This documented misclassification is an important counterweight to the strong aggregate accuracy figures reported elsewhere, as well as reinforcing that these tools function as triage aids rather than autonomous diagnostic replacements.

## 4. Large Language Models in Radiology Reporting

Since the wider availability of GPT-4 in 2023, a substantial body of literature has tested LLMs on radiology-specific tasks. In neuroradiology specifically, a study testing GPT-4 against 140 diagnostic cases from a major neuroradiology journal's teaching archive found an overall diagnostic performance of 57.86%, materially lower than the near-ceiling performance reported for narrower diagnostic tasks elsewhere in this review, indicating the greater difficulty of open-ended diagnostic reasoning compared to binary lesion detection.[5] Beyond diagnosis, LLMs have shown strong performance on report infrastructure tasks: GPT-4 detected clinically significant speech-recognition transcription errors in radiology reports with 76.9% exactness and 100% recall.[6] A systematic review of LLM applications to radiology reports found the literature concentrated heavily on chest radiography (60% of studies) relative to MRI or CT, with substantial heterogeneity in evaluation methodology limiting cross-study comparability.[7,8]

## 5. Real-World and Registry-Based Evidence

Large federated real-world electronic health record networks have enabled several population-scale studies relevant to AI-era stroke care pathways. A retrospective analysis of 90,740 patients examined outcomes of endovascular thrombectomy in patients with low NIHSS scores, a population underrepresented in pivotal randomized trials.[9] A separate cohort study examined the impact of atrial fibrillation on intracerebral hemorrhage and mortality after thrombectomy.[10] These registry-scale datasets are well suited to evaluating downstream clinical outcomes of AI-accelerated care pathways at a scale no single-center prospective study can match. However, reliance on aggregated real-world data introduces important limitations, including inaccuracies in clinical coding, heterogeneity across contributing sites, a lack of granular clinical detail, and residual confounding that can continue despite careful propensity score matching.[11,12] Selection bias may arise if data are missing non-randomly or if participating sites differ in ways that affect both AI deployment and clinical outcomes. Furthermore, temporal changes in clinical practice and data collection standards can complicate interpretation of trends over time. Dealing with these shortcomings in subsequent research will require combining registry data with richer sources, such as imaging archives and detailed chart reviews, applying standardized data definitions across sites, and planning prospective observational studies where AI exposure is clearly documented. Notably, no available registry-based study has used AI deployment itself, rather than a downstream treatment-pathway variable, as its primary exposure, which is a clear gap in the current literature given the demonstrated scale of these platforms.

## 6. Generalizability and External Validation

A recurring and arguably the single most important finding across this literature is that internally validated AI performance in neuroimaging systematically overstates externally validated, real-world performance. A systematic review of AI diagnostic models across CT, MRI, and radiography found internal validation AUCs ranging from 0.76 to 0.95. Still, external validation showed a median AUC decline of approximately 0.03 and a substantially larger decline in specificity of up to 24 percentage points on some external datasets.[13] A related systematic review found that 70 of 86 (81%) deep learning radiology algorithms underperformed on external datasets, with 24% showing substantial performance diminishment.<sup>14</sup> Underlying this pattern, deep learning studies frequently fail to report the demographic or clinical composition of their development datasets in sufficient detail to assess selection bias.[15]

## 7. Conclusion

AI has achieved genuine, measurable clinical value in neuroradiology, most convincingly in stroke and ICH triage, where large-scale real-life deployment data now exist alongside ongoing evidence of imperfect performance in practice. In other domains, benchmark-level accuracy figures should be interpreted with caution pending external validation, and the field's most pressing methodological need is a shift from internally validated accuracy reporting to prospective, externally validated, clinical-workflow-integrated evaluation. A decade on from Hinton's prediction, the evidence assembled here suggests the forecast was directionally right about AI's diagnostic potential but wrong about its trajectory: the technology did not replace radiologists on any five-year timeline, and the reasons why—an internal-to-external validation gap, a continuing need for human monitoring of triage flags, and a performance ceiling on open-ended reasoning that narrow pattern recognition was never designed to clear—are themselves the most durable findings of this review. Going forward, future research needs to focus on formulating robust, generalizable models through multicenter collaboration and broad datasets, as well as creative approaches that improve explainability and address bias in clinical practice. Advances in regulatory science, ongoing assessment of long-term clinical outcomes of AI deployment, and integration of multimodal data—including imaging, clinical notes, and genomics—constitute promising directions for the field. For students and emerging researchers, there are important opportunities to contribute to these problems as well as lead the next wave of evidence-based innovation in neuroradiology.

## Conflict of Interest

The authors declare no conflicts of interest.

## Funding

None.

## References

1. Krag, C. H., Müller, F. C., Gandrup, K. L., Raaschou, H., Andersen, M. B., Brejnebo, M. W., Sagar, M. V., Bojsen, J. A., Rasmussen, B. S., Graumann, O., Nielsen, M., Kruuse, C., & Boesen, M. (2023). Diagnostic test accuracy study of a commercially available deep learning algorithm for ischemic lesion detection on brain MRIs in suspected stroke patients from a non-comprehensive stroke center. *European Journal of Radiology*, 168, 111126. <https://doi.org/10.1016/j.ejrad.2023.111126>
2. Gete, K. Y., & Ayele, A. A. (2025). Deep-learning-based non-contrast CT for detecting acute ischemic stroke: A systematic review and HSROC meta-analysis of patient-level diagnostic accuracy. *BMC Neurology*, 25(1), 507. <https://doi.org/10.1186/s12883-025-04528-3>
3. Chavoshi, M., Mansuri, A., Bala, W., Brown-Mulry, B., Dapamede, T., Isaac, R., Khosravi, B., et al. (2025). Real-world performance evaluation of a commercial deep learning model for intracranial hemorrhage detection. *NPJ Digital Medicine*, 9(1), 66. <https://doi.org/10.1038/s41746-025-02244-3>
4. Zhang, K., Khosravi, B., Vahdati, S., & Erickson, B. J. (2023). FDA review of radiologic AI algorithms: Process and challenges. *Radiology*, 310(1). <https://doi.org/10.1148/radiol.230242>
5. Suthar, P. P., Kounsai, A., Chhetri, L., Saini, D., & Dua, S. G. (2023). Artificial intelligence (AI) in radiology: A deep dive into ChatGPT 4.0's accuracy with the American Journal of Neuroradiology's (AJNR) "Case of the Month." *Cureus*, 15(8), e43958. <https://doi.org/10.7759/cureus.43958>
6. Schmidt, R. A., Seah, J. C. Y., Cao, K., Lim, L., Lim, W., & Yeung, J. (2024). Generative large language models for detection of speech recognition errors in radiology reports. *Radiology: Artificial Intelligence*, 6(2), e230205. <https://doi.org/10.1148/ryai.230205>
7. Artsi, Y., Klang, E., Collins, J. D., Glicksberg, B. S., Nadkarni, G. N., Korfiatis, P., & Sorin, V. (2025). Large language models in radiology reporting: A systematic review of performance, limitations, and clinical implications. *Intelligence-Based Medicine*, 12, 100287. <https://doi.org/10.1016/j.ibmed.2025.100287>

8. Lee, R. C., Hadidchi, R., Coard, M. C., Rubinov, Y., Alamuri, T., Liaw, A., Chandrupatla, R., & Duong, T. Q. (2026). Use of large language models on radiology reports: A scoping review. *Journal of the American College of Radiology*, 23(3), 437–454. <https://doi.org/10.1016/j.jacr.2025.10.005>
9. Malhotra, A., Thorman, I., Kiss, A., Sacknovitz, A., Spirollari, E., Jain, A., & Al-Mufti, F. (2025). Outcomes of endovascular thrombectomy for stroke in patients with low NIHSS scores: A retrospective cohort study of 90,740 patients [Conference abstract]. *Neurology*, 104(7\_Suppl\_1), 5565. <https://doi.org/10.1212/WNL.000000000000212447>
10. Alobaida, M., Harrison, S. L., Lane, D. A., Underhill, P., Hill, A., & Lip, G. Y. H. (2023). Outcomes in patients with ischaemic stroke undergoing endovascular thrombectomy: Impact of atrial fibrillation. *Journal of Stroke and Cerebrovascular Diseases*, 32(2), 106917. <https://doi.org/10.1016/j.jstrokecerebrovasdis.2022.106917>
11. Nassar, M., Abosheaishaa, H., Elfert, K., Beran, A., Ismail, A., Mohamed, M., Misra, A., Essibayi, M. A., Altschul, D. J., & Azzam, A. Y. (2025). TriNetX and real-world evidence: A critical review of its strengths, limitations, and bias considerations in clinical research. *ASIDE Internal Medicine*, 1(2), 24–33. <https://doi.org/10.71079/aside.im.03222516>
12. Ludwig, R. J., Anson, M., Zirpel, H., Thaci, D., Olbrich, H., Bieber, K., Kridin, K., Dempfle, A., Curman, P., Zhao, S. S., & Alam, U. (2025). A comprehensive review of methodologies and application to use the real-world data and analytics platform TriNetX. *Frontiers in Pharmacology*, 16, 1516126. <https://doi.org/10.3389/fphar.2025.1516126>
13. Suleman, M. U., Mursaleen, M., Khalil, U., Saboor, A., Bilal, M., Khan, S. A., Subhani, M. A., Hussnain, M. A., Tabassum, S. N., & Tahir, M. (2025). Assessing the generalizability of artificial intelligence in radiology: A systematic review of performance across different clinical settings. *Annals of Medicine and Surgery (London)*, 87(12), 8803–8811. <https://doi.org/10.1097/MS9.00000000000004166>
14. Soliman, D. (2026). Beyond benchmarks: A framework for post-deployment validation of CT lung nodule detection AI. *arXiv*. <https://arxiv.org/abs/2603.26785>
15. Yu, A. C., Mohajer, B., & Eng, J. (2022). External validation of deep learning algorithms for radiologic diagnosis: A systematic review. *Radiology: Artificial Intelligence*, 4(3), e210064. <https://doi.org/10.1148/ryai.210064>
16. Zeineldin, R. A., Karar, M. E., Elshaer, Z., Coburger, J., Wirtz, C. R., Burgert, O., & Mathis-Ullrich, F. (2024). Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI. *Scientific Reports*, 14(1), 3713. <https://doi.org/10.1038/s41598-024-54186-7>
17. Zakariah, M., Al-Razgan, M., & Alfakih, T. (2024). Dual vision Transformer-DSUNET with feature fusion for brain tumor segmentation. *Heliyon*, 10(18), e37804. <https://doi.org/10.1016/j.heliyon.2024.e37804>
18. Usuzaki, T., Inamori, R., Shizukuishi, T., Morishita, Y., Takagi, H., Ishikuro, M., Obara, T., & Takase, K. (2024). Predicting isocitrate dehydrogenase status among adult patients with diffuse glioma using patient characteristics, radiomic features, and magnetic resonance imaging: Multi-modal analysis by variable vision transformer. *Magnetic Resonance Imaging*, 111, 266–276. <https://doi.org/10.1016/j.mri.2024.05.012>
19. Kanzawa, J., Yasaka, K., Fujita, N., Fujiwara, S., & Abe, O. (2024). Automated classification of brain MRI reports using fine-tuned large language models. *Neuroradiology*, 66(12), 2177–2183. <https://doi.org/10.1007/s00234-024-03427-7>
20. Schramm, S., Preis, S., Metz, M.-C., Jung, K., Schmitz-Koep, B., Zimmer, C., Wiestler, B., Hedderich, D. M., & Kim, S. H. (2025). Impact of multimodal prompt elements on diagnostic performance of GPT-4V in challenging brain MRI cases. *Radiology*, 314(1), e240689. <https://doi.org/10.1148/radiol.240689>

**Copyright:** © 2026 All rights reserved by Okoye O and Filatov A. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.