What Radiology Teaches Dermatology About AI

Benchmarks, Bedside, and the Gap Between

Shira Lipton, MD
Board-Certified Dermatologist
March 2026

A clinical perspective on what a decade of radiology AI adoption reveals about the future of artificial intelligence in dermatology.

Executive Summary

Radiology has been the longest-running natural experiment in clinical AI adoption. The specialty combines digital images, clear diagnostic benchmarks, and repeatable pattern-recognition tasks. By conventional logic, it should have been the first medical field to see significant workforce displacement. Instead, after nearly a decade of AI models that match or exceed human performance on standardized tests, demand for radiologists in the United States is at an all-time high. Residency positions reached a record 1,208 in 2025, vacancy rates have never been greater, and average compensation has risen 48 percent since 2015.[1]

Dermatology occupies a strikingly similar position. It is a visual specialty with well-defined diagnostic categories, large labeled image datasets, and a growing portfolio of AI tools trained to classify skin lesions. The same predictions that were made about radiology a decade ago are now being made about dermatology. The evidence from radiology suggests that those predictions will follow a similar trajectory: strong benchmark performance, slow clinical integration, and persistent demand for the human clinician.

This paper examines what the radiology experience reveals about the likely path of AI in dermatology. It draws on findings from the ARISE State of Clinical AI Report 2026,[2] current evidence on dermatology AI tools, and the structural factors that determine whether benchmark capability translates into bedside utility. The goal is to offer dermatologists a grounded framework for evaluating AI claims and adoption decisions, informed by the only medical specialty where the experiment has actually been running long enough to produce results.

The Radiology Paradox

In 2016, Geoffrey Hinton declared that medical schools should stop training radiologists because AI would soon make them obsolete. The logic appeared sound. Radiology operates on digital images that can be fed directly into neural networks. Diagnostic questions are often binary or categorical. Performance can be measured against clear ground truth. By 2017, a model called CheXNet could detect pneumonia on chest X-rays with greater accuracy than a panel of board-certified radiologists. Since then, companies have released models that can detect hundreds of diseases across multiple scan types, and the FDA has cleared more than 700 radiology AI devices, accounting for roughly three-quarters of all medical AI authorizations.[3]

The prediction was wrong. Three structural factors explain why.

The Performance Gap

Models that perform well on curated benchmarks frequently degrade when deployed in real clinical environments. A radiology AI model tested on data from a single hospital can lose as much as 20 percentage points in accuracy when applied to images from a different institution.[1] Training datasets are assembled with strict inclusion criteria: unambiguous diagnoses, clean images, optimal angles. Real-world images are noisier, more ambiguous, and drawn from patient populations that differ from the training set. The ARISE report confirms this pattern across clinical AI more broadly, noting that model capability is accelerating while evidence of real clinical impact remains limited.[2]

The Regulatory and Liability Gap

The FDA maintains a sharp divide between assistive tools, which require physician oversight, and autonomous tools, which do not. The bar for autonomous clearance is substantially higher, requiring developers to demonstrate that the system can recognize and reject cases outside its competence. Malpractice insurers compound this constraint. Standard policy language now often excludes coverage for diagnoses generated autonomously by software. Without malpractice coverage, hospitals cannot permit algorithms to sign reports, regardless of technical performance. One insurer has adopted what it calls an “Absolute AI Exclusion” clause.[1]

The Workflow Gap

A 2012 study of staff radiologists found that only 36 percent of their time was dedicated to direct image interpretation. The majority was spent on activities that no current AI system can perform: consulting with referring clinicians, communicating results to patients, overseeing imaging examinations, teaching trainees, and reviewing imaging orders. Even a model that could read every scan in the department would displace only a fraction of the radiologist’s actual workload.[1]

The net effect of these three gaps is that AI in radiology has followed a pattern of augmentation rather than replacement. The technology has increased throughput and, in some cases, diagnostic accuracy, but demand for human radiologists has grown alongside it. When digital imaging made scans faster to read in the early 2000s, American imaging utilization rose 60 percent rather than producing layoffs. AI appears to be following the same trajectory.

Dermatology: The Next Test Case

To AI developers and investors, dermatology looks like radiology’s natural successor. The parallels are immediately apparent. Both specialties rely heavily on visual pattern recognition. Both have well-established classification systems. Both generate large volumes of digital images that can be used to train deep learning models. And both have produced AI systems that match or exceed dermatologist-level performance on curated image sets.

In melanoma detection, for example, systematic reviews report pooled sensitivity of 86 percent for AI classifiers compared with 65 percent for generalist clinicians. Convolutional neural networks trained on dermoscopic images have achieved sensitivity above 90 percent for several skin cancer subtypes. The FDA granted De Novo authorization to DermaSensor in January 2024, making it the first AI-enabled device for skin cancer detection in primary care.[6] In Europe, Skin Analytics received a Class III CE Mark for DERM, the first autonomous AI system approved for clinical decisions in dermatology without physician oversight.[7]

If the radiology experience were simply a matter of technical maturity, dermatology might be on a faster track. The technology exists, clearances are beginning to accumulate, and the clinical need for diagnostic support in primary care settings is well-documented.

The radiology experience, however, demonstrates that technical maturity is necessary but not sufficient. Dermatology faces the same three structural gaps that have slowed radiology AI adoption. It also faces additional challenges that are specific to the specialty.

Dermatology's Additional Complications

The Skin Tone Bias Problem

The most extensively documented challenge specific to dermatology AI is performance disparity across skin tones. Daneshjou and colleagues demonstrated in 2022 that state-of-the-art AI models showed a 27 to 36 percent reduction in ROC-AUC when tested on a diverse clinical image set compared with their originally reported performance.[4] All models performed substantially worse on dark skin tones and uncommon diseases. The underlying problem is structural: the datasets used to train these models underrepresent darker skin tones, and the dermatologists who generated the training labels themselves perform worse on images of dark skin, embedding human bias into the algorithmic baseline.

Groh and colleagues, in a 2024 study published in Nature Medicine, extended this finding into the clinical workflow.[5] In a study of 389 dermatologists and 459 primary care physicians across 39 countries, AI decision support improved accuracy for both groups. However, for primary care physicians, AI assistance exacerbated skin-tone-based disparities by five percentage points. Accuracy improved more on light skin than on dark skin. The mechanism was revealing: specialists used AI contextually, integrating it with their own clinical judgment, while primary care physicians tended to follow AI suggestions more uncritically, amplifying the model's existing biases.

This finding carries direct implications for how dermatology AI will be deployed. The tools are most likely to be adopted first in primary care and triage settings, where clinicians have less dermatologic training and are therefore more susceptible to automation bias. The population-level consequence of deploying a biased tool through biased adoption patterns could worsen existing disparities in skin cancer detection among patients with darker skin tones.

Context Beyond the Image

Dermatologic diagnosis frequently depends on information that no photograph can capture: the evolution of a lesion over time, associated symptoms such as pruritus or pain, patient history of sun exposure and immunosuppression, family history of melanoma, and the tactile qualities of a lesion on palpation. A clinician evaluating a suspicious pigmented lesion integrates visual pattern recognition with clinical context, risk stratification, and a differential diagnosis that may include entities not representable in a single image.

Current AI systems in dermatology are predominantly trained on single images with diagnostic labels. They lack access to the longitudinal, multimodal data that clinicians use. This is not merely a technical limitation to be solved by larger datasets. It represents a fundamental mismatch between what the models are trained to do and what clinical dermatology requires.

The Patient-Facing Dimension

Dermatology differs from radiology in a way that has significant implications for AI integration: dermatologists interact directly with patients in a manner that radiologists typically do not. The clinical encounter in dermatology involves visual examination, patient education, shared decision-making about biopsies and treatment plans, and longitudinal management of chronic conditions such as psoriasis, eczema, and acne. These patient-facing responsibilities cannot be delegated to an algorithm.

Patient attitudes reinforce this constraint. Survey data indicate that 73.8 percent of patients trust dermatologist-guided AI, while only 1.5 percent trust standalone AI applications for dermatologic assessment. The clinical encounter is not simply a site where diagnoses are delivered; it is where trust is established, context is gathered, and therapeutic relationships are maintained.

What the Broader Evidence Shows

The ARISE State of Clinical AI Report 2026 provides a comprehensive view of the current landscape that extends well beyond dermatology. Several of its findings bear directly on how dermatologists should evaluate AI tools.[2]

Benchmarks do not predict clinical safety. The NOHARM benchmark evaluated 31 large language models on 100 authentic specialist consult cases and found that even top-performing models produced potentially severely harmful advice in 10 to 22 percent of cases. Medical knowledge scores on standard benchmarks correlated only moderately with clinical safety performance.[2] For dermatology AI, this means that a model's published accuracy on a curated image set is an incomplete measure of its fitness for clinical use.

Models are overconfident and fragile under uncertainty. Frontier AI models show strong performance on structured tasks but break down when facing uncertainty, missing information, or changed context. When researchers modified standard medical questions so that "none of the other answers" became the correct response, model accuracy dropped by 9 to 38 percent.[8] In dermatology, where clinical presentations are frequently ambiguous and where the correct answer is sometimes "monitor and re-evaluate," overconfident outputs represent a specific clinical risk.

Human-AI collaboration has not yet been optimized. A meta-analysis of 52 studies found that human-AI teaming improved diagnostic reliability compared with clinicians working alone, but full complementarity was rarely achieved.[9] In some configurations, AI assistance actually reduced human performance compared to the model working alone, suggesting that the interface between clinician and tool matters as much as the model's accuracy. The Groh skin-tone study illustrates this precise dynamic in dermatology: how clinicians use the tool determined whether it helped or harmed.

Automation bias and deskilling are documented risks. In a randomized controlled trial, AI-trained physicians exposed to deliberately flawed AI recommendations saw their diagnostic accuracy drop from 85 percent to 73 percent compared with those receiving error-free advice.[10] Separately, endoscopists who regularly used AI for polyp detection showed a 6 percent decline in adenoma detection rates when they later performed colonoscopies without AI, suggesting that continuous AI exposure may erode independent clinical vigilance.[11] These findings are directly relevant to dermatology, where pattern recognition is a core clinical skill that depends on ongoing practice.

The strongest evidence of real-world benefit comes from narrowly scoped, well-integrated tools. The most compelling clinical AI results reported in the ARISE review involved tools designed for specific tasks within existing workflows. In radiology itself, a prospective study across 12 sites and 463,094 women found that AI-supported mammography reading increased breast cancer detection rates by 17.6 percent without a significant increase in recall rates.[12] A randomized trial of 70 physicians found that a collaborative AI system improved clinician diagnostic accuracy by 10 percentage points when used as either a first or second opinion.[13] In stroke imaging, the NHS's deployment of AI across 26 hospitals was associated with a doubling of endovascular thrombectomy rates at AI-enabled sites.[14] In each case, the tool succeeded because it addressed a defined clinical problem, operated within an existing workflow, and preserved clinician oversight.

Implications for Dermatologists

The radiology experience and the broader clinical AI evidence base point toward several practical conclusions for dermatologists evaluating AI tools and the claims made about them.

Benchmark performance is necessary but not sufficient. A model that achieves high sensitivity on a curated test set has demonstrated technical capability. It has not demonstrated clinical utility. Before adopting any AI tool, dermatologists should ask what populations it was validated on, whether it has been tested prospectively in real clinical settings, and whether its performance has been evaluated on diverse skin tones. A tool that has been validated only on datasets skewed toward lighter skin tones and unambiguous diagnoses has not been validated for the full scope of dermatologic practice.

Workflow integration determines value more than model accuracy. The most successful clinical AI deployments are those that fit into existing workflows rather than requiring clinicians to adopt new processes. For dermatology, this suggests that AI tools embedded within dermoscopy platforms, teledermatology triage systems, or electronic health record workflows are more likely to deliver value than standalone applications that exist outside the clinical encounter.

Clinician training on AI limitations is as important as training on AI use. The automation bias findings from the ARISE report indicate that clinicians who understand when and how AI fails are better positioned to use it effectively. Dermatologists and primary care physicians who use AI diagnostic tools should be trained not only on how to interpret the tool's outputs but also on its known failure modes, its performance limitations on specific skin tones, and the types of clinical presentations where it is most likely to produce inaccurate or overconfident results.

Regulatory clearance is not clinical endorsement. The majority of FDA-cleared AI medical devices, including radiology tools, have received authorization through the 510(k) pathway, which requires only demonstrated equivalence to an existing device rather than prospective evidence of clinical benefit.[2] For dermatology, this means that an FDA-cleared AI tool has met a regulatory threshold for market entry, not a clinical threshold for patient benefit. Dermatologists should evaluate the published evidence behind any tool independently of its regulatory status.

The near-term value of AI in dermatology lies in augmentation, not autonomy. The evidence consistently shows that AI tools add the most value when they operate within a clinician-supervised workflow. In dermatology, this means tools that help triage teledermatology referrals, flag suspicious lesions for closer examination, or assist less experienced clinicians in primary care settings. Fully autonomous diagnostic AI in dermatology remains premature. The workflow, liability, and bias challenges are unresolved, and the patient-facing nature of the specialty makes clinician involvement not merely a safety requirement but a clinical one.

Conclusion

Radiology has provided the medical profession with a decade of evidence about what happens when a visual, pattern-recognition-intensive specialty encounters AI. The answer is not obsolescence. Models have improved, clearances have accumulated, and adoption has gradually increased. But radiologists are busier and better-compensated than they have ever been, because the factors that determine whether AI translates from benchmark to bedside extend far beyond the model's accuracy on a test set.

Dermatology is now entering the same phase. The tools are improving. The datasets are growing. The investment is flowing. And the same predictions that were wrong about radiology are being made about dermatology. The radiology experience does not suggest that dermatologists should ignore AI or resist its integration. It suggests the opposite: that dermatologists are precisely the professionals who should be leading the evaluation, implementation, and governance of these tools within their specialty. The clinicians who understand both the clinical context and the technology's limitations are the ones best positioned to ensure that AI serves patients rather than vendor narratives.

The models will continue to improve. The question that matters is not whether AI can match a dermatologist on a benchmark. The question is whether it improves outcomes for the patient sitting in the examination room. That question can only be answered by clinicians willing to engage with the technology critically, demand rigorous evidence, and insist that adoption decisions are driven by clinical need rather than technological enthusiasm.

Disclosures

The author reports no conflicts of interest related to the content of this paper. No funding was received for this work.

References

1. Mousa D. AI isn't replacing radiologists. Works in Progress. September 25, 2025.
2. Brodeur PG, Goh E, Tat E, et al. State of Clinical AI 2026. ARISE Network. January 2026.
3. US Food and Drug Administration. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. Updated 2025.
4. Daneshjou R, Vodrahalli K, Novoa RA, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. Sci Adv. 2022;8(32):eabq6147.
5. Groh M, Harris C, Soenksen L, et al. Deep learning-aided decision support for diagnosis of skin disease across skin tones. Nat Med. 2024;30:573-583.
6. DermaSensor. FDA De Novo Authorization. January 17, 2024. DEN230008.
7. Skin Analytics. DERM receives Class III CE Mark under European Medical Device Regulation (MDR 2017/745). 2024.
8. Bedi S, Shah N, et al. None of the other answers: An LLM weakness. JAMA Network Open. August 2025.
9. Liu C, Chen X, et al. Human-machine collaboration is not yet optimized: A meta-analysis. NPJ AI. December 2025.
10. Qazi A, Alizai A, et al. AI-trained physicians exhibit automation bias. medRxiv. September 2025.
11. Budzyn M, Mori Y, et al. Are there concerns for deskilling? Lancet Gastroenterol Hepatol. October 2025.
12. Eisemann N, Katalinic A, et al. AI improves breast cancer detection without more false alarms. Nat Med. January 2025.
13. Everett H, Horvitz E, et al. Collaborative AI improves clinician diagnostic accuracy by 10%. medRxiv. June 2025.
14. Nagaratnam K, Harston G, et al. Nationwide AI for stroke imaging doubles thrombectomy rates in England. Lancet. December 2025.