The Journal of High School Research

Real-World Diagnostic Accuracy of Fully Autonomous, U.S. Food and Drug Administration-Approved Artificial Intelligence Systems for Diabetic Retinopathy Screening in the United States: A Systematic Review and Meta-analysis

Shreya Parimoo*

Submitted: 03 August 2026 **Accepted:** 10 August 2026 **Publication date:** 21 August 2026

DOI: 10.70671/4js25h02

Abstract: Diabetic retinopathy (DR) is a leading cause of preventable vision loss, and fully autonomous, U.S. Food and Drug Administration (FDA)-cleared artificial intelligence (AI) systems—LumineticsCore (formerly IDx-DR) and EyeArt—have been developed to expand DR screening access. While these systems performed well in their pivotal regulatory trials, their diagnostic accuracy when deployed in real-world U.S. clinical practice, outside controlled trial conditions, is less well characterized. This systematic review and meta-analysis searched PubMed/MEDLINE, the Cochrane Library, and IEEE Xplore (the Embase database was inaccessible) for U.S.-based, real-world evaluations of FDA-cleared LumineticsCore/IDx-DR or EyeArt that reported diagnostic accuracy against a human-grader reference standard, deliberately excluding pivotal trials and any evaluations conducted before FDA clearance. Only four eligible studies, together comprising approximately 1,429 screening encounters or eyes, were identified. This is a small evidence base, and the resulting pooled estimates should be interpreted as imprecise and hypothesis-generating rather than definitive. Pooled sensitivity for detecting referable DR was high, while pooled specificity was lower and considerably more variable across studies, with very high statistical heterogeneity that likely reflects differences among study sites in fundus camera systems, patient case mix, and image-acquisition protocols (for example, dilated versus non-mydriatic photography), as well as differences in reference-standard rigor. Certainty of the evidence, assessed using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework, was rated Low for both sensitivity and specificity. Because every included study was conducted within the U.S. regulatory and healthcare context, these findings are specific to U.S. deployment and should not be extrapolated to other countries or healthcare systems, which may use different devices, regulatory thresholds, populations, or care pathways. Site-level validation before and during clinical deployment remains essential.

Author keywords: Diabetic retinopathy; artificial intelligence; diagnostic accuracy; real-world evidence; autonomous screening

Introduction

Diabetic retinopathy (DR) remains a leading cause of preventable vision loss among working-age adults with diabetes. Fully autonomous artificial intelligence (AI) systems that interpret retinal photographs without clinician oversight have been developed to expand screening access beyond the limited supply of eye-care specialists. Two systems hold U.S. Food and Drug Administration (FDA) clearance for fully autonomous DR diagnosis: LumineticsCore (formerly IDx-DR; Digital Diagnostics; De Novo clearance April 2018) and EyeArt (Eyenuk; 510(k) clearance August 2020). In their pivotal regulatory trials, both systems demonstrated strong diagnostic accuracy against rigorous reference standards.[1,2]

Pivotal trials, however, used standardized imaging protocols and enriched populations under controlled research conditions. Real-world deployments, by contrast, often use non-mydriatic (undilated) cameras in primary-care settings, encounter a broader case mix of patients, and may operate under different referral thresholds. Whether the high diagnostic accuracy demonstrated in pivotal trials is preserved once these systems are deployed in routine U.S. clinical practice is an open and clinically important question: over- or under-performance relative to pivotal-trial figures has direct implications for patient safety, referral burden, and the credibility of autonomous AI diagnostics more broadly.

This review therefore asks: among U.S. studies evaluating FDA-cleared LumineticsCore/IDx-DR or EyeArt in real-world, non-pivotal-trial clinical deployment, what is the

*Corresponding Author: Shreya Parimoo.
Email: shreyaparimoo@gmail.com
Stockdale High School, Bakersfield, California, United States

This review was not registered in the International Prospective Register of Systematic Reviews (PROSPERO). It was conducted as an independent student research project

pooled sensitivity and specificity for referable (more-than-mild) diabetic retinopathy, and how much confidence can be placed in that estimate given the size and consistency of the available evidence? Establishing this real-world evidence base is expected to help clinicians, health systems, and regulators anticipate how these systems perform outside the enriched conditions of regulatory trials and to identify where additional site-level validation or evidence generation is most needed.

Description/Methodology/Proposed Research Approach

Protocol and registration

This review was not pre-registered with the International Prospective Register of Systematic Reviews (PROSPERO), a public database in which systematic review protocols are recorded before conduct to reduce reporting bias. It was conducted as a single-reviewer student research project. Dual independent screening was not performed. These limitations are disclosed and reflected in the downgraded GRADE certainty ratings (see Discussion).

Eligibility criteria

Studies were eligible if they: (1) evaluated LumineticsCore/IDx-DR or EyeArt in a version that was FDA-cleared at the time of the study, or in post-clearance commercial deployment (pre-clearance predecessor versions of subsequently cleared products were excluded); (2) were conducted in a real-world U.S. clinical deployment and were not a pivotal or regulatory registration trial; (3) reported sensitivity and/or specificity against a human-grader reference standard; and (4) were peer-reviewed primary research with extractable diagnostic accuracy data.

Information sources and search strategy

Database-restricted searches were run against PubMed/MEDLINE, the Cochrane Library, and IEEE Xplore. Embase was inaccessible and is disclosed as a gap. Search terms combined diabetic retinopathy, artificial intelligence/autonomous/deep learning, and sensitivity and specificity/diagnostic accuracy, restricted where possible to real-world terminology and U.S. settings. Reference lists were hand-searched. Full search strings are in the Appendix. Study flow is reported following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement.[3]

Study selection and data extraction

Title/abstract and full-text screening and data extraction were performed by a single reviewer, a recognized limitation relative to the dual-reviewer standard recommended for systematic reviews. Extracted diagnostic accuracy values were verified against the original publications; one extraction error (Mehra et al. sensitivity) was corrected. One large study[4] that evaluated EyeArt before its August 2020 FDA clearance was excluded at the eligibility-refinement stage to maintain consistency with the FDA-cleared criterion (see Discussion).

Risk of bias assessment

Each included study was assessed with the Quality Assessment of Diagnostic Accuracy Studies–2 (QUADAS-2) tool,[5] a validated instrument that rates risk of bias and applicability concerns across four domains: patient selection, index test, reference standard, and flow and timing.

Data synthesis

Sensitivity and specificity were pooled using a random-effects model with the DerSimonian–Laird estimator of between-study variance,[6] applied on the logit scale. A random-effects model was selected a priori rather than a fixed-effect model because the included studies were expected to differ meaningfully in patient populations, imaging equipment, and reference standards; a fixed-effect model assumes a single true underlying sensitivity and specificity shared by all studies, an assumption that is unlikely to hold for real-world deployments across different sites. The DerSimonian–Laird approach estimates this between-study variance (τ^2) using a moment-based method and incorporates it into the study weights, so that no single study dominates the pooled estimate purely because it has a large sample size. A continuity correction was applied to studies with perfect (100%) sensitivity since the logit transformation is undefined at 0% or 100%. Results are reported with 95% confidence intervals (CI), the range within which the true pooled value is expected to fall with 95% confidence given the observed data, and with the I^2 statistic,[7] which quantifies the percentage of total variability across studies attributable to genuine between-study heterogeneity rather than chance (I^2 values above roughly 75% are conventionally considered very high). The primary analysis used inverse-variance study-level weights. Positive and negative likelihood ratios (LR+ and LR−) and the diagnostic odds ratio (DOR) were derived from the pooled sensitivity and specificity estimates. Certainty of evidence was assessed with the GRADE approach for diagnostic tests.[8]

Study Selection Results

Database searches identified 29 records; six additional records came from citation tracking. After de-duplication, 32 records were screened. Twenty-one full-text articles were assessed; 17 were excluded (pivotal trials n = 3, non-U.S. n = 5, not an FDA-cleared autonomous system including pre-clearance EyeArt n = 5, no extractable sensitivity/specificity n = 4). Four studies met the inclusion criteria (Fig. 1). This is a small number of studies and a modest total sample (approximately 1,429 screening encounters or eyes); the implications of this limited evidence base for the precision and interpretability of the results are discussed explicitly below.

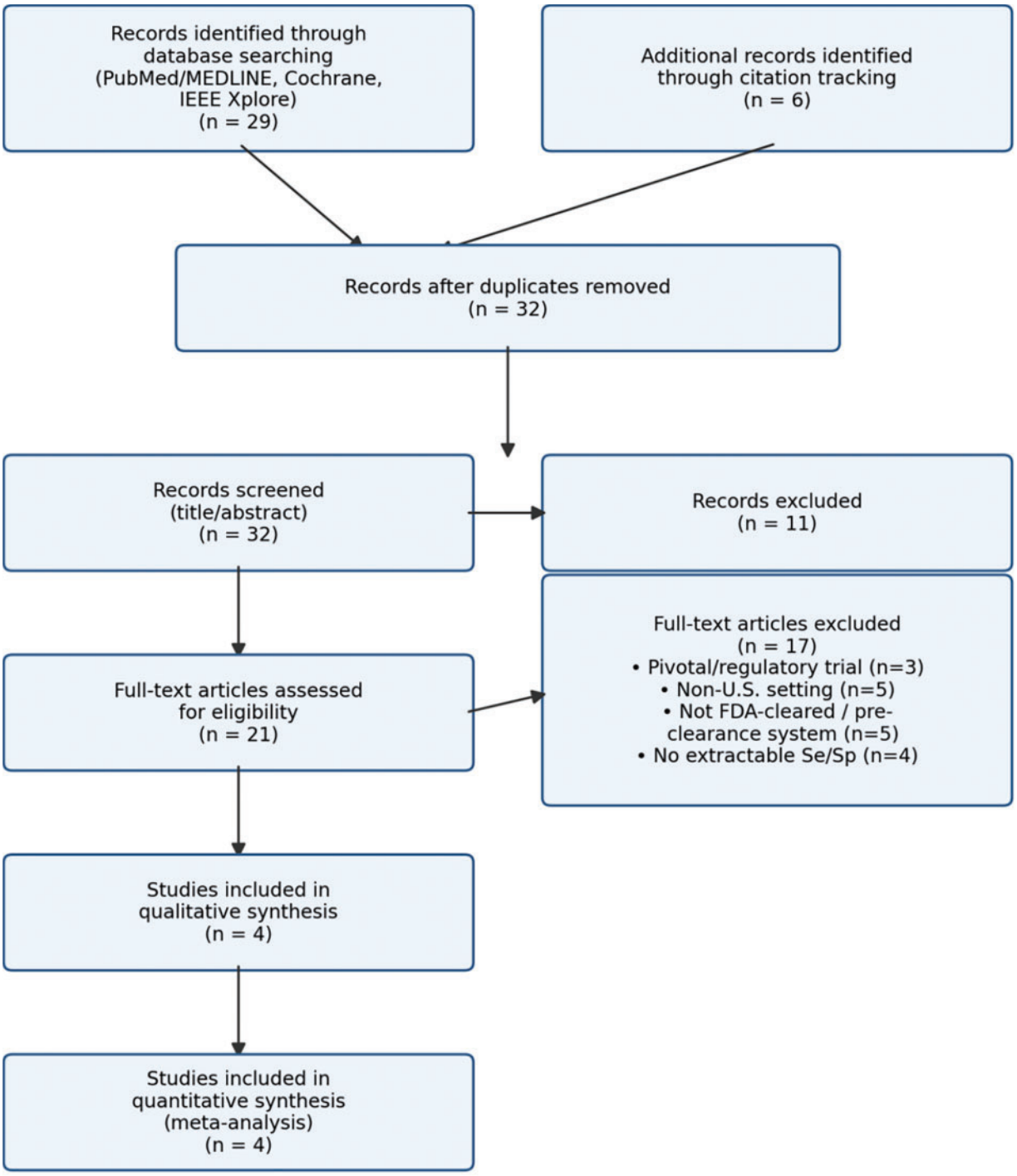

Figure 1. PRISMA 2020 flow diagram. Bhaskaranand et al.[4] was excluded because it evaluated EyeArt before FDA clearance. Original figure generated by the author from the review's own screening counts

Study Characteristics

The four included studies comprised approximately 1,429 screening encounters or eyes and evaluated FDA-cleared IDx-DR/LumineticsCore (two studies) or EyeArt (two studies) in real-world U.S. settings (Table 1).

Risk of Bias Results

No study was low risk across all QUADAS-2 domains except Mehra et al. Common concerns were unclear reference-standard rigor and patient selection in smaller or less representative samples (Fig. 2, presented at the point of first

Table 1. Characteristics of included studies (FDA-cleared systems only; values verified)

Study	Setting	N	AI system	Reference standard	Sensitivity	Specificity
[9]	FL, Mayo Clinic	965*	IDx-DR	Manual overread (all)	100% (90.8–100)	89.2% (87.0–91.1)
[10]	PA, Temple Univ.	260	EyeArt	Optometrist ± retina adj.	100%	77.8%
[11]	NJ, community	124 eyes	EyeArt	Retina specialist	74%	87%
[12]	CA, Stanford	80	IDx-DR	In-person retina exam	95.5% (86.7–100)	60.3% (47.7–72.9)

Note: *Gradable subset of 1,052. Mehra sensitivity confirmed as 100% from source paper.

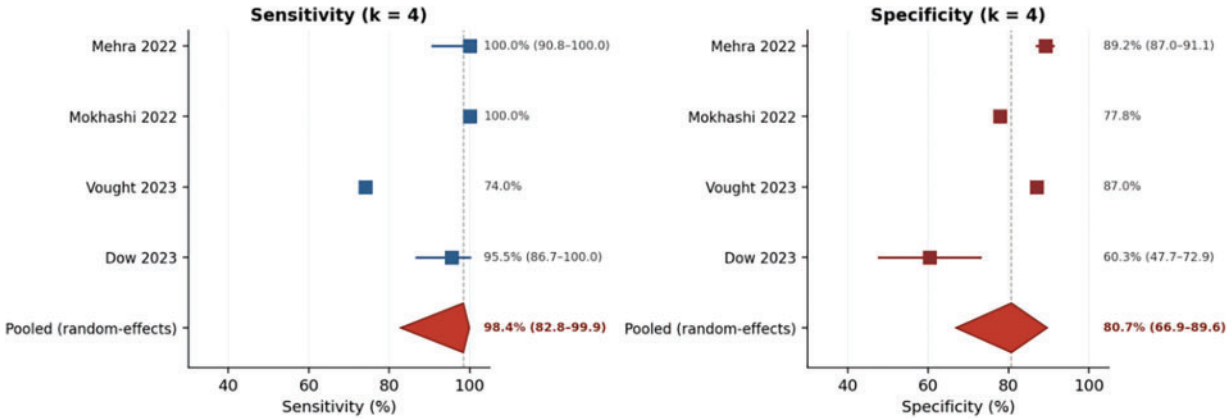

Figure 2. QUADAS-2 risk-of-bias summary (k = 4) across the patient selection, index test, reference standard, and flow-and-timing domains, based on the author's own domain-level judgments as described in the text. Original figure generated by the author

Table 2. Pooled diagnostic accuracy (k = 4)

Outcome	Pooled estimate	95% CI	I^2	τ^2
Sensitivity	98.4%	82.8%–99.9%	93.9%	—
Specificity	80.7%	66.9%–89.6%	94.7%	—

mention below, reflects the domain-level judgments underlying this summary; it is shown here for reference and repeated in the Discussion, where risk of bias is interpreted alongside certainty of evidence).

Primary Meta-Analysis

Random-effects pooling yielded a pooled sensitivity of 98.4% (95% CI: 82.8–99.9%; I^2 = 93.9%) and a pooled specificity of 80.7% (95% CI: 66.9–89.6%; I^2 = 94.7%) (Table 2, Fig. 3). Both outcomes showed very high heterogeneity, meaning that the differences in observed sensitivity and specificity across the four studies are far larger than would be expected from chance alone. Specificity ranged from 60.3% (Dow) to 89.2% (Mehra), a 29-percentage-point spread across only four studies. Possible sources of this heterogeneity are examined below.

Bivariate/SROC Exploration

Pooled estimates from bivariate exploration were essentially identical to the univariate results, and the covariance parameters were imprecise (Fig. 4). This is an expected consequence of the small number of studies (k = 4). Bivariate and hierarchical summary receiver operating characteristic (HSROC) models estimate a between-study correlation between sensitivity and specificity, in addition to their individual means and variances; reliably estimating that correlation typically requires substantially more than four studies. With only four data points, the model has limited ability to distinguish a genuine sensitivity–specificity trade-off from ordinary sampling variation, so the bivariate and univariate pooled point estimates converge, while the correlation and covariance parameters carry very wide, largely uninformative uncertainty. This result should be read as a caution about the limits of the current evidence base rather than as evidence that no trade-off exists.

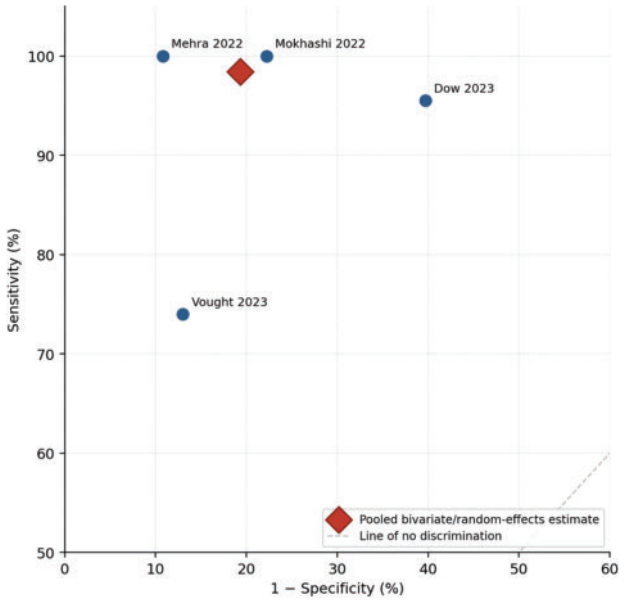

Figure 3. Forest plots of sensitivity and specificity (k = 4), random-effects (DerSimonian–Laird) model. Red diamond = pooled estimate with a 95% confidence interval; point estimates without a plotted interval reflect studies for which the source publication did not report a confidence interval. Original figure generated by the author from the review's own extracted data

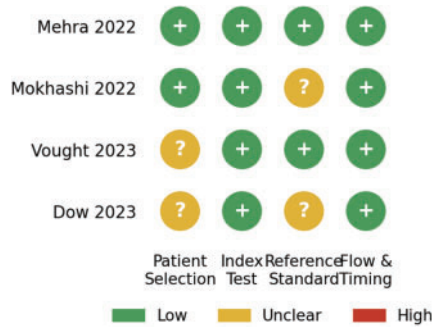

Figure 4. Summary ROC (SROC) space (k = 4) showing individual studies and the pooled estimate. With only four studies, the covariance between sensitivity and specificity is imprecisely estimated (see above). Original figure generated by the author from the review's own extracted data

Impact of the Limited Evidence Base on the Results

The small number of eligible studies (k = 4) and modest pooled sample (~1,429 encounters/eyes) affect these results in several concrete ways. First, precision is low: the 95% CI for pooled sensitivity spans nearly 17 percentage points (82.8–99.9%) and for specificity spans almost 23 percentage points (66.9–89.6%), which is wide enough that clinically meaningfully different "true" values remain compatible with the data. Second, the pooled estimates are numerically dominated by the largest study (Mehra et al., n = 965), so the overall result may not represent smaller or differently configured deployment sites as well as it represents high-volume academic telemedicine programs. Third, with only four studies, it was not possible to conduct meaningful subgroup or meta-regression analyses (for example, by AI system, camera type, or reference standard) that could otherwise help explain the very high heterogeneity described above. Fourth, the small k directly limited the bivariate/SROC modeling described above. Finally, the GRADE certainty for both outcomes was downgraded for imprecision on this basis (see Discussion). Taken together, these results should be regarded as an early, low-certainty signal rather than a stable estimate of real-world performance, and they motivate continued real-world data collection as more FDA-cleared systems are deployed.

Sources of Heterogeneity

Several plausible, non-mutually-exclusive sources likely contributed to the very high I^2 values observed for both sensitivity (93.9%) and specificity (94.7%):

Fundus camera systems and image acquisition protocols. Studies differed in whether imaging was performed with non-mydriatic (undilated) cameras versus dilated photography, and in the number and field of view of images captured per eye. Non-mydriatic imaging, typical of point-of-care primary-care deployment, tends to produce more ungradable or borderline images than dilated, standardized photography, which can shift both false-positive and false-negative rates.

Patient populations and case mix. The four studies were conducted in settings ranging from a large academic telemedicine network (Mehra, Mayo Clinic) to a single urban health system (Mokhashi, Temple), a community screening event (Vought), and an academic ophthalmology hybrid workflow (Dow, Stanford). Differences in underlying DR prevalence and disease severity distribution across these populations can materially shift observed sensitivity and, especially, specificity through a spectrum effect, even when the underlying AI system performance is constant.

Reference standard rigor. Reference standards ranged from full manual overread of all images (Mehra) to optometrist grading with selective retina-specialist adjudication (Mokhashi), dedicated retina-specialist examination (Vought), and in-person retina examination (Dow). Reference standards with different sensitivity to mild disease will systematically shift the AI system's apparent performance relative to that standard.

Referral thresholds and workflow design. Some deployments used AI results as part of a hybrid workflow with reflex dilation or human overread (Mehra, Mokhashi), while others compared AI output more directly to specialist examination (Vought, Dow); differing thresholds for what counts as a "referable" result affect both sensitivity and specificity.

Table 3. Likelihood ratios and diagnostic odds ratio

Metric	Estimate	Interpretation
Positive likelihood ratio (LR+)	5.09	Moderate increase in post-test odds given a positive result
Negative likelihood ratio (LR−)	0.02	Large decrease in post-test odds given a negative result
Diagnostic odds ratio (DOR)	253	Overall discriminatory ability

Table 4. Comparison of pivotal-trial and real-world diagnostic accuracy

AI system	Trial type	Setting	N	Sensitivity	Specificity
IDx-DR	Pivotal (FDA registration)[1]	Primary care, enriched population, 10 U.S. sites	900	87.2% (81.8–91.2)	90.7% (88.3–92.7)
IDx-DR	Real-world (this review)	Mayo Clinic FL; Stanford CA	965/80	95.5–100%	60.3–89.2%
EyeArt	Pivotal (FDA registration)[2], mtmDR endpoint	Primary/eye care, 15 U.S. sites	942	95.5% (92.4–98.5)	85.0% (82.6–87.4)
EyeArt	Real-world (this review)	Temple Univ. PA; NJ community	260/124 eyes	74–100%	77.8–87%
Both systems, pooled	Real-world random-effects meta-analysis (this review)	Four U.S. sites, mixed system	∼1,429	98.4% (82.8–99.9)	80.7% (66.9–89.6)

Note: [1] Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med. 2018;1:39. [2] Ipp E, Liljenquist D, Bode B, et al. Pivotal evaluation of an artificial intelligence system for autonomous detection of referrable and vision-threatening diabetic retinopathy. JAMA Netw Open. 2021;4(11):e2134254 (mtmDR endpoint reported here; values as published, not company press-release figures).

Because only four studies were available, these explanations cannot be formally tested through meta-regression or subgroup meta-analysis; they are offered as plausible, literature-consistent explanations for the observed heterogeneity rather than as statistically confirmed causes. This heterogeneity, combined with the small number of studies, is the principal reason the GRADE certainty rating for both outcomes was downgraded to Low (see Discussion) and is a central reason the pooled estimates should be interpreted cautiously.

Discussion and Recommendations for Future Work

Principal findings

In real-world U.S. deployments of FDA-cleared autonomous AI systems for DR screening, pooled sensitivity is very high (98.4%) while specificity is more variable (60.3–89.2%; pooled 80.7%) and, on the whole, lower than typical pivotal-trial values (Table 4). The negative likelihood ratio of 0.02 supports a strong rule-out role (Table 3). The positive likelihood ratio of 5.09 indicates only a moderate increase in post-test probability after a positive result, consistent with meaningful false-positive burden in some real-world settings. As emphasized above, these findings rest on only four studies and should be read with that limitation in mind.

Comparison with pivotal-trial performance

Table 4 places the published FDA pivotal-trial results for each system alongside the real-world results identified in this review.

This comparison shows that pooled real-world sensitivity in this review (98.4%) was numerically higher than either pivotal trial's sensitivity, while pooled real-world specificity (80.7%) was lower than both pivotal-trial specificity values (90.7% for IDx-DR and 85.0% for EyeArt). At the individual-study level, the lowest real-world specificity (60.3%, Dow) fell roughly 30 percentage points below the corresponding IDx-DR pivotal-trial figure. This pattern is consistent with the sources of heterogeneity discussed above: pivotal trials used standardized, often dilated imaging and enriched, prospectively consented populations, whereas real-world deployments contend with more variable image quality, broader case mix, and differing reference standards, all of which can lower observed specificity even when the underlying algorithm is unchanged. Given that this comparison draws on only four real-world studies and two pivotal trials, it should be treated as descriptive rather than as a formal statistical comparison.

Table 5. GRADE certainty assessment

Outcome	Risk of bias	Inconsistency	Indirectness	Imprecision	Certainty
Sensitivity	Serious	Very serious	Not serious	Serious	LOW
Specificity	Serious	Very serious	Not serious	Serious	LOW

Exclusion of Bhaskaranand et al.[4]

Bhaskaranand et al.[4] evaluated EyeArt v2.0 in a large EyePACS telemedicine network (N ≈ 100,000) and reported a sensitivity of 91.3% and a specificity of 91.1%. EyeArt did not receive FDA clearance until August 2020. To maintain a strict FDA-cleared eligibility criterion, this study was excluded. Its inclusion would have numerically dominated inverse-variance pooling and raised pooled specificity; its exclusion produces more conservative specificity estimates based only on post-clearance deployments, but it is also part of why the final evidence base is so small.

Implications for implementation and equity

Although the diagnostic decision is rendered autonomously, these systems operate inside human clinical workflows. High sensitivity supports their use as rule-out tools, but variable specificity increases false-positive referrals, clinic burden, and potential patient anxiety. These effects may fall disproportionately on under-resourced safety-net and rural settings. Successful deployment depends on image quality, staff training, connectivity, electronic health record integration, and closed-loop referral pathways.

Certainty of evidence (GRADE)

Certainty was rated Low for both outcomes (Table 5) using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework for diagnostic test accuracy, due to serious risk of bias, very serious inconsistency (I^2 > 90%), and serious imprecision (wide confidence intervals; only four studies).

Limitations

This review has several important limitations. Most centrally, only four studies met the strict FDA-cleared, real-world, U.S.-only eligibility criteria, and the pooled sample of roughly 1,429 encounters/eyes is small for a diagnostic meta-analysis; this limits statistical precision, prevented formal subgroup analysis of heterogeneity sources, and constrained the bivariate/SROC modeling. Embase was inaccessible. Screening and extraction were performed by a single reviewer rather than in duplicate. Full 2 × 2 tables were not available for every study, requiring reliance on reported summary sensitivity and specificity. Heterogeneity was very high across both outcomes. The review was not registered on PROSPERO. Finally, because all four included studies were conducted in U.S. clinical and regulatory settings, the findings describe U.S. real-world performance specifically; they are not applicable to other countries or healthcare systems, which may use different AI devices or device versions, different regulatory clearance pathways, different reference-standard conventions, and different patient populations. Pooled estimates should be interpreted as rough, low-certainty central tendencies specific to the U.S. context rather than as precise or globally generalizable performance figures.

Recommendations for future work

Building directly on the limitations identified above, four directions would most improve the evidence base. First, more real-world U.S. studies are needed, ideally with dual independent screening and full 2 × 2 data reported, so that future meta-analyses can achieve narrower confidence intervals and support formal subgroup and meta-regression analysis of heterogeneity sources (camera type, reference standard, and patient population). Second, prospective multi-site studies that hold camera hardware and reference-standard protocol constant across sites would help isolate how much of the observed heterogeneity is attributable to imaging versus population versus reference-standard differences. Third, future studies should report downstream outcomes beyond sensitivity and specificity alone, including referral completion rates, time to treatment, and performance stratified by clinic resource level and patient demographics, to clarify the real-world equity implications discussed above. Fourth, as additional autonomous AI systems obtain FDA clearance, this review should be updated and, ideally, prospectively registered on PROSPERO with dual independent screening to strengthen the certainty of future conclusions.

Conclusions

This review systematically identified and meta-analyzed real-world U.S. evidence on two FDA-cleared, fully autonomous AI systems for diabetic retinopathy screening. Among the four eligible studies (≈1,429 encounters/eyes), pooled sensitivity was high (98.4%, 95% CI 82.8–99.9%) while specificity was substantially more variable (pooled 80.7%, 95% CI 66.9–89.6%; study range 60.3–89.2%) than pivotal-trial data alone would suggest (Table 4). This conclusion is based on a small and heterogeneous evidence base (I^2 > 90% for both outcomes) that yields wide confidence intervals and Low GRADE certainty for both sensitivity and specificity; it should therefore be treated as an early, imprecise signal rather than a definitive estimate of real-world performance. Because every included study was conducted in the United States, these conclusions apply to the U.S. regulatory and clinical context specifically and should not be assumed to

generalize to other countries or healthcare systems. Site-level validation before and during deployment, and continued accumulation of real-world evidence across a larger and more diverse set of U.S. sites, remain essential.

Acknowledgements

The author thanks Rohina Swaroop, MD (Ophthalmology), for her mentorship and guidance throughout this project. Her clinical expertise and thoughtful feedback strengthened the design, interpretation, and presentation of this work. Any remaining errors are the author's own.

References

[1] Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med*. 2018;1:39. doi:10.1038/s41746-018-0040-6.

[2] Ipp E, Liljenquist D, Bode B, et al. Pivotal evaluation of an artificial intelligence system for autonomous detection of referrable and vision-threatening diabetic retinopathy. *JAMA Netw Open*. 2021;4(11):e2134254. doi: 10.1001/jamanetworkopen.2021.34254.

[3] Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA, 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. doi: 10.31222/osf.io/v7gm2_v1.

[4] Bhaskaranand M, Ramachandra C, Bhat S, et al. The value of automated diabetic retinopathy screening with the EyeArt system: a study of more than 100,000 consecutive encounters from people with diabetes. *Diabetes Technol Ther*. 2019;21(11):635–643. doi: 10.1089/dia.2019.0164.

[5] Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529–536. doi: 10.7326/0003-4819-155-8-201110180-00009.

[6] DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177–188. doi:10.1016/j.cct.2015.09.002.

[7] Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327(7414):557–560. doi:10.1136/bmj.327.7414.557.

[8] Schünemann HJ, Oxman AD, Brozek J, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ*. 2008;336(7653):1106–1110. doi: 10.1136/bmj.39500.677199.ae.

[9] Mehra AA, Softing A, Guner MK, Hodge DO, Barkmeier AJ. Diabetic retinopathy telemedicine outcomes with artificial intelligence-based image analysis, reflex dilation, and image overread. *Am J Ophthalmol*. 2022;244:125–132. doi: 10.1016/j.ajo.2022.08.008.

[10] Mokhashi N, Grachevskaya J, Cheng L, et al. A comparison of artificial intelligence and human diabetic retinal image interpretation in an urban health system. *J Diabetes Sci Technol*. 2022;16(4):1003–1007. doi: 10.1177/1932296821999370.

[11] Vought R, Vought V, Shah M, Szirth B, Bhagat N. EyeArt artificial intelligence analysis of diabetic retinopathy in retinal screening events. *Int Ophthalmol*. 2023;43(12):4851–4859. doi:10.1007/s10792-023-02887-9.

[12] Dow ER, Khan NC, Chen KM, et al. AI-human hybrid workflow enhances teleophthalmology for the detection of diabetic retinopathy. *Ophthalmol Sci*. 2023;3(4):100330. doi: 10.1016/j.xops.2023.100330.

Figure and Data Originality Statement

All four figures in this manuscript (Figs. 1–4) were generated originally by the author from the review's own extracted screening counts and diagnostic accuracy data; none are reproduced, adapted, or copied from any journal, website, or third-party source, and no permissions are therefore required for their use.

Appendix A. Search Strategies

PubMed, Cochrane Library, and IEEE Xplore search strings as previously documented. Embase not searched (subscription unavailable). Bhaskaranand et al.[4] was identified but excluded for pre-clearance status of EyeArt (FDA clearance August 2020).

Appendix B. Data Verification and Eligibility Log

Mehra et al.[9]—IDx-DR, post-clearance, sensitivity 100%, specificity 89.2%: included. Mokhashi et al.[10]—EyeArt, post-clearance: included. Vought et al.[11]—EyeArt, post-clearance: included. Dow et al.[12]—IDx-DR, post-clearance: included. Bhaskaranand et al.[4]—EyeArt v2.0, pre-FDA-clearance (clearance Aug 2020): excluded.

About the Author

Shreya Parimoo is a 10th-grade student at Stockdale High School in Bakersfield, California, with research interests in healthcare, psychology, and the application of technology to advance patient care. She is the founder of ShreyaVocabLabs.com, a free vocabulary-learning platform for high school students grounded in principles of cognitive science, work profiled in Seven Oaks Greet Magazine ("Meet Shreya Parimoo: She Turned a Study Frustration into Something Brilliant," May 2026). Her writing on technology, identity, and mental health has appeared in Teen Ink.