

International Journal of Applied Technology in Medical Sciences
Vol 5 Issue 1 (2026)
Pages 4 - 12
Available at
www.emiratesscholar.com

Diagnostic Performance of Deep Learning in Medical Imaging: An Analytical Review and Quantitative Synthesis of the Evidence

Wei Zhang[1], Chinedu Okafor[2]
[1]Tsinghua University - China, [2]University of Nigeria - Nigeria

ARTICLE INFO

KEYWORDS
deep learning; medical imaging; diagnostic accuracy; sensitivity and specificity; meta-analysis; external validation

HOW TO CITE
Zhang, W., & Okafor, C. (2026). Diagnostic Performance of Deep Learning in Medical Imaging: An Analytical Review and Quantitative Synthesis of the Evidence. International Journal of Applied Technology in Medical Sciences, 5(1), 4-12.
https://doi.org/10.54878/mcr68n77

ABSTRACT

Deep learning has produced a large body of studies reporting expert-level accuracy in the diagnostic interpretation of medical images, yet the headline numbers are easily misread. This analytical review examines the quantitative evidence with the tools of diagnostic-test methodology, asking not only how accurate deep-learning models are but what their reported accuracy means and how far it can be trusted. We define and apply the core metrics, sensitivity, specificity, the area under the receiver operating characteristic curve (AUC), Youden's index, likelihood ratios, and the diagnostic odds ratio, and we assemble reported pooled estimates from the principal meta-analyses across imaging domains. The best available synthesis of externally validated studies places the pooled sensitivity of deep-learning models at 87.0% and specificity at 92.5%, statistically indistinguishable from the 86.4% and 90.5% of health-care professionals assessed on the same samples, while domain-specific meta-analyses report AUCs from roughly 0.86 for respiratory imaging to above 0.93 for retinal disease. Behind these figures, however, lies a consistent methodological problem that our analysis foregrounds: extreme between-study heterogeneity, a scarcity of external validation and prospective testing, small and non-representative human comparators, poor adherence to reporting standards, and susceptibility to dataset shift and confounding, all of which tend to inflate apparent performance. We argue that the central analytical lesson is a gap between reported diagnostic accuracy and demonstrated clinical validity, and we set out the metrics, study designs, and reporting practices, including AI-specific reporting guidelines and prospective external validation, needed to close it. The review is quantitative and methodological in emphasis and is weighted toward classification tasks in radiology, ophthalmology, pathology, and dermatology.

Double-blind peer review by independent reviewers.
Received: 26 Jan 2026, Accepted: 24 May 2026, Published: 25 Jun 2026
*Corresponding author: wei.zhang@tsinghua.edu.cn
Copyright: © 2026 by the author. Licensee Emirates Scholar Center for Research & Studies, UAE.
Open access under the Creative Commons Attribution (CC BY) license
https://creativecommons.org/licenses/by/4.0

1. Introduction

Few claims in contemporary medicine have been made as often, or as confidently, as the claim that deep-learning algorithms can match or exceed expert clinicians at reading medical images. Since the demonstration that a convolutional neural network could detect diabetic retinopathy from retinal photographs with an area under the receiver operating characteristic curve of 0.99 (Gulshan et al., 2016) and classify skin lesions at dermatologist level (Esteva et al., 2017), the literature has grown into thousands of studies spanning radiology, ophthalmology, pathology, and dermatology, and the rhetoric surrounding it has grown faster still. For an applied-medical-technology audience the pressing question is not whether such results exist but how they should be interpreted: what exactly the reported metrics measure, how consistent they are, and how far the accuracy demonstrated in a study predicts usefulness in the clinic.

This is an analytical rather than a descriptive question, and it is the one this review takes up. Diagnostic-test evaluation has a mature methodology, developed over decades for laboratory and imaging tests, and applying that methodology to deep-learning studies both clarifies what has been achieved and exposes what has not. Our aim is threefold: to define precisely the metrics by which diagnostic performance is judged; to assemble and interpret the quantitative evidence, drawing on the principal meta-analyses that pool results across studies; and to analyse the methodological features, heterogeneity, external validation, comparator quality, reporting, and bias, that determine whether a reported number can be believed. Throughout, we distinguish sharply between two things that the enthusiastic literature tends to conflate, diagnostic accuracy as measured on a dataset and clinical validity as demonstrated in practice, and we argue that the gap between them is the field's central unresolved problem. The analysis is directly relevant to health systems worldwide, including in low-resource and Global South settings where the promise of automated image interpretation is greatest and the evidence base thinnest, and it connects to the broader engagement of clinical communities with AI-supported diagnosis (Sulthan et al., 2025).

2. Metrics of Diagnostic Performance

A disciplined reading of the evidence begins with the metrics, because much confusion stems from their loose use. For a binary classifier compared against a reference standard, the fundamental quantities are derived from the counts of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Sensitivity and specificity are then

$$\text{Sensitivity} = TP / (TP + FN), \quad \text{Specificity} = TN / (TN + FP),$$

the proportions of diseased and non-diseased cases correctly classified. Because these two move in opposite directions as the decision threshold changes, a single pair describes only one operating point; the receiver operating characteristic (ROC) curve traces the pair across all thresholds, and the area under it (AUC) summarizes discrimination as the probability that the model ranks a random diseased case above a random non-diseased one, with 0.5 denoting chance and 1.0 perfect separation. A convenient one-number summary of a chosen operating point is Youden's index,

$$J = \text{Sensitivity} + \text{Specificity} - 1,$$

which ranges from 0 to 1. For clinical reasoning the likelihood ratios are more directly useful, since they update the odds of disease: $LR+ = \text{Sensitivity} / (1 - \text{Specificity})$ and $LR- = (1 - \text{Sensitivity}) / \text{Specificity}$. Their ratio is the diagnostic odds ratio,

$$DOR = LR+ / LR- = (\text{Sensitivity} \times \text{Specificity}) / ((1 - \text{Sensitivity})(1 - \text{Specificity})),$$

a single global measure of discriminatory power that is convenient for meta-analysis but that, by collapsing sensitivity and specificity into one number, hides the trade-off clinicians actually care about. Two properties of these metrics are essential to interpretation and are frequently neglected. First, sensitivity and specificity are properties of the test, but predictive values, the probabilities a clinician wants, depend on disease prevalence, so a model with excellent specificity can still generate mostly false alarms when disease is rare. Second, AUC is threshold-independent and prevalence-independent, which makes it a good measure of discrimination but a poor guide to clinical utility at any particular operating point. Table 1 collects these definitions.

Table 1. Core metrics of diagnostic performance.

Metric	Definition	What it captures / caveat
Sensitivity (recall)	TP / (TP + FN)	Detection of disease; ignores false positives.

ISSN: 3014-5197 - https://doi.org/10.54878/mcr68n77

Specificity	TN / (TN + FP)	Correct exclusion; ignores false negatives.
AUC (ROC)	$P(score_+ > score_-)$	Threshold-free discrimination; not utility.
Youden's J	Se + Sp − 1	Single operating-point summary (0–1).
LR+ / LR−	Se/(1−Sp) · (1−Se)/Sp	Updates odds of disease; clinically intuitive.
Diagnostic odds ratio	LR+ / LR−	Global discrimination; hides Se/Sp trade-off.
Predictive values	Depend on prevalence	What clinicians see; vary with setting.

2.1 A worked example: why prevalence changes everything

The prevalence dependence of predictive values is worth making concrete, because it is the single most common source of over-optimism when a model moves from study to clinic. Consider a model with sensitivity 0.90 and specificity 0.95, respectable and typical figures. Its positive predictive value, the probability that a flagged case truly has disease, follows from Bayes' rule,

$$PPV = (Se \times p) / (Se \times p + (1 - Sp)(1 - p)) ,$$

where p is disease prevalence. At the 50% prevalence of an enriched research dataset, this model achieves a PPV of about 0.95; but at the 1% prevalence typical of a screening population, the same model, with identical sensitivity and specificity, achieves a PPV of only about 0.15, meaning roughly six of every seven positive flags are false alarms. Nothing about the model has changed; only the setting has. Table 2 traces this collapse across prevalences and illustrates why a model validated on a balanced dataset can disappoint in a real screening programme, and why sensitivity and specificity, not predictive values, are the transferable properties that studies should report.

Table 2. Positive and negative predictive value of a model (Se 0.90, Sp 0.95) across disease prevalence.

Prevalence	PPV	NPV	Setting
50%	0.95	0.91	Balanced research set
10%	0.67	0.99	Enriched clinic
5%	0.49	0.99	High-risk group
1%	0.15	1.00	General screening

3. Quantitative Synthesis of Reported Performance

We now assemble the quantitative evidence, drawing on the meta-analyses that have pooled results across the primary literature, and we interpret it with the caveats developed above. The most rigorous synthesis to date restricted its pooled analysis to studies with out-of-sample external validation that compared models and clinicians on the same samples, and found a pooled sensitivity of 87.0% (95% CI 83.0–90.2) and specificity of 92.5% (95% CI 85.1–96.4) for deep-learning models, against 86.4% (79.9–91.0) and 90.5% (80.6–95.7) for health-care professionals, a difference that is not statistically significant (Liu et al., 2019). This is the single most important number in the field, and it supports a carefully bounded claim: on the best-designed comparisons available, deep-learning models are roughly equivalent to, not dramatically better than, expert clinicians.

Domain-specific meta-analyses fill in the picture and reveal wide variation. A synthesis of 503 studies reported pooled AUCs between 0.933 and 1.0 for ophthalmic diagnoses on fundus photographs and optical coherence tomography, between 0.864 and 0.937 for lung nodules and cancer on chest radiography and CT, and between 0.868 and 0.909 for breast cancer across mammography, ultrasound, MRI, and tomosynthesis (Aggarwal et al., 2021). A focused meta-analysis of breast and cervical cancer detection pooled a sensitivity of 88% (85–90), a specificity of 84% (79–87), and an AUC of 0.92 (0.90–0.94) (Xue et al., 2022). Figure 1 places the reported operating points in ROC space, and Figure 2 shows representative pooled AUCs by domain; Table 2 gives the numbers. Two patterns are analytically important. First, discrimination

ISSN: 3014-5197 - https://doi.org/10.54878/mcr68n77

is genuinely high across domains, with most AUCs near or above 0.90. Second, performance is far from uniform: retinal imaging, with its standardized acquisition and clear targets, sits near the top, while tasks with noisier images and more ambiguous targets sit lower, a gradient that reflects the difficulty of the task as much as the power of the method.

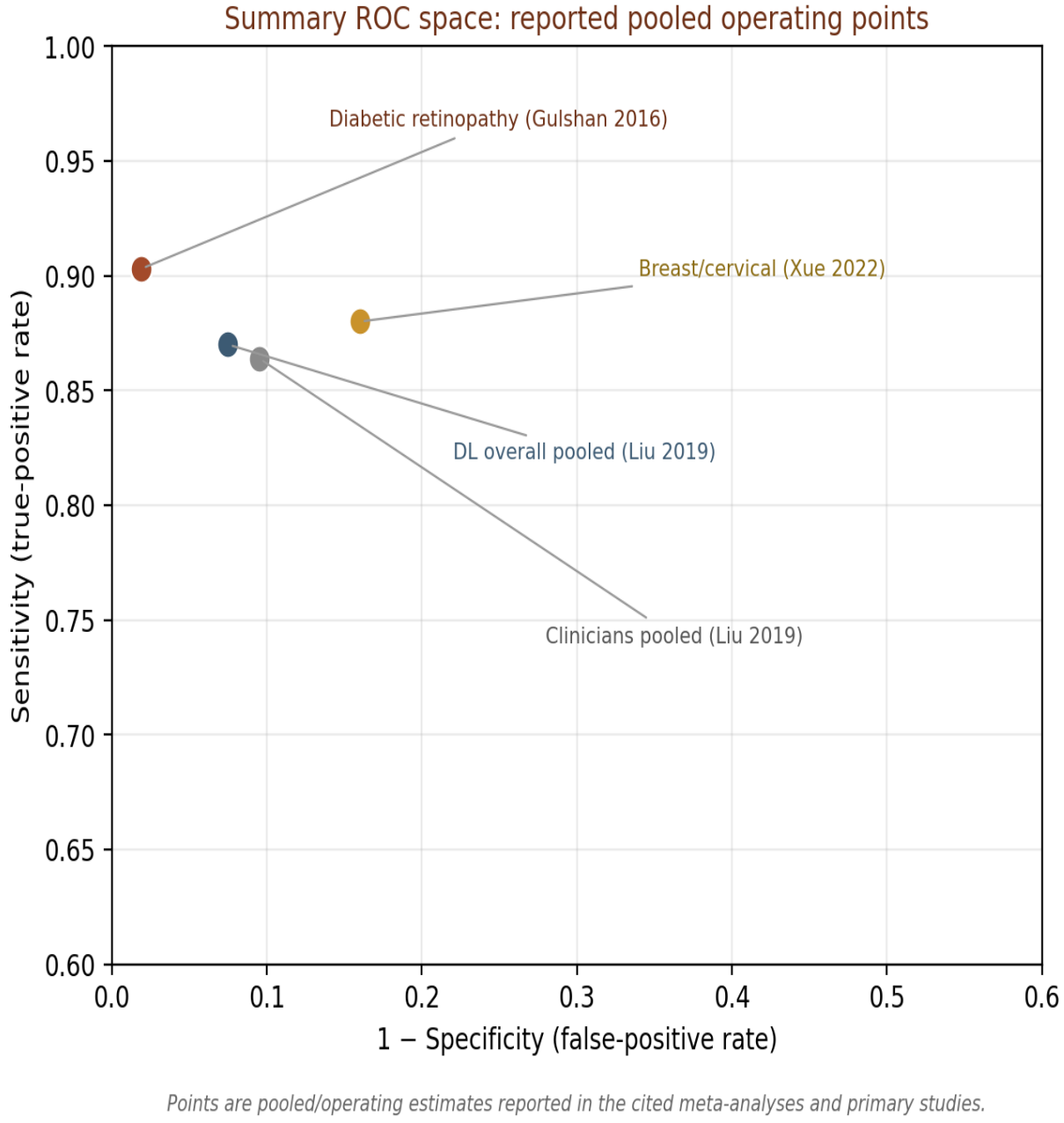


Figure 1. Reported pooled operating points in summary ROC space; deep-learning and clinician estimates are closely comparable (Liu et al., 2019; Gulshan et al., 2016; Xue et al., 2022).

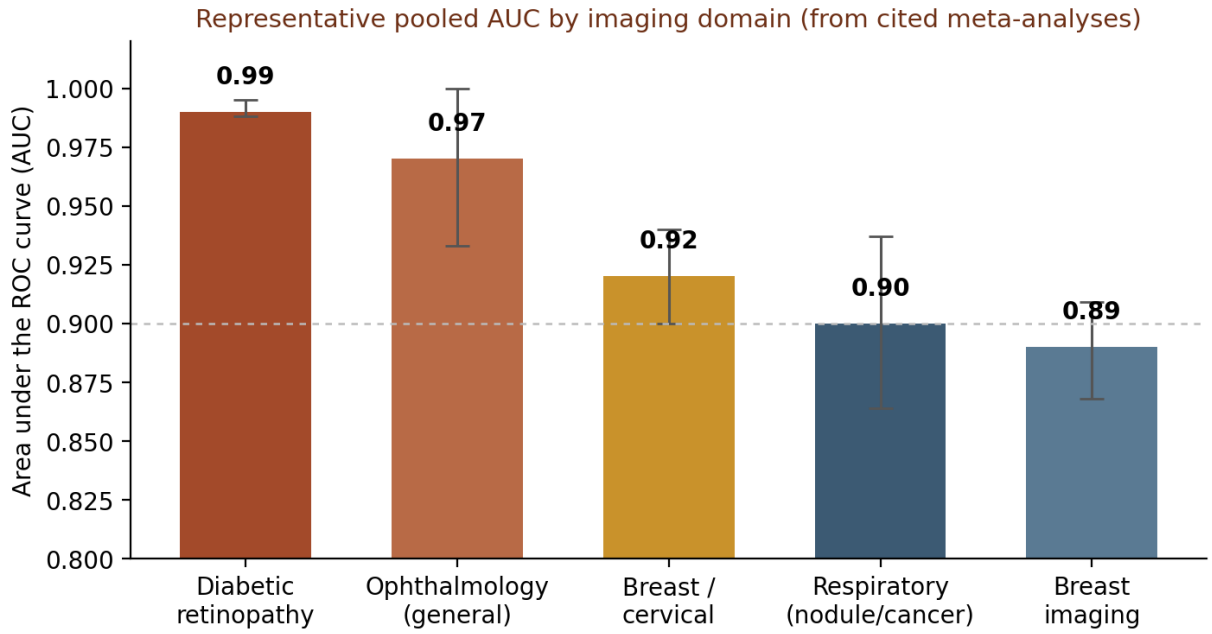


Figure 2. Representative pooled AUC by imaging domain, with reported ranges (Aggarwal et al., 2021; Xue et al., 2022; Gulshan et al., 2016).

It is instructive to look past the pooled averages at individual landmark studies, because they show both the ceiling of what is achievable and the conditions required to approach it. In breast-cancer screening, a large international evaluation reported that an AI system reduced both false positives and false negatives relative to radiologists, with absolute reductions of several percentage points in each, and performed comparably to a double-reading protocol, results obtained on large, multi-site datasets that lend them unusual credibility (McKinney et al., 2020). In digital pathology, where whole-slide images are enormous and expert time is scarce, deep learning has advanced from proof of concept toward clinical-grade tools, though authorities in the field emphasize that the path to routine deployment runs through rigorous validation and regulatory clearance rather than benchmark accuracy alone (van der Laak, Litjens, & Ciompi, 2021). These exemplary studies matter analytically not because they are typical, they are not, but because they define what adequately validated evidence looks like, and thereby throw the weaknesses of the median study into relief.

Table 3. Reported pooled diagnostic performance by domain and source.

Domain / task	Reported estimate	Source
Overall (externally validated)	Se 87.0%, Sp 92.5%	Liu et al. (2019)
Clinicians (same samples)	Se 86.4%, Sp 90.5%	Liu et al. (2019)
Diabetic retinopathy	AUC 0.991; Se 90.3%, Sp 98.1%	Gulshan et al. (2016)
Ophthalmology (general)	AUC 0.933–1.0	Aggarwal et al. (2021)
Respiratory (nodule/cancer)	AUC 0.864–0.937	Aggarwal et al. (2021)
Breast imaging	AUC 0.868–0.909	Aggarwal et al. (2021)
Breast / cervical cancer	Se 88%, Sp 84%, AUC 0.92	Xue et al. (2022)

4. Heterogeneity and the Limits of the Numbers

The pooled estimates above must be read against the variability behind them, and here the analysis turns from optimism to caution. Every major meta-analysis reports substantial between-study heterogeneity, formally the proportion of variation across studies that exceeds what chance would produce, often summarized by the I^2 statistic,

$I^2 = (Q - df) / Q \times 100\%$,

where Q is Cochran's heterogeneity statistic and df its degrees of freedom; values above 75% are conventionally deemed high, and diagnostic meta-analyses in this field routinely exceed that. High heterogeneity means the pooled number is an average over studies so different in population, imaging modality, reference standard, and threshold that the average may describe no real-world setting well, which is why sophisticated diagnostic meta-analyses fit bivariate or hierarchical summary-ROC models rather than pooling a single sensitivity and specificity. The practical consequence is that a headline pooled AUC should be treated as a rough indication of what is achievable under favourable conditions, not as a performance guarantee.

The methodology used to pool such data is itself worth understanding, because it shapes what the summary numbers mean. Naively averaging sensitivities and specificities across studies is inappropriate, since the two are negatively correlated through the threshold each study chose; the accepted approach is the bivariate random-effects model, or the equivalent hierarchical summary-ROC (HSROC) model, which jointly estimates sensitivity and specificity while allowing both to vary randomly across studies and modelling their correlation. These models yield a summary operating point with a confidence region and a summary ROC curve rather than a single misleading pair, and they are what the more rigorous meta-analyses in this field employ. Even they, however, cannot repair a biased evidence base. Publication bias, the tendency for flattering results to be published and disappointing ones to vanish, is difficult to detect in diagnostic meta-analyses and is rarely formally assessed, and where it operates it inflates every pooled estimate. The analyst's stance toward a headline number should therefore be conditioned less by its magnitude than by the quality and completeness of the studies behind it.

Heterogeneity is compounded by a deeper problem of internal and external validity. A systematic appraisal of studies comparing deep learning with clinicians found that few were prospective, most were at high risk of bias, human comparator groups were small (a median of four experts), data and code were almost never available, and yet the large majority claimed performance at least comparable to clinicians, often without calling for the further prospective validation that their design demanded (Nagendran et al., 2020). The scarcity of external validation is especially consequential because deep-learning models are prone to learn spurious, dataset-specific correlations; a model trained to detect pneumonia on chest radiographs from one hospital system was shown to generalize poorly to others, having in part learned to recognize the hospital rather than the disease (Zech et al., 2018). Reported accuracy measured on internal test data therefore systematically overstates the accuracy a model will achieve on new populations, and the size of that optimism is frequently unmeasured because the external test is never done.

One further subtlety deserves an analyst's attention: the reference standard against which a model is judged is itself imperfect, and this quietly distorts the metrics. Deep-learning studies are frequently evaluated against expert labels rather than against an independent ground truth such as biopsy or long-term follow-up, so a model is really being measured on its agreement with clinicians, not on its correctness. When the reference is the same kind of judgement the model is meant to replace, two biases follow. If the model was trained on labels from the very experts who then serve as the standard, their shared errors are invisible, flattering the apparent accuracy; and where the reference standard is genuinely uncertain, disagreements that may reflect model correctness are scored as model errors. Rigorous evaluation therefore favours hard endpoints where they exist, and treats agreement with clinicians as a weaker form of evidence than concordance with outcomes. This is not a peripheral quibble: it means that some reported accuracy reflects the reproduction of human judgement, including its mistakes, rather than a superior reading of the underlying biology, and distinguishing the two requires reference standards the literature too rarely supplies.

5. From Diagnostic Accuracy to Clinical Validity

The analysis so far supports a single organizing conclusion: the field has abundant evidence of diagnostic accuracy and sparse evidence of clinical validity, and the two are not the same. Accuracy is a property of a model on a dataset; clinical validity is a demonstrated effect on the decisions clinicians make and the outcomes patients experience, and it depends on factors that accuracy metrics do not capture, how the model performs on the local population, how its output is presented, whether clinicians act on it appropriately, and whether the net effect on care is positive. Bridging the two requires a sequence that most of the literature has not yet completed: external validation on representative data, prospective evaluation in the clinical workflow, and ultimately randomized comparison of outcomes with and without the model (Kelly et al., 2019). Randomized trials of diagnostic AI remain rare, and the reporting standards needed to make even non-randomized studies

ISSN: 3014-5197 - https://doi.org/10.54878/mcr68n77

trustworthy have had to be created specifically for the field, in the form of AI-specific extensions to established guidelines (Liu et al., 2020).

Two further considerations bear directly on validity and on equity. First, calibration, whether a model's stated probabilities match observed frequencies, matters as much as discrimination for clinical use, yet it is reported far less often than AUC; a model can discriminate well while being poorly calibrated, and clinicians who trust its probabilities will then be misled. Second, performance measured in one population may not transfer to another, and because training data are drawn disproportionately from a few well-resourced health systems, models may underperform precisely in the settings, including many in the Global South, where they are most needed, quietly widening rather than narrowing disparities. These are not reasons for dismissal but specifications of the work required: the value of a model is established not by its AUC on a benchmark but by validated, calibrated, equitable performance where it is deployed. Table 3 summarizes the analytical gaps and the study designs that address them.

Table 4. Gaps between reported accuracy and clinical validity, and the designs that address them.

Gap	Why it inflates apparent performance	Corrective
No external validation	Internal accuracy overstates generalization.	External validation on representative data.
High heterogeneity	Pooled numbers describe no single setting.	Hierarchical/SROC modelling; subgroup analysis.
Weak comparators	Small, non-representative expert groups.	Adequate, prospective clinician comparison.
Poor reporting	Bias and non-reproducibility hidden.	AI-specific reporting guidelines.
Calibration ignored	Good AUC can mask miscalibration.	Report and correct calibration.
Population shift & equity	Models fail on under-represented groups.	Local validation; representative training data.

6. Discussion

Read analytically, the evidence on deep learning in medical imaging tells a coherent and, in its way, encouraging story, provided the numbers are interpreted correctly. On the best-designed comparisons available, deep-learning models achieve diagnostic discrimination roughly equivalent to expert clinicians across a range of imaging tasks, and in some well-defined, standardized problems such as retinal screening they perform at a very high level (Liu et al., 2019; Gulshan et al., 2016). This is a genuine scientific achievement, and it establishes that automated image interpretation of clinical quality is possible. What the same evidence does not establish, and what the enthusiastic reception of it often assumes, is that these models will deliver equivalent performance, or improve care, when deployed in real and diverse clinical settings. The gap between those two propositions is not a technicality; it is the difference between a promising result and a validated medical technology (Topol, 2019).

The analytical implications are clear. For researchers, the priority is to shift effort from producing yet another internally validated accuracy figure toward the external validation, prospective evaluation, calibration reporting, and, where feasible, randomized outcome studies that establish clinical validity, and to adhere to the AI-specific reporting standards that make studies trustworthy and comparable (Nagendran et al., 2020; Liu et al., 2020). For those deploying models, in pathology, radiology, and beyond, the priority is local validation on the population of use and vigilance against dataset shift and miscalibration (van der Laak, Litjens, & Ciompi, 2021). And for health systems, especially in resource-limited settings where automated diagnosis could extend scarce expertise furthest, the priority is to demand evidence of validity in context rather than to import benchmark accuracy on faith, so that a technology capable of reducing disparities does not instead entrench them. Realizing the promise the numbers hint at is, in the end, less a modelling problem than a problem of evidence and evaluation.

Table 5. An analytical checklist for appraising a deep-learning diagnostic-accuracy study.

Dimension	Question to ask	Red flag
Validation	Was performance tested on external, out-of-sample data?	Only internal test set reported.
Comparator	Were clinicians adequate in number and prospective?	Few experts; retrospective; unmatched.
Metrics	Are Se, Sp, and calibration reported, not AUC alone?	AUC only; no operating point or calibration.
Prevalence	Does the test set reflect real-world prevalence?	Artificially balanced case/control set.
Reporting	Does the study follow AI-specific guidelines?	No protocol; code/data unavailable.
Generalisability	Was dataset shift and subgroup performance examined?	Single site; no equity analysis.

It is worth situating diagnostic imaging within the broader trajectory of machine learning in medicine, because the analytical lessons generalize. Across clinical prediction, risk stratification, and decision support, the same pattern recurs: impressive performance on retrospective data, followed by the harder and slower work of prospective validation and integration, with the gap between the two determining whether a model helps or harms (Rajkomar, Dean, & Kohane, 2019; Yu, Beam, & Kohane, 2018). Imaging has led the field partly because its tasks are well posed and its data plentiful, but the evidentiary standards it is now being held to, external validation, prospective evaluation, calibration, and equity, apply to medical AI as a whole. Reading the imaging literature analytically is thus a rehearsal for reading the much larger body of clinical AI that is following it.

Looking forward, two developments will test the methodology assembled here. Foundation models trained on vast, heterogeneous data promise generalist medical AI that could ease the data and generalization problems that constrain today's task-specific models, but they also complicate evaluation, since a model that can perform many tasks resists the clean, single-task accuracy metrics on which this review has relied. Generative models raise further questions still, of factual reliability and of how to measure the quality of an output that is not a simple class label. In both cases the response should be the same in spirit: to insist that capability be demonstrated through rigorous, prospective, externally validated evaluation reported to explicit standards, rather than asserted from benchmark scores. The metrics and cautions developed for diagnostic classification are not made obsolete by these advances; they are the foundation on which the evaluation of more capable systems must be built.

7. Limitations

This review is analytical and interpretive rather than a fresh meta-analysis; it synthesizes reported pooled estimates from existing meta-analyses rather than re-extracting primary data, and the figures we present are drawn from those sources and inherit their limitations, including the very heterogeneity and reporting problems we discuss. Our selection of evidence, though centred on the most rigorous and highly cited syntheses, reflects judgement, and the field moves quickly enough that specific estimates will date. We concentrate on binary classification in a few imaging domains, so conclusions transfer less securely to segmentation, detection, multi-class, and non-imaging tasks, and to the emerging generation of foundation and generative models whose evaluation raises further methodological questions. These constraints qualify the specific numbers without, we believe, unsettling the central analytical conclusion regarding the gap between accuracy and validity.

8. Conclusion

Deep learning has demonstrated, across a substantial and quantitatively impressive body of work, that algorithms can interpret medical images with an accuracy comparable to that of expert clinicians, reaching pooled sensitivity and specificity near 87% and 92% on the best externally validated comparisons and AUCs above 0.90 in many domains. Read with the discipline of diagnostic-test methodology, however, these numbers carry an essential qualification: they are measures of accuracy on datasets, achieved amid high heterogeneity and often without external validation, and they are not yet measures of clinical validity in practice. The decisive task for the field is therefore analytical and evidentiary rather than purely computational, to close the gap between reported

ISSN: 3014-5197 - https://doi.org/10.54878/mcr68n77

accuracy and demonstrated clinical benefit through external validation, prospective evaluation, honest reporting, attention to calibration and equity, and, ultimately, outcome studies. Only when that evidentiary work is done will the remarkable numbers of the deep-learning literature translate into the improvements in diagnosis, and in access to diagnosis, that they promise.

References

Aggarwal, R., Sounderajah, V., Martin, G., Ting, D. S. W., Karthikesalingam, A., King, D., … Darzi, A. (2021). Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis. npj Digital Medicine, 4, 65. https://doi.org/10.1038/s41746-021-00438-z

Alshamsi, M., & Alhammadi, N. (2025). Next-generation AI in neurodevelopment: Multi-omics applications from diagnosis to care. International Journal of Rehabilitation & Disability Studies, 1(2), 13–24. https://doi.org/10.54878/yrp30g15

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, 115–118. https://doi.org/10.1038/nature21056

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., … Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1), 195. https://doi.org/10.1186/s12916-019-1426-2

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., … Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. The Lancet Digital Health, 1(6), e271–e297. https://doi.org/10.1016/S2589-7500(19)30123-2

Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., & Denniston, A. K. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Nature Medicine, 26(9), 1364–1374. https://doi.org/10.1038/s41591-020-1034-x

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., … Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577, 89–94. https://doi.org/10.1038/s41586-019-1799-6

Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., … Maruthappu, M. (2020). Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. BMJ, 368, m689. https://doi.org/10.1136/bmj.m689

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

Sulthan, N., et al. (2025). The role of AI in early diagnosis, management, and clinical research of primary immunodeficiency diseases: Insights from GCC immunologists. International Journal of Applied Technology in Medical Sciences, 4(1).

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

van der Laak, J., Litjens, G., & Ciompi, F. (2021). Deep learning in histopathology: The path to the clinic. Nature Medicine, 27(5), 775–784. https://doi.org/10.1038/s41591-021-01343-4

Xue, P., Wang, J., Qin, D., Yan, H., Qu, Y., Seery, S., … Qiao, Y. (2022). Deep learning in image-based breast and cervical cancer detection: A systematic review and meta-analysis. npj Digital Medicine, 5, 19. https://doi.org/10.1038/s41746-022-00559-z

Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. Nature Biomedical Engineering, 2(10), 719–731. https://doi.org/10.1038/s41551-018-0305-z

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Medicine, 15(11), e1002683. https://doi.org/10.1371/journal.pmed.1002683