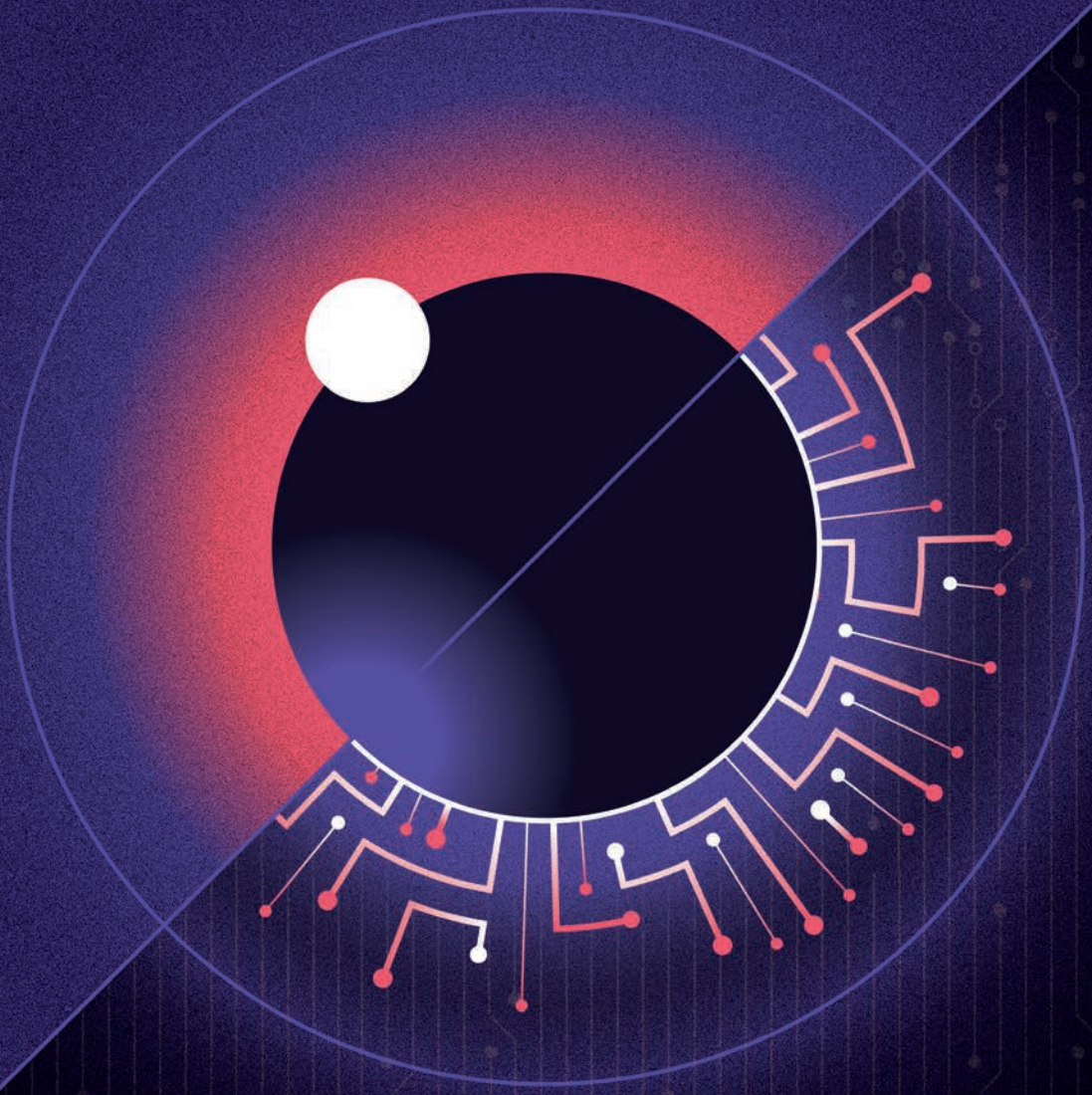
AI: From Hype to Real-World Clinical Value



Examining ophthalmic AI through the lens of clinical value rather than technical novelty

By Kai Jin and Andrzej Grzybowski

Artificial intelligence (AI) has attracted exceptional interest in ophthalmology because the specialty sits at the intersection of high-volume imaging, pattern recognition, and unmet screening demand. The landmark deep-learning study by Gulshan and colleagues (1) showed that retinal fundus photographs could be classified for diabetic retinopathy with expert-level performance, helping define ophthalmology as an early proving ground for medical AI. Soon after, deep learning systems for OCT-based referral decisions suggested that AI could support more complex triage pathways beyond single-disease screening (2). These studies fueled a decade of optimism that ophthalmology would become one of the first specialties to realize broad clinical gains from AI.

The optimism was not misplaced, but it was incomplete. Much of the early literature focused on retrospective discrimination metrics, especially the area under the receiver operating characteristic curve, sensitivity, and specificity. These metrics matter, but they do not answer the questions that determine real-world clinical value. Does the system expand access to care? Does it reduce clinician workload without increasing downstream errors? Does it improve triage speed, treatment timeliness, or health equity? Does it work across cameras, institutions, disease prevalence settings, and patient populations different from the development set? A model that performs well on a curated test set may still fail clinically if it is ungradable too often, is difficult to integrate into workflow, produces outputs clinicians do not trust, or increases false-positive referrals that overwhelm specialist capacity.

This distinction between technical promise and clinical usefulness is now central to the field. In ophthalmology, AI is no longer a purely experimental idea. Autonomous diabetic retinopathy screening systems have undergone prospective evaluation, achieved regulatory authorization, and entered real clinical workflows (3-7). At the same time, utilization remains modest, and many promising ophthalmic AI tools remain far from routine deployment (7). The gap between capability and adoption is where hype accumulates.

This review examines ophthalmic AI through the lens of real-world clinical value rather than technical novelty. The central argument is straightforward: the field should stop asking only whether AI can classify images accurately and start asking whether it improves the delivery of eye care in measurable, durable, and equitable ways.

WHERE OPHTHALMIC AI ALREADY DELIVERS VALUE

Diabetic retinopathy screening is the clearest success case

If one application has moved furthest from hype to practice, it is diabetic retinopathy (DR) screening. The reasons are structural as much as technical. DR is common, early treatment matters, image acquisition is standardized, and screening often happens outside specialist clinics where workforce shortages are greatest. These characteristics create a clinically meaningful use case for automation.

The early development literature established technical feasibility. Gulshan et al. showed strong diagnostic performance for referable DR on fundus photographs (1). Subsequent community and smartphone-based work demonstrated that AI-enabled DR screening could operate beyond tertiary eye hospitals, including settings with constrained infrastructure (4). The decisive shift, however, came from prospective and autonomous evaluation. Abramoff et al. reported a pivotal trial of an autonomous AI diagnostic system in primary care offices (3), showing that ophthalmic AI could function as a regulated clinical device rather than only as a research model. Ipp et al. later showed high sensitivity and specificity for autonomous detection of both more-than-mild DR and vision-threatening DR in a prospective multicenter study, with high imageability after a dilate-if-needed protocol (5).

Importantly, the translation story does not stop at accuracy. In a cluster-randomized trial, autonomous AI deployment increased specialist clinic productivity, providing rare interventional evidence that AI can improve operational performance in routine care (6). This is much closer to real-world value than retrospective benchmark gains. In addition, recent utilization data from the US show that AI-based DR detection is no longer hypothetical, even if adoption remains limited. Since 2021, claims-based use has increased, but uptake is still low relative to the size of the eligible diabetic population (7). That finding is sobering: regulatory authorization and positive trials do not automatically create scale.

Taken together, DR screening shows what meaningful ophthalmic AI translation looks like: a clinically important problem, prospective validation, regulatory pathway, workflow fit, and early operational benefit. It also shows that implementation science, reimbursement, and adoption are now the limiting factors.

OCT triage and retinal referral are highly promising but still more site-dependent

AI for OCT interpretation has also produced influential results. De Fauw et al. developed a deep-learning system for diagnosis and referral in retinal disease that used OCT input and generated clinically actionable referral recommendations (2). Conceptually, this work mattered because it moved ophthalmic AI beyond binary screening toward richer, multi-condition triage. In principle, such systems could help prioritize retina referrals, support virtual clinics, and reduce bottlenecks in high-volume services.

However, OCT-based AI has been harder to generalize than DR screening in primary care. OCT devices vary by vendor,

Table 1. Representative ophthalmic AI use cases and current translational maturity

Use case	*Main input*	*Strongest current evidence*	*Current translational maturity*	*Main barrier to wider adoption*
Autonomous diabetic retinopathy screening	Fundus photography	Prospective multicenter studies, regulatory clearance, early real-world deployment (3,5-7)	High	Reimbursement, implementation scale, referral integration
OCT-based retinal triage	OCT volumes	Strong retrospective and referral studies (2)	Moderate	Device heterogeneity, workflow integration, prospective multicenter implementation
Rural/community retinal screening	Ultra-widefield or smartphone fundus imaging	Community and rural screening studies (4,8)	Moderate	Infrastructure, image quality assurance, local referral pathways
Myopic maculopathy classification	Fundus photography	Public challenge and multicenter evaluation (9)	Moderate-early	Management-linked prospective validation
Retinopathy of prematurity support	Wide-field neonatal fundus imaging	Early deep-learning classification studies (10)	Early	Safety thresholds, expert oversight, medicolegal risk
LLM-supported documentation/education	Text or multimodal clinical data	Early observational and commentary-level evidence (12)	Early	Hallucination, governance, privacy, supervision

acquisition protocols differ, annotation is labor-intensive, and referral decisions may depend on local pathways rather than on image features alone. As a result, many OCT models remain strong demonstrations of capability without yet producing the same depth of prospective multicenter implementation evidence seen in autonomous DR screening. Their value is likely to emerge first in tightly integrated health systems or teleophthalmology networks where acquisition quality, referral thresholds, and specialist escalation pathways are already standardized.

AI may improve access in community and underserved settings

A second domain of tangible value is outreach and community screening. Cui et al. showed strong deep-learning performance for ultra-widefield fundus imaging in rural screening settings, reinforcing the idea that ophthalmic AI can help extend specialist-grade triage to regions with limited ophthalmologist availability (8). Similar logic underpins smartphone-based and offline DR systems, which may be especially relevant in low-resource settings where network dependence and specialist grading are major constraints (4).

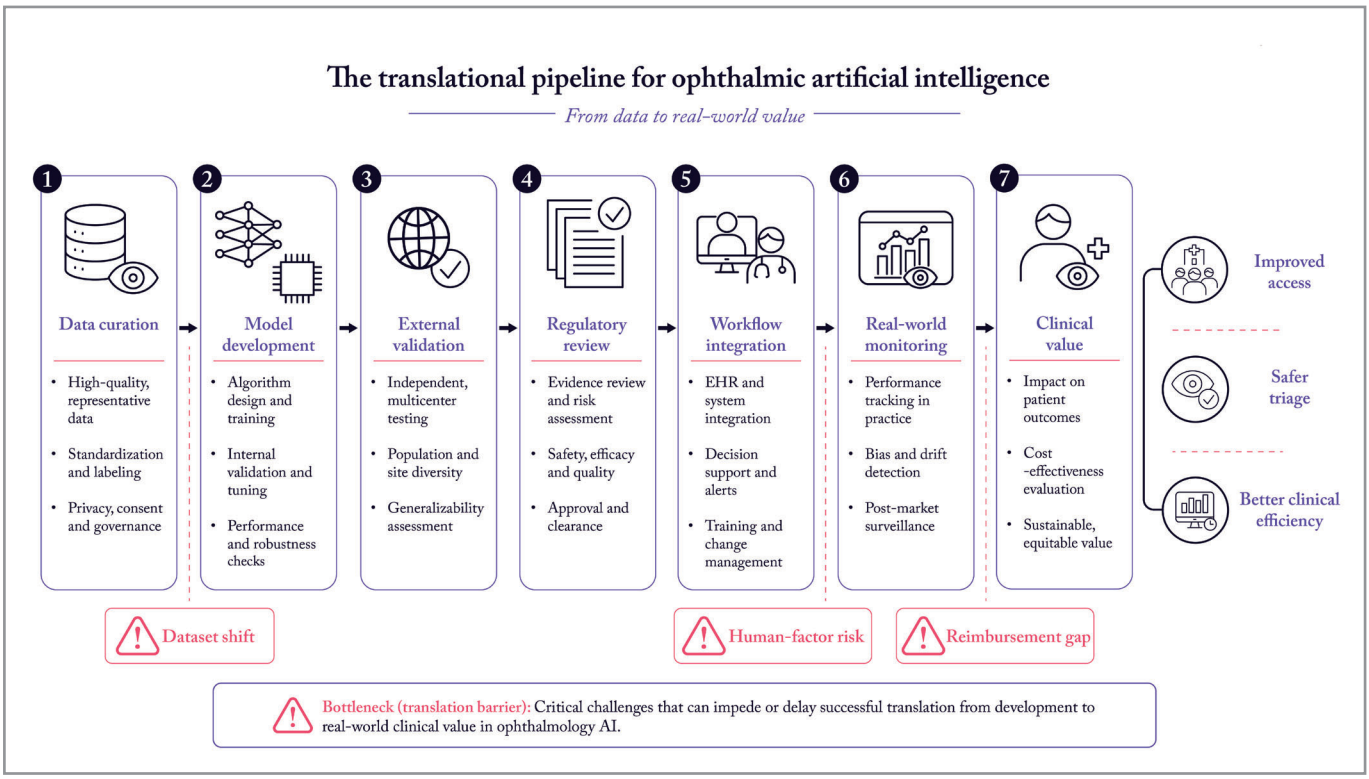
The key point is that AI value is contextual. A modestly accurate model that functions reliably in a rural or primary care pathway may be more clinically valuable than a technically superior model that requires expensive hardware, specialist oversight, and ideal image capture. Ophthalmology should therefore judge AI not only by discrimination metrics but also by whether it lowers the threshold for reaching currently underserved populations.

WHERE THE HYPE STILL EXCEEDS THE EVIDENCE

Many models are still retrospective, narrow, and fragile

Despite the growth of the literature, most ophthalmic AI studies remain retrospective and development-focused. Many are trained and tested on data captured under similar conditions, sometimes from the same institutions or device ecosystems. This raises familiar concerns: spectrum bias, label leakage, limited external validity, and performance erosion under distribution shift. Even when external

Figure 1. From algorithm performance to real-world clinical value in ophthalmology



Conceptual translation pathway showing how ophthalmic AI moves from dataset development to bedside value. The figure highlights where many models stall: external validity, workflow mismatch, and lack of prospective evidence.

test sets are used, the externality may be weaker than it appears if image acquisition standards, disease prevalence, or referral definitions remain closely aligned with the development environment.

This is one reason clinical adoption lags behind publication volume. An ophthalmologist does not need another high-performing model in principle; they need a system that works when images are imperfect, patients are complex, and clinic schedules are full. This is also why recent reporting guidance matters. TRIPOD+AI, CONSORT-AI, and SPIRIT-AI all emphasize transparent description of data provenance, model behavior, human-AI interaction, and prospective evaluation (13-15). These frameworks are not bureaucratic add-ons; they are essential if ophthalmic AI is to move from research claims to trustworthy clinical tools.

Human factors remain underappreciated

A second source of hype is the assumption that better model performance automatically leads to better human performance. That assumption is unsafe. In practice, clinicians and allied eye-care professionals do not respond to AI outputs mechanically. They weigh confidence displays, visual explanations, time pressure, prior beliefs, and workflow demands. Carmichael et al. showed that ambiguous deep-learning outputs can influence diagnostic decisions in nontrivial ways, highlighting the risk that poorly designed AI interfaces may confuse rather than assist users (11).

This matters because many ophthalmic AI systems are likely to function as decision support, not as fully autonomous devices. If output formats are vague, overconfident, or poorly calibrated, they may increase cognitive burden or induce automation bias. In other words, the clinical unit of analysis is often not the algorithm alone but the human-AI team. Ophthalmology needs more prospective studies that evaluate that team directly.

Large language models are useful, but their clinical role is narrower than the hype suggests

The newest wave of enthusiasm centers on large language models (LLMs). In ophthalmology, LLMs can already support drafting clinic communications, summarizing records, generating patient education materials, and organizing literature. These are plausible near-term use cases because they target language-intensive but lower-risk tasks. By contrast, unsupervised diagnostic reasoning from free text or multimodal input remains much less mature.

Young and Zhao argued that LLMs have reached the shoreline of ophthalmology, but not yet the interior of routine clinical decision-making (12). That framing is useful. For now, LLMs

should be treated as assistive tools whose outputs require oversight, especially where hallucination, omitted nuance, or fabricated citations could mislead clinicians and patients. Their real-world value is likely to emerge incrementally through workflow support rather than abrupt replacement of specialist judgment.

A PRACTICAL FRAMEWORK FOR REAL-WORLD CLINICAL VALUE

To move beyond hype, ophthalmic AI should be judged against five translational questions:

1. Is the clinical problem important and poorly served by the current pathway?

AI adds the most value where disease burden is high, delay matters, and specialist access is constrained. DR screening is the clearest example. By contrast, deploying AI into already efficient, specialist-rich workflows may yield little incremental benefit unless it meaningfully reduces workload or improves triage accuracy.

2. Does the model solve the right task?

Many ophthalmic AI models predict image labels; fewer solve operational decisions. Yet clinicians often need referral prioritization, interval prediction, treatment urgency classification, or longitudinal risk stratification rather than disease/no-disease outputs alone. Qian et al. showed impressive diagnostic performance for AI algorithms in myopic maculopathy classification, but the path to clinical value will depend on whether those outputs alter management decisions in real settings (9). The same principle applies to retinopathy of prematurity, where automated classification may be valuable only if it improves access to timely expert review without creating unsafe false reassurance (10).

3. Has the system been tested prospectively in workflow?

The strongest ophthalmic AI evidence comes from studies that evaluate the system where it will actually be used. Prospective design, multicenter sampling, pragmatic endpoints, and workflow-level outcomes should become standard. Productivity gains, referral completion, time to treatment, imageability, false referral burden, and missed disease rates are often more informative than small changes in area under the curve.

4. Does the system work equitably across settings and populations?

Ophthalmology should resist adopting systems that perform well only in narrow subgroups or well-resourced centers. Device variation, pigmentation differences, coexisting ocular disease, and differences in image quality may all affect performance. Equity is not a downstream issue to be checked after deployment; it is part of whether a tool has clinical value at all.

5. Is there a plan for governance after deployment?

Clinical value is not fixed at launch. Models drift, workflows change, cameras are replaced, and prevalence shifts. The FDA's AI-enabled device framework and recent transparency principles signal the direction of travel: lifecycle oversight, user-facing transparency, and clearer communication about intended use and limitations (16, 17). Ophthalmology programs implementing AI need post-deployment monitoring, escalation pathways for failure cases, and mechanisms for periodic recalibration or retraining.

WHAT THE NEXT PHASE SHOULD LOOK LIKE

The next phase of ophthalmic AI should be less about proving that deep learning can recognize eye disease and more about proving where it changes care for the better. Three priorities stand out.

First, prospective comparative studies should measure service outcomes, not only algorithmic outcomes. For example, does AI reduce missed follow-up, shorten treatment delays, increase screening completion, or allow retina clinics to absorb more urgent referrals without sacrificing quality? The cluster-randomized evidence from autonomous DR screening provides a model for this type of study (6).

Second, reporting quality must improve. Ophthalmology should expect AI studies to align with TRIPOD+AI for prediction modeling and CONSORT-AI/SPIRIT-AI for interventional evaluation (13-15). Poorly described datasets, opaque exclusion criteria, or vague descriptions of human oversight should no longer be acceptable in a field that seeks bedside impact.

Third, the field should broaden its concept of value. Not every useful AI tool will be autonomous or diagnostic. Some of the most scalable gains may come from image quality control, worklist prioritization, documentation assistance, patient messaging, and clinical trial enrichment. These uses may not be as glamorous as "doctor-level diagnosis," but they may create more reliable value in everyday eye care.

CONCLUSION

Artificial intelligence in ophthalmology has progressed beyond pure hype, but only in selected domains. The field's most convincing success to date is autonomous diabetic retinopathy screening, where prospective validation, regulatory authorization,

Table 2. A practical checklist for judging whether an ophthalmic AI tool has real-world clinical value

Domain	*Question*	*Minimum evidence expected before broad clinical use*
Clinical need	Does the tool address a meaningful bottleneck or unmet need?	Clear target population and workflow problem
Technical validity	Does performance hold across sites, devices, and populations?	External validation across heterogeneous settings
Workflow fit	Can the tool be used without disrupting care delivery?	Usability and implementation testing in routine practice
Human factors	Do users interpret outputs correctly and safely?	Prospective human-AI interaction studies
Outcome relevance	Does it improve access, efficiency, referral quality, or patient outcomes?	Pragmatic or prospective comparative evidence
Governance	Is there post-deployment monitoring and transparency?	Defined monitoring, escalation, and update plan
Equity	Does the system perform fairly across subgroups?	Prespecified subgroup and imageability analyses

and early workflow benefits now support genuine clinical value. Outside that area, many ophthalmic AI systems remain promising yet insufficiently proven. Their limitations are no longer mainly computational; they are translational. Weak external validation, incomplete reporting, workflow mismatch, human-factor risks, uncertain reimbursement, and insufficient post-deployment governance continue to prevent many strong models from becoming strong clinical tools.

The practical standard for the next decade should be simple. Ophthalmic AI deserves adoption when it improves access, accuracy, efficiency, equity, or outcomes in routine care more than existing alternatives do, and when those gains are demonstrated prospectively rather than assumed from retrospective performance. If the field keeps that standard in view, it can move from excitement about what AI might do to evidence about where it already matters.

References

1. *V Gulshan et al., "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," JAMA, 316 (2016). PMID: 27898976.*
2. *J De Fauw et al., "Clinically applicable deep learning for diagnosis and referral in retinal disease," Nat Med., 24, 1342 (2018). PMID: 30104768.*
3. *MD Abramoff et al., "Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices," npj Digit Med, 1, 39 (2018). PMID: 31304320.*
4. *S Natarajan et al., "Diagnostic accuracy of community-based diabetic retinopathy screening with an offline artificial intelligence system on a smartphone," JAMA Ophthalmol, 137, 1182 (2019). PMID: 31429903.*
5. *E Ipp et al., "Pivotal evaluation of an artificial intelligence system for autonomous detection of referrable and vision-threatening diabetic retinopathy," JAMA Netw Open, 4, e2134254 (2021). PMID: 34787626.*
6. *MD Abramoff et al., "Autonomous artificial intelligence increases real-world specialist clinic productivity in a cluster-randomized trial," npj Digit Med, 6, 184 (2023). PMID: 37821691.*
7. *SA Shah et al., "Use of artificial intelligence-based detection of diabetic retinopathy in the US," JAMA Ophthalmol, 142, 1171 (2024). PMID: 39475749.*
8. *T Cui et al., "Deep learning performance of ultra-widefield fundus imaging for screening retinal lesions in rural locales," JAMA Ophthalmol, 141, 1045 (2023). PMID: 37728940.*
9. *B Qian et al., "A competition for the diagnosis of myopic maculopathy by artificial intelligence algorithms," JAMA Ophthalmol, 142, 1006 (2024). PMID: 39259384.*
10. *EK Yenice et al., "Automated detection of type 1 ROP, type 2 ROP and A-ROP based on deep learning," Eye (Lond), 38, 2644 (2024). PMID: 38664166.*
11. *J Carmichael et al., "Diagnostic decisions of specialist optometrists exposed to ambiguous deep-learning outputs," Sci Rep, 14, 6775 (2024). PMID: 38531716.*
12. *BK Young, PY Zhao, "Large language models and the shoreline of ophthalmology," JAMA Ophthalmol, 142, 375 (2024). PMID: 38353968.*
13. *GS Collins et al., "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, 385, e078378 (2024). PMID: 38631719.*
14. *X Liu et al., "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension," Nat Med, 26, 1364 (2020). PMID: 32855431.*
15. *S Cruz Rivera et al., "Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension," Nat Med, 26, 1351 (2020). PMID: 32855427.*
16. *US Food and Drug Administration, "Artificial intelligence-enabled medical devices," FDA website (2024). Accessed May 8, 2026. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices*
17. *US Food and Drug Administration, "CDRH issues guiding principles for transparency of machine learning-enabled medical devices," FDA website (2024). Published June 13, 2024. Accessed May 8, 2026. https://www.fda.gov/medical-devices/medical-devices-news-and-events/cdrh-issues-guiding-principles-transparency-machine-learning-enabled-medical-devices*
18. *HE Moss, "Deep learning to improve diagnosis must also not do harm," JAMA Ophthalmol, 142, 1079 (2024). PMID: 39301272.*