

SYSTEMATIC REVIEW / META-ANALYSIS

Telemedicine and Artificial Intelligence for Retinopathy of Prematurity Screening in Low-Resource Settings: A Systematic Review and Meta-Analysis of Diagnostic Accuracy with Cost-Effectiveness Perspectives

I Wayan Eka Sutyanawan¹, Ni Made Ayu Surasmiati¹, Ivana Beatrice Alberta^{2*}, Calvin Prasetyo²

¹ Pediatric Ophthalmology and Strabismus Division, Department of Ophthalmology, Faculty of Medicine, Universitas Udayana, Denpasar, Indonesia

² Resident of Ophthalmology, Faculty of Medicine, Universitas Udayana, Denpasar, Indonesia

ARTICLE HISTORY

Received: August 20, 2026

Revised: September 15, 2026

Accepted: October 2, 2026

Available online: October 9, 2026

Published: October 12, 2026

KEYWORDS

Artificial intelligence
Cost-benefit analysis
Retinopathy of prematurity
Sensitivity and specificity
Telemedicine

CORRESPONDING AUTHOR

Ivana Beatrice Alberta
ivanabeatricealberta@gmail.com

DOI

<https://doi.org/10.37275/cmej.v7i2.953>

ABSTRACT

Background: Screening for retinopathy of prematurity (ROP) in low- and middle-income countries (LMICs) is limited by the shortage of trained examiners.

Objective: We assessed the diagnostic accuracy of telemedicine and artificial intelligence (AI) for ROP screening in LMICs and summarised economic evaluations of screening.

Methods: PubMed, Europe PMC, OpenAlex and Scopus were searched for accuracy studies from LMICs comparing remote grading of retinal images (telemedicine) or AI with ophthalmoscopy or expert grading, and for economic evaluations. Sensitivity and specificity were pooled with bivariate random-effects models; risk of bias was assessed with QUADAS-2 and certainty with GRADE.

Results: Twenty-four accuracy studies from nine countries and 13 economic evaluations were included. For clinically significant ROP, telemedicine (6 datasets; 927 units) had a sensitivity of 92.8% (95% CI 82.6–97.3) and a specificity of 95.8 (87.6–98.7)%; AI (6 datasets; 7,346 units) had a sensitivity of 93.0 (81.5–97.6)% and a specificity of 77.3 (55.6–90.3)%. Specificity was higher in upper-middle-income settings. For any-stage ROP and plus disease, pooled sensitivities were 92.2% and 81.4%, and specificities 97.6% and 99.1%. Fifteen studies were at high risk of bias, and certainty was low to very low. Telemedicine was cost-effective in all five cost-utility analyses; no economic evaluation of AI from LMICs was found.

Conclusion: Telemedicine detected clinically significant ROP with high sensitivity and specificity and was cost-effective in LMICs, although the certainty of evidence was low. AI achieved similar sensitivity with more false-positive referrals; its use as a triage step within telemedicine programmes requires prospective evaluation.

Copyright: © 2026 The Author(s). Published by HM Publisher in collaboration with CMHC Research Center. The authors retain copyright and grant the journal the right of first publication. **Access license:** This is an open-access article under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), <https://creativecommons.org/licenses/by-nc-sa/4.0/>, which permits non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited and derivative works are shared under the same license.

1. Introduction

Retinopathy of prematurity (ROP) is a vasoproliferative disorder of the developing retina and one of the leading causes of avoidable childhood blindness. Analyses of the Global Burden of Disease data showed that ROP-related visual impairment has risen with the survival of preterm infants and that the burden is concentrated in low- and middle-income countries, where neonatal care has expanded faster than eye care for preterm infants.^{1–3} In a fifteen-year comparison from Mexico, ROP remained the leading

cause of blindness among children attending schools for the blind, illustrating how incomplete screening translates into lifelong disability.⁴

The same pattern has been documented across Asia and Africa. In Indonesia, a national multicentre survey found that the incidence of ROP had declined after the introduction of a national protocol and national health insurance. The incidence was probably underestimated because of the high mortality of very preterm infants, and not all at-risk infants were screened because the number of paediatric ophthalmologists trained to examine and treat them was limited.^{5,6} Screening

criteria and practices differed widely between eight Asian neonatal networks, and infants presenting with stage 4B or 5 disease in South India had frequently never been screened.^{7,8} Cohorts from Tanzania, Ethiopia and Vietnam reported a substantial incidence of ROP under screening criteria that included larger and more mature infants than those used in high-income countries, and the incidence in China varied with regional economic development.^{9–12} In Africa, half of the surveyed professionals reported no formal referral pathway, and a South African register showed that a considerable proportion of eligible infants were not screened and examined whether screening began before the onset of severe disease.^{13–15} When disease is detected late, infants require urgent treatment such as intravitreal anti-vascular endothelial growth factor therapy for aggressive ROP, which is available in only a few referral centres.¹⁶

Screening by binocular indirect ophthalmoscopy, performed by an ophthalmologist experienced in ROP and graded according to the International Classification of Retinopathy of Prematurity, remains the reference approach.¹⁷ Surveys showed variable exposure of ophthalmology fellows to ROP training and uncertain intentions to provide ROP care, and identified gaps in knowledge and practical challenges among neonatal nurses in both on-site and telemedicine-based screening.^{18,19} Telemedicine separates image acquisition from interpretation: trained nurses, technicians or non-ophthalmologist physicians capture retinal images with wide-field contact cameras, portable non-contact cameras or smartphones, and experts grade them remotely. Telemedicine and technician-assisted programmes were implemented in India, Guatemala and China, including during the COVID-19 pandemic, when travel to neonatal units was restricted.^{20–23} Newly established programmes in Pakistan, rural Egypt and the Philippines reported a considerable burden of ROP and the need for locally tailored screening criteria and systems that ensure timely examination.^{24–26}

Artificial intelligence (AI) does not replace this workflow but adds to it: a deep learning algorithm classifies the images captured in a telemedicine programme, and only suspicious examinations are sent to a human grader. An AI-derived vascular severity score was used to describe the epidemiology of ROP in large Indian telemedicine programmes and improved the consistency of plus disease diagnosis among experts.^{27–29} However, image-based algorithms can learn features associated with retinal pigmentation and self-reported race, and the field of view and image enhancement alter the grading of plus disease even among experts, so that algorithms developed in high-income countries may perform differently with other

populations and cameras.^{30–32} Edge-cloud architectures and target product profiles for low-cost imaging devices have been proposed to make such systems usable in resource-limited settings.^{33,34}

The choice between these strategies is also an economic decision. A United States model suggested that autonomous AI could be as effective as and less costly than other ROP screening strategies, and screening for diabetic retinopathy with teleophthalmology or AI was cost-effective in India, Hong Kong, mainland China and Australia, with results that depended strongly on the sensitivity and specificity of the algorithm.^{35–39} Combined screening for blinding eye diseases and prevention of ROP with oxygen blenders were likewise cost-effective in China and Uganda.^{40,41} Whether the accuracy and cost-effectiveness of ROP telemedicine and AI hold in low- and middle-income countries, where infants are larger, imaging is often performed by non-physicians and portable cameras are common, has not been synthesised; the searches for this review identified no previous synthesis restricted to data from these settings.

The novelty of this study lies in restricting the synthesis of diagnostic accuracy to data collected in low- and middle-income countries, in comparing telemedicine grading and AI within the same bivariate framework for three target conditions, and in placing these estimates alongside the economic evaluations conducted in the same settings. The aim of this study was to estimate the sensitivity and specificity of telemedicine and AI for detecting clinically significant ROP, any-stage ROP and plus disease in low- and middle-income countries, to explore sources of heterogeneity, and to summarise the cost-effectiveness of ROP screening strategies in these settings.

2. Methods

2.1. Study design and reporting

This systematic review and meta-analysis of diagnostic test accuracy followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement and the methods of the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy.^{42,43} Economic evaluations were appraised with the items of the Consolidated Health Economic Evaluation Reporting Standards (CHEERS 2022).⁴⁴ The review used published data only, and ethical approval was not required.

2.2. Eligibility criteria

Diagnostic accuracy studies were eligible when they enrolled preterm infants who met local ROP screening criteria in a low- or middle-income country, defined by

the World Bank classification at the time of data collection; low-resource settings were operationalised in this way. Two index tests were considered: telemedicine, defined as retinal imaging with any wide-field contact, portable non-contact or smartphone-based camera followed by remote or deferred grading by a human reader; and AI, defined as an automated algorithm that classified ROP images and was evaluated on data that had not been used for training (temporal, external or hold-out test sets). The reference standard was binocular indirect ophthalmoscopy by an ophthalmologist, a clinical diagnosis recorded in the screening programme, or consensus grading of images by ROP experts.

The World Bank grouping of low-, lower-middle- and upper-middle-income countries was used as the operational definition of low-resource settings for three reasons: it is reproducible across studies and years, most of the current burden of ROP-related blindness lies in middle-income countries, and access to screening within upper-middle-income countries remains uneven between regions.^{1,12,25} Because resource levels differ within this group, all analyses were also reported by income group.

Three target conditions were analysed. Clinically significant ROP comprised treatment-requiring ROP (type 1 ROP or ROP requiring treatment) and referral-warranted ROP (type 2 ROP or worse, stage 3 ROP, plus disease or more-than-mild ROP); the threshold used by each study was recorded. Any-stage ROP was defined as stage 1 or worse, and plus disease as plus versus no plus, with pre-plus counted as negative.¹⁷ Studies were excluded when they reported data from a high-income country, when a two-by-two table of true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN) could not be obtained or reconstructed consistently with the reported estimates, when AI was evaluated by cross-validation only, or when the same unmasked examiner performed the index test and the reference standard.

Economic evaluations were eligible when they reported costs, or costs together with health outcomes, of an ROP screening strategy in a low- or middle-income country, including cost-utility, cost-effectiveness, cost-benefit and cost analyses of telemedicine, AI or ophthalmoscopy-based programmes.

2.3. Information sources and search strategy

PubMed, Europe PMC, OpenAlex and Scopus were searched from inception without language restrictions. The diagnostic accuracy search combined terms for ROP with terms for telemedicine, digital retinal imaging, smartphone imaging, deep learning and AI;

the economic search combined ROP terms with terms for costs, cost-effectiveness, cost-utility and economic evaluation. Medical Subject Headings were combined with free-text synonyms, and the reference lists of included studies were screened.

Records in languages other than English were retained when they were indexed in the searched databases; three included studies were published in Chinese with English abstracts, and their full texts were read in the original with the assistance of translation, with the two-by-two tables checked against the published tables. Embase, Web of Science and Chinese databases such as the China National Knowledge Infrastructure were not searched.

2.4. Study selection and data extraction

Records were deduplicated by digital object identifier, PubMed identifier and normalised title. Titles and abstracts were screened by one reviewer, and every full-text eligibility decision and every extracted two-by-two table was verified by a second reviewer against the source document; disagreements were resolved by discussion. Data were extracted into a piloted form with 23 items covering bibliographic details, country and income group, design, setting, participants (gestational age and birth weight), imaging device, person acquiring the images, reader or algorithm, reference standard, target condition, threshold, unit of analysis and the two-by-two counts.

Two-by-two tables were taken directly from the reports or reconstructed from reported counts, sensitivities, specificities and denominators; every reconstructed table was checked by recomputing the published estimates. When a study reported several readers, the first or consensus reader was used in the primary analysis and the second reader in a sensitivity analysis. When both referral-warranted and treatment-requiring thresholds were reported, the referral-warranted threshold was used in the primary analysis because it reflects the screening decision. Data were analysed at the unit reported, with infant-level tables preferred over eye-, examination- or image-level tables when several were available. Economic data comprised the type of evaluation, perspective, time horizon, discount rate, strategies compared, costs, effects, incremental cost-effectiveness ratios and willingness-to-pay thresholds. All two-by-two tables, including those of second readers and alternative thresholds, are listed in Supplementary Table S2.

Units with ungradable images were analysed as handled by each study. Some programmes counted them as screen positive and referred the infant, which lowers specificity, whereas others excluded them, which can raise both sensitivity and specificity; this

was recorded in the flow and timing domain of the risk-of-bias assessment. Datasets from independent populations reported in one article, such as the Nepalese and Mongolian cohorts of one AI study, were analysed as separate datasets. One study reported telemedicine and AI results for overlapping infants; both datasets entered the analyses by index test, and the pooled analysis of all datasets was treated as secondary to the index-specific estimates.

2.5. Risk of bias and applicability

Risk of bias and applicability concerns were assessed with QUADAS-2 in four domains: patient selection, index test, reference standard, and flow and timing.⁴³ Because no validated AI-specific version was available, the signalling questions were tailored for AI studies. For patient selection, consecutive or complete programme samples were rated at low risk, whereas curated, case-control or class-balanced image sets were rated at high risk. For the index test, the test set had to be independent of the training data at the infant level and the operating threshold had to be fixed before testing; image-level splits in which images of the same infant could appear in both sets were rated at high risk. For the reference standard, ophthalmoscopy or a programme diagnosis masked to the index test was preferred over image labels assigned by the developers. For flow and timing, the handling of ungradable images and the interval between the index test and the reference standard were assessed. A study was rated at high overall risk when any domain was at high risk, at low risk when all domains were at low risk, and unclear otherwise. The reporting of economic evaluations was appraised against the CHEERS 2022 items for the perspective, time horizon, discounting and sensitivity analysis, and it was recorded whether the full text or only the abstract was available.⁴⁴

2.6. Statistical analysis

For each dataset, sensitivity and specificity were calculated from the two-by-two table as shown in Equation 1, with exact Clopper-Pearson 95% confidence intervals (CI).

$$Se = \frac{TP}{TP+FN}, Sp = \frac{TN}{TN+FP} \quad (1)$$

Sensitivity and specificity were pooled with the bivariate binomial-normal random-effects model. The model describes the number of true positives among units with the target condition and the number of true negatives among units without it with binomial distributions (Equation 2). The logit-transformed sensitivity and specificity of each study are assumed to follow a bivariate normal distribution N with mean vector (μ_A, μ_B) and between-study covariance matrix Σ (Equation 3), where A denotes sensitivity, B specificity, σ^2 the between-study variances and σ_{AB} their

covariance. The model was fitted by maximum likelihood with the Laplace approximation. When the correlated model was singular or did not converge, the covariance term was set to zero.

$$y_{Ai} \sim \text{Bin}(n_{Ai}, p_{Ai}), y_{Bi} \sim \text{Bin}(n_{Bi}, p_{Bi}) \quad (2)$$

$$\begin{pmatrix} \text{logit}(p_{Ai}) \\ \text{logit}(p_{Bi}) \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \Sigma \right), \Sigma = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix} \quad (3)$$

Positive and negative likelihood ratios and the diagnostic odds ratio were derived from the pooled estimates (Equation 4), and their 95% CIs were obtained from 20,000 draws from the bivariate normal distribution of the estimated logit means. Hierarchical summary receiver operating characteristic (HSROC) parameters were derived from the bivariate estimates (Equation 5), and the HSROC curve (Equation 6) was plotted with the 95% confidence region of the summary point and the 95% prediction region.

$$LR^+ = \frac{Se}{1-Sp}, LR^- = \frac{1-Se}{Sp}, DOR = \frac{LR^+}{LR^-} \quad (4)$$

$$\beta = \ln \left(\frac{\sigma_B}{\sigma_A} \right), \Lambda = \left(\frac{\sigma_B}{\sigma_A} \right)^{1/2} \mu_A + \left(\frac{\sigma_A}{\sigma_B} \right)^{1/2} \mu_B \quad (5)$$

$$\text{logit}(Se) = \Lambda e^{-\beta/2} + e^{-\beta} \text{logit}(1-Sp) \quad (6)$$

Heterogeneity was assessed mainly from the between-study variances of the bivariate model and the 95% prediction region. The I^2 statistic of univariate random-effects logistic models (Equation 7) was reported as a descriptive measure, recognising that it tends to be high in accuracy reviews with large samples.

$$I^2 = \max \left\{ 0, \frac{Q-df}{Q} \right\} \times 100\% \quad (7)$$

Sources of heterogeneity were explored by bivariate meta-regression, adding each covariate to the model as a fixed effect on logit sensitivity and logit specificity and comparing it with the model without the covariate by a likelihood-ratio test with two degrees of freedom (Equation 8). The prespecified covariates were the index test, income group, reference standard (ophthalmoscopy or clinical diagnosis versus expert image grading), imaging device, person acquiring the images, unit of analysis, overall risk of bias and, for clinically significant ROP, the threshold. These analyses were exploratory; to account for multiple testing, covariates were regarded as robustly associated with accuracy only when their p value was below the Bonferroni threshold (0.05 divided by the number of covariates tested, 0.0063 for clinically significant ROP).

$$G^2 = -2(\ell_0 - \ell_1) \sim \chi^2_2 \quad (8)$$

Sensitivity analyses excluded studies at high risk of bias, studies with an image-based reference standard, non-infant-level data and image-level data. Further analyses replaced the primary reader with the second reader, applied the treatment threshold to all studies,

excluded the prognostic AI model, and excluded case-control or class-balanced designs and studies without reported masking of readers. Finally, each dataset was omitted in turn.

Small-study effects were assessed with Deeks' funnel plot asymmetry test when at least ten datasets were available, by regressing the log diagnostic odds ratio on the inverse square root of the effective sample size (Equation 9), where n_1 and n_0 are the numbers of units with and without the target condition; a p value below 0.10 indicated asymmetry.

$$\ln\text{DOR}_i = \alpha + \beta_1 \frac{1}{\sqrt{\text{ESS}_i}}, \text{ESS}_i = \frac{4n_{1i}n_{0i}}{n_{1i}+n_{0i}} \quad (9)$$

Analyses were performed in R version 4.5.3 with the lme4 (version 2.0.1), meta (version 8.2.1) and MASS (version 7.3.65) packages, and figures were drawn in Python with matplotlib (version 3.10.8). Economic evaluations were not pooled because they differed in perspective, comparators, currencies and price years. They were synthesised narratively in their original currencies, and each incremental cost-effectiveness ratio was interpreted against the willingness-to-pay threshold used in its own study, because several studies did not report a price year and conversion to a common currency would have required additional assumptions.

2.7. Certainty of the evidence

The certainty of the evidence for clinically significant ROP was rated with the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach for diagnostic tests, starting at high certainty for cross-sectional accuracy studies and downgrading for risk of bias, indirectness, inconsistency, imprecision and publication bias. Publication bias was judged from Deeks' test, taking its limited power into account. The

consequences of screening were expressed as true and false results per 1000 infants at pre-test probabilities of 5%, 10% and 20%. These values spanned most of the observed range of prevalence of clinically significant ROP in the included datasets (2.0% to 24.0%; median 9.1%). False negatives and false positives were calculated as shown in Equation 10, where π is the pre-test probability.

$$\text{FN}_{1000} = 1000\pi(1 - \text{Se}), \text{FP}_{1000} = 1000(1 - \pi)(1 - \text{Sp}) \quad (10)$$

3. Results

3.1. Study selection

The diagnostic accuracy searches identified 2,376 records, of which 986 remained after removal of duplicates. After screening of titles and abstracts, 77 reports were sought, 28 could not be retrieved and 49 were assessed in full; 25 were excluded, most often because two-by-two data were not reported or the reported accuracy measures were internally inconsistent. The economic searches identified 512 records (212 after deduplication), and 19 reports were assessed in full. In total, 24 diagnostic accuracy studies and 13 economic evaluations were included, as detailed in Figure 1.

Most of the 28 reports that could not be retrieved described AI algorithms published in engineering or computing journals; the others evaluated smartphone or digital imaging devices. The eight reports excluded because of internal inconsistency included both telemedicine and AI studies; in these, the reported sensitivity and specificity could not be reproduced from any integer two-by-two table with the stated denominators, or the counts in the text and the tables contradicted each other.

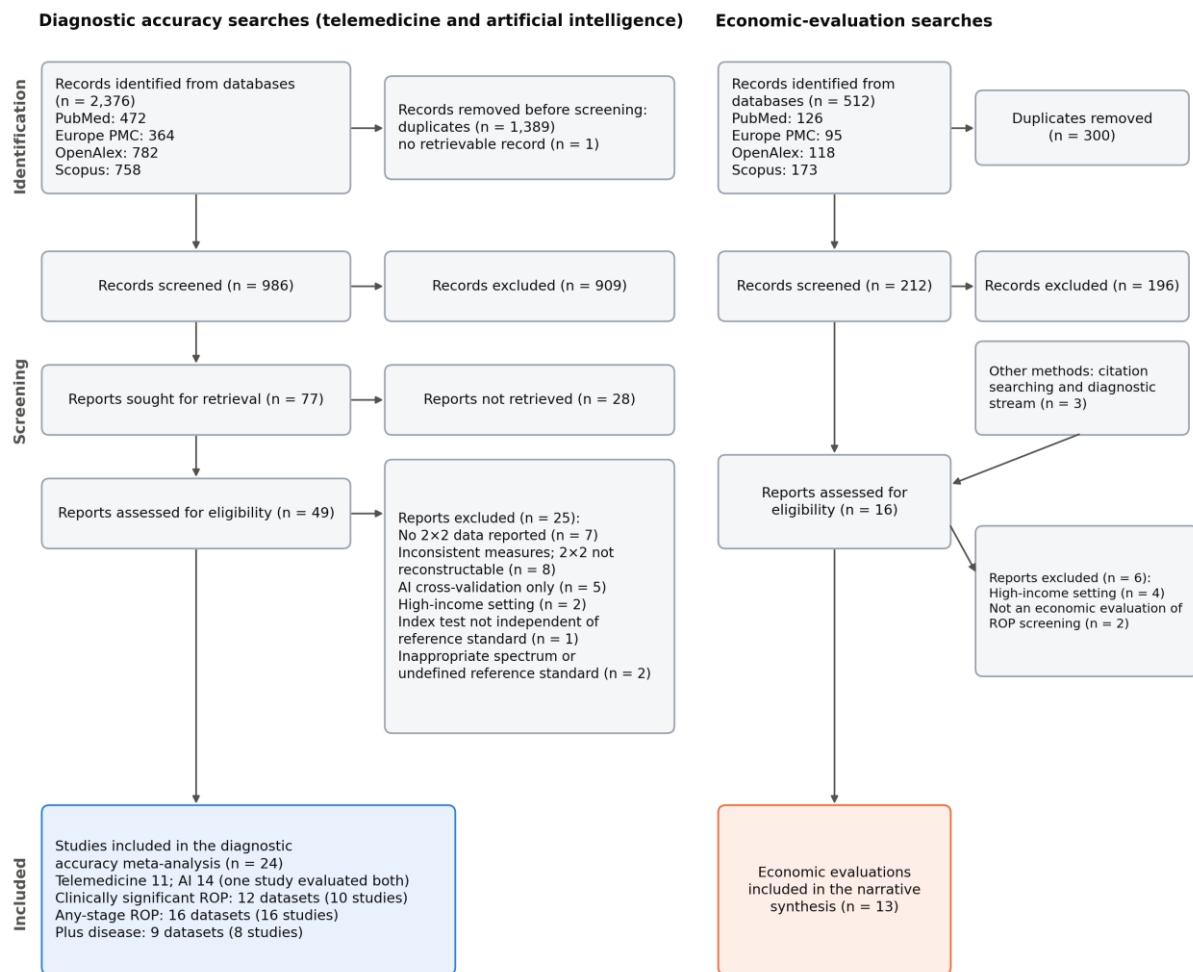


Figure 1. PRISMA 2020 flow diagram of the identification, screening and inclusion of diagnostic accuracy studies (left) and economic evaluations (right). AI, artificial intelligence; ROP, retinopathy of prematurity.

3.2. Characteristics of the included studies

The twenty-four diagnostic accuracy studies were conducted in nine countries, as summarised in Table 1. China contributed 9 studies, India 8, Thailand 2 and Nepal 2; Turkey and Brazil contributed one study each. One AI study contributed separate cohorts from Nepal and Mongolia, and another combined images from

Argentina and Colombia. Eleven studies evaluated telemedicine grading and fourteen evaluated AI; one study assessed both on paired images. Eleven studies were from low- or lower-middle-income settings and thirteen from upper-middle-income settings; only the Nepalese data came from a low-income country. The eighteen studies that reported the number of infants contributed 9,519 infants.

Table 1. Characteristics of the included diagnostic accuracy studies.

Study	Country (income group)	Design and setting	Infants (n); GA; BW	Units analysed	Index test: device; image acquisition; reading	Reference standard	Targets	RoB
Telemedicine (image-based human grading)								
Shah 2006 ⁴⁵	India (lower-middle)	Prospective comparative, single centre	27; GA 32.3 (28–36); BW 1469 (900–2050)	87 examinations	RetCam 120; acquisition: ophthalmologist; reading: masked deferred reading by vitreoretinal surgeon	BIO with scleral depression (vitreoretinal surgeon)	Any-stage	High
Wu 2010 ⁴⁶	China (lower-middle)	Prospective paired comparison, single centre	112; GA 31.7 ± 2.7; BW 1645 ± 420	224 eyes	RetCam II; acquisition: ophthalmologist; reading: ophthalmologist; masking not reported	BIO with scleral depression (experienced ophthalmologist)	Any-stage	Unclear

Study	Country (income group)	Design and setting	Infants (n); GA; BW	Units analysed	Index test: device; image acquisition; reading	Reference standard	Targets	RoB
Sekeroglu 2013 ⁴⁷	Turkey (upper-middle)	Prospective, consecutive, masked, single centre	58; GA 30 (24–32); BW 1335 (760–2000)	248 eye-examinations	RetCam Shuttle; acquisition: ophthalmologist; reading: imaging ophthalmologist masked to BIO (different examiner)	BIO with scleral depression (paediatric ophthalmologist)	CS, any-stage, plus	Low
Vinekar 2014 ²⁰	India (lower-middle)	Prospective programme data (KIDROP), multicentre	NR; GA NR; BW NR	454 imaging sessions (level III technicians)	RetCam Shuttle; acquisition: non-physician; reading: accredited non-physician technicians (real-time triage); remote expert reporting	BIO with scleral depression by ROP expert, masked	CS, any-stage	Unclear
Lin 2018 ⁴⁸	China (upper-middle)	Retrospective, all screened infants, single centre	600; GA 32.3 ± 1.9; BW 1810 ± 460	1200 eyes	RetCam III; acquisition: ophthalmologist; reading: fundus specialist; masking not reported	BIO by experienced ophthalmologist	Any-stage	Unclear
Prakalapakorn 2018 ⁴⁹	Thailand (upper-middle)	Prospective pilot, single centre	56; GA 28.1 (23–33); BW 1156 (585–2000)	50 infants (accuracy)	Pictor (45°, non-contact); acquisition: non-physician; reading: two remote paediatric ophthalmologists (grader 1 primary)	BIO by retina specialist	CS, plus	High
Wongwai 2018 ⁵⁰	Thailand (upper-middle)	Prospective, consecutive, multicentre	100; GA 31.1 ± 3.0; BW 1511 ± 421	100 infants	RetCam Shuttle; acquisition: non-physician; reading: two remote masked ophthalmologists (reader A primary)	BIO with scleral indentation (expert), masked	CS, any-stage, plus	Unclear
Tian 2019 ⁵¹	China (upper-middle)	Retrospective, remote-reading network, multicentre	1402; GA NR; BW NR	1402 infants	RetCam III / Nautilus; acquisition: non-ophthalmologist physician; reading: remote readers (Shenzhen reading centre)	BIO by trained physicians at partner units	Any-stage	High
Adhikari 2021 ⁵²	Nepal (low)	Cross-sectional validation, all new eligible infants	100; GA 32.6 ± 2.5; BW 1300 ± 200	200 eyes	Smartphone + Paxos Scope, 20 D lens; acquisition: non-physician; reading: two remote vitreoretinal specialists (grader 1 primary), masked	BIO (paediatric ophthalmologist)	CS	Low
Lanzelotte 2022 ⁵³	Brazil (upper-middle)	Cross-sectional convenience sample, 6 centres	200; GA 30 ± 2; BW 1265 ± 315	712 eye-examinations	Pictor Plus (40°, non-contact); acquisition: ophthalmologist; reading: two remote ophthalmologists (observer 3 primary)	BIO (ophthalmologist who acquired the images)	Plus	High
Young 2023 ⁵⁴	India (lower-middle)	Prospective paired comparison, telemedicine programme	156; GA 33.0 ± 3.0; BW 1685 ± 419	156 infants (AI test set 29)	Smartphone (MII RetCam or Keeler MIO, 28 D lens); acquisition: non-physician; reading: consensus of two masked telemedicine graders; ResNet-18 (AI)	Telemedicine grading of paired RetCam images	CS, any-stage	High
Artificial intelligence								
Wang 2018 ⁵⁵	China (upper-middle)	Retrospective development; temporally separate test set	264; GA NR; BW NR	552 eye-examinations	RetCam-type wide-field images; acquisition: NR; reading: deepROP (Id-Net/Gr-Net)	Consensus image labels of three experts	CS, any-stage	Unclear
Zhang 2019 ⁵⁶	China (upper-middle)	Retrospective, telemedicine platform; image-level hold-out	NR; GA 32.0 (25–36); BW 1500 (780–2000)	1742 images	RetCam 2/3; acquisition: non-physician; reading: VGG-16 deep neural network	Paediatric ophthalmologists' consensus labels	Any-stage	High
Wang 2021 ⁵⁷	China (upper-middle)	Retrospective, 4 centres; patient-level hold-out	1448; GA 32.9 ± 3.1; BW 1925 ± 774	1448 infants (test)	RetCam II/III; acquisition: NR; reading: J-PROP multidimensional platform	Consensus image labels (junior/senior/expert)	Any-stage, plus	High

Study	Country (income group)	Design and setting	Infants (n); GA; BW	Units analysed	Index test: device; image acquisition; reading	Reference standard	Targets	RoB
Cole 2022 ⁵⁸	Nepal; Mongolia (low; lower-middle)	Retrospective external validation, screening programmes	637; GA 33.3 ± 2.5; 30.4 ± 2.1; BW 1960 ± 557; 1515 ± 384	391 + 467 examinations	3nethra neo (Nepal); RetCam Portable (Mongolia); acquisition: NR; reading: i-ROP deep learning (US-trained, not retrained)	Clinical diagnosis (medical record); expert reference for plus	CS, plus	Unclear
Coyner 2022 ⁵⁹	India (lower-middle)	External validation of prognostic AI model, telemedicine programme	762; GA 33.5 ± 2.9 (untreated); BW 1734 ± 467 (untreated)	762 infants (test)	RetCam Shuttle; acquisition: non-physician; reading: vascular severity score + gestational age risk model	Subsequent treatment-requiring ROP (telemedicine clinical diagnosis)	CS	High
Li 2022 ⁶⁰	China (upper-middle)	Retrospective, single centre; separate test set	567; GA 30.4 ± 5.8; BW 1442 ± 517	3680 images	RetCam 3; acquisition: NR; reading: denseNet deep convolutional network	Consensus of three ROP experts on images	Any-stage	High
Peng 2022 ⁶¹	China (upper-middle)	Retrospective; examination-level hold-out	NR; GA NR; BW NR	311 examinations	RetCam 3; acquisition: NR; reading: ADS-Net attention network	Three paediatric ophthalmologists (concordant labels only)	Any-stage	High
Liu 2023 ⁶²	China (upper-middle)	Retrospective, single centre; patient-level hold-out	131; GA 30.9 ± 3.8; BW 1546 ± 710	131 infants (test)	RetCam II/III; acquisition: NR; reading: denseNet-121 / ResNet-18 system	Two senior ophthalmologists with expert arbitration (images)	Any-stage	High
Raja Sankari 2023 ⁶³	India (lower-middle)	Retrospective case-control image set; hold-out	200; GA 29.8–31.8; BW 1380–1700	50 images	3nethra neo; acquisition: non-physician; reading: quantum support vector machine	Certified ophthalmologists' image labels	Any-stage	High
Rao 2023 ⁶⁴	India (lower-middle)	Retrospective, KIDROP tele-ROP programme; image-level hold-out	NR; GA NR; BW NR	7489 images	RetCam III/Shuttle and 3nethra neo; acquisition: non-physician; reading: medios AI-ROP (EfficientNet-B0)	Trained non-physician tele-ROP graders' registers	Any-stage	High
Subramaniam 2023 ⁶⁵	Argentina; Colombia (upper-middle)	Retrospective convenience image set; hold-out	NR; GA NR; BW NR	50 images	Smartphone + 20/28 D lens; acquisition: physician; reading: googLeNet after image harmonisation	In-person BIO by expert physician	Plus	High
Coyner 2024 ⁶⁶	India (lower-middle)	Retrospective external validation, all programme examinations	2699; GA 33.5 ± 2.9; BW 1735 ± 450	5145 examinations	RetCam Shuttle or 3nethra; acquisition: non-physician; reading: autonomous i-ROP DL (vascular severity score ≥3.1)	Telemedicine clinical diagnosis (ROPE-SOS programme)	CS	Unclear
Agrawal 2025 ⁶⁷	India (lower-middle)	Curated class-balanced public dataset (HVDROPDB); hold-out	NR; GA NR; BW NR	39 images	RetCam and 3nethra neo; acquisition: NR; reading: multiCNN_LSTM classifier	Hospital ophthalmologists' image annotation	Plus	High

^aYoung 2023 evaluated both telemedicine grading and an artificial intelligence algorithm on paired images. Income group is the World Bank classification at the time of data collection. BIO, binocular indirect ophthalmoscopy; BW, birth weight in grams (mean ± SD or median/range as reported); CS, clinically significant ROP (treatment-requiring or referral-warranted); any-stage, ROP of any stage; plus, plus disease; GA, gestational age in weeks; NR, not reported; RoB, overall QUADAS-2 risk of bias; ROP, retinopathy of prematurity.

Telemedicine studies were mostly prospective paired comparisons with ophthalmoscopy, whereas AI studies were retrospective evaluations on temporally separate, external or hold-out image sets. Images were acquired with wide-field contact cameras in most studies and with portable non-contact cameras or smartphones in the others, frequently by nurses or technicians. Clinically significant ROP was reported in 12 datasets from 10 studies (8,273 units; 589 with the target condition), any-stage ROP in 16 datasets from 16 studies (19,272 units) and plus disease in 9 datasets from 8 studies (4,898 units). The unit of analysis was

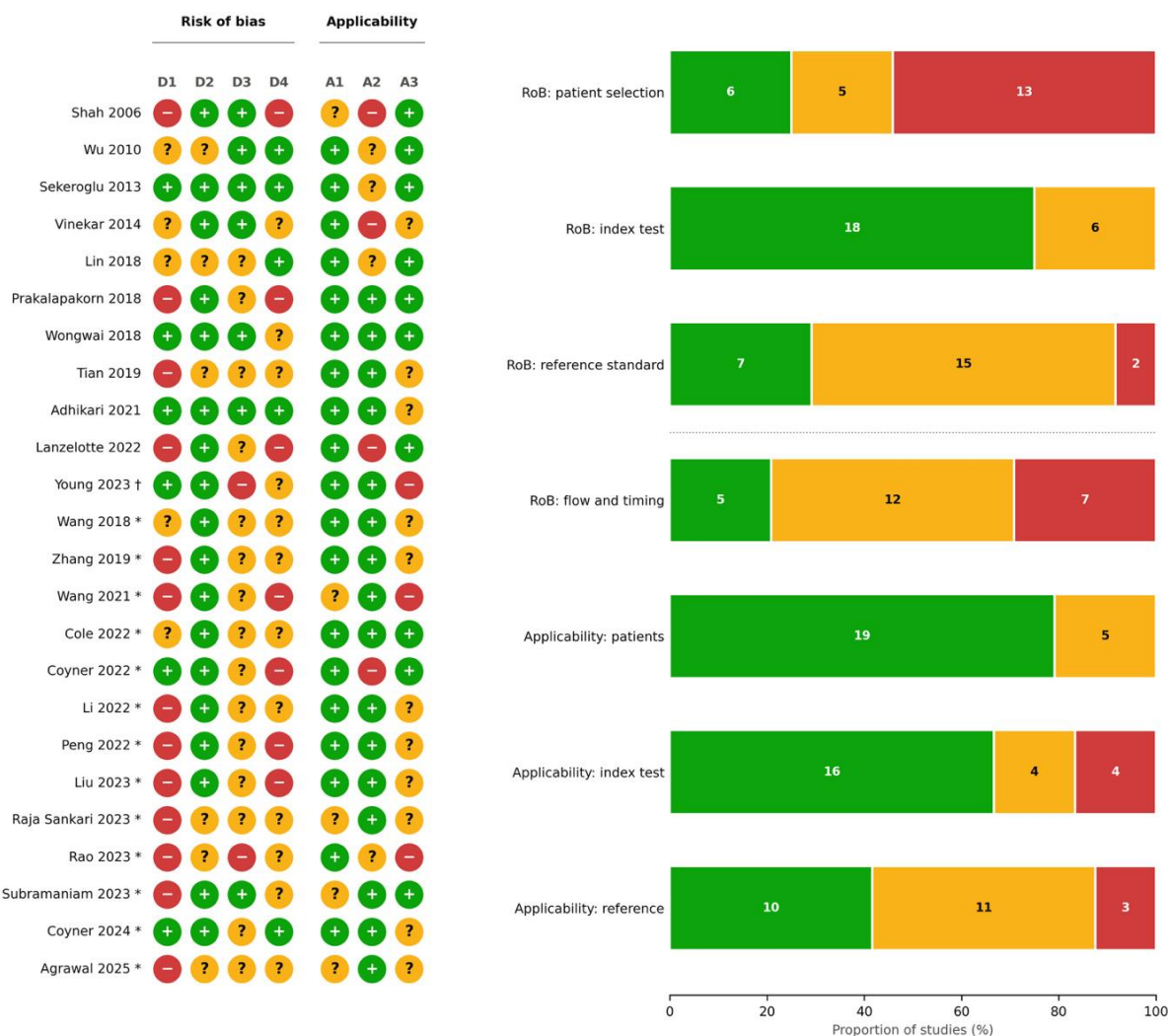
the infant in 5 of the 12 datasets for clinically significant ROP and the eye or examination in the remainder.

3.3. Risk of bias and applicability

Fifteen studies were at high overall risk of bias, seven at unclear risk and two at low risk, as illustrated in Figure 2. High risk arose mainly from patient selection in AI studies that used curated, case-control or class-balanced image sets and from flow and timing when not all units received the same reference standard. The reference standard domain was frequently unclear because image-based expert labels were used or masking was not reported. Applicability concerns for

the index test were high in four studies, in which images were acquired by ophthalmologists rather than within a typical remote programme, technicians graded

their own images in real time, or the AI model predicted future rather than current disease.



D1, patient selection; D2, index test; D3, reference standard; D4, flow and timing; A1-A3, applicability concerns for patient selection, index test and reference standard.
* Artificial intelligence study; † study evaluating both telemedicine grading and artificial intelligence; numbers in the bars are counts of studies.

Figure 2. Risk of bias and applicability concerns of the included diagnostic accuracy studies (QUADAS-2). Left, judgements for each study and domain, shown by colour and symbol (+, low; ?, unclear; -, high); right, proportion of studies with low, unclear and high risk or concern in each domain; numbers in the bars are counts of studies.

3.4. Accuracy for clinically significant ROP

Study-level sensitivity for clinically significant ROP ranged from 80.8–100.0% and specificity from 54.3–100.0%, as shown in Figure 3. Telemedicine had a pooled sensitivity of 92.8% (95% CI 82.6–97.3) and specificity of 95.8% (95% CI 87.6–98.7), whereas AI

had a sensitivity of 93.0% (95% CI 81.5–97.6) and a specificity of 77.3% (95% CI 55.6–90.3), as detailed in Table 2. The corresponding positive likelihood ratios were 22.3 (95% CI 7.4–70.8) and 4.1 (95% CI 2.1–9.1), and the negative likelihood ratios 0.07 (95% CI 0.03–0.19) and 0.09 (95% CI 0.04–0.22).

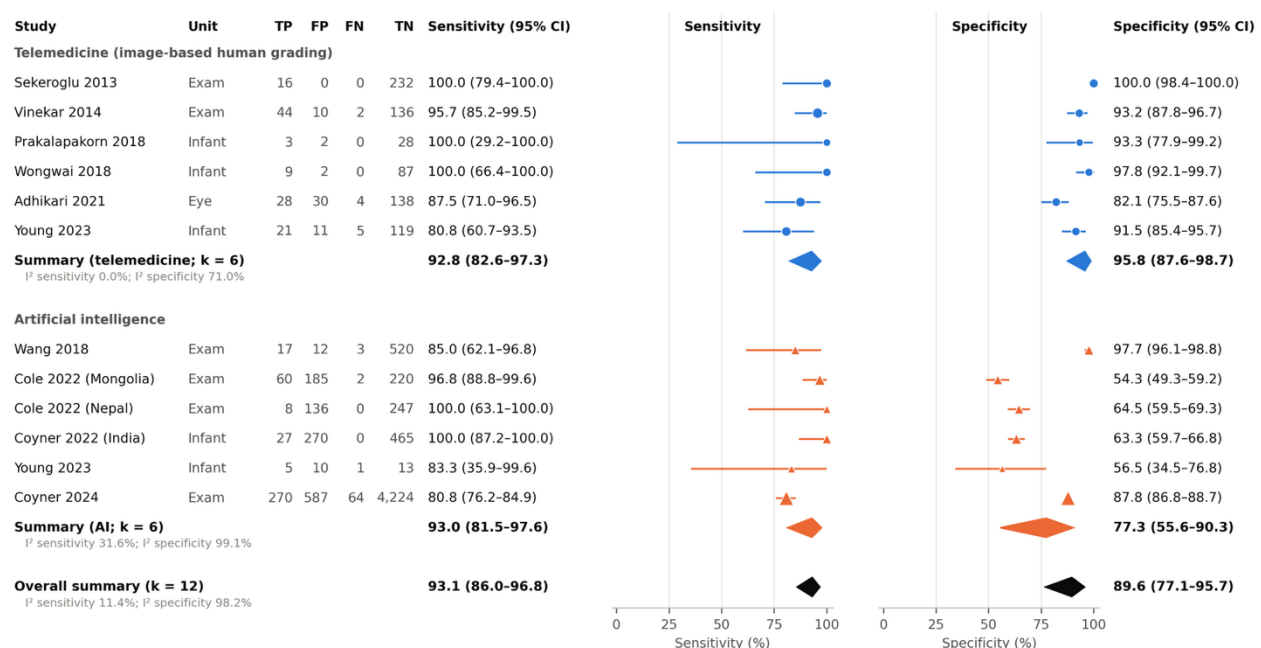


Figure 3. Forest plot of sensitivity and specificity for clinically significant retinopathy of prematurity, grouped by index test. Points show study estimates with exact binomial (Clopper-Pearson) 95% confidence intervals (marker size proportional to the number of units); diamonds show summary estimates from the bivariate random-effects model, and I^2 values come from univariate models. CI, confidence interval; FN, false negatives; FP, false positives; k, number of datasets; TN, true negatives; TP, true positives.

Table 2. Pooled diagnostic accuracy of telemedicine and artificial intelligence by target condition.

Target condition and index test	k	Units (positive)	Sensitivity, % (95% CI)	Specificity, % (95% CI)	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)	I^2 sensitivity/sp ecificity (%)
Clinically significant ROP								
All datasets	12	8,273 (589)	93.1 (86.0–96.8)	89.6 (77.1–95.7)	9.0 (4.0–21.4)	0.08 (0.04–0.16)	117 (34–394)	11.4/98.2
Telemedicine	6	927 (132)	92.8 (82.6–97.3)	95.8 (87.6–98.7)	22.3 (7.4–70.8)	0.07 (0.03–0.19)	298 (62–1410)	0.0/71.0
Artificial intelligence	6	7,346 (457)	93.0 (81.5–97.6)	77.3 (55.6–90.3)	4.1 (2.1–9.1)	0.09 (0.04–0.22)	45 (16–126)	31.6/99.1
Any-stage ROP								
All datasets	16	19,272 (4,424)	92.2 (85.0–96.1)	97.6 (95.5–98.8)	39.0 (20.4–73.8)	0.08 (0.04–0.15)	490 (195–1234)	93.2/92.2
Telemedicine	8	3,869 (768)	88.3 (68.6–96.3)	99.0 (95.7–99.8)	86.7 (20.4–360.1)	0.12 (0.04–0.32)	734 (119–4458)	93.3/78.8
Artificial intelligence	8	15,403 (3,656)	94.9 (90.9–97.2)	95.6 (93.7–97.0)	21.8 (14.9–31.8)	0.05 (0.03–0.10)	411 (175–963)	83.6/95.2
Plus disease								
All datasets	9	4,898 (306)	81.4 (68.8–89.6)	99.1 (96.8–99.7)	88.7 (25.5–307.6)	0.19 (0.10–0.31)	471 (124–1782)	72.2/88.9
Telemedicine	4	1,233 (98)	75.4 (45.2–91.9)	99.5 (96.9–99.9)	162.0 (22.0–1079.6)	0.25 (0.08–0.55)	655 (62–6839)	0.0/0.0
Artificial intelligence	5	3,665 (208)	89.4 (84.5–92.9)	98.1 (92.9–99.5)	47.0 (12.5–181.4)	0.11 (0.07–0.16)	436 (103–1845)	11.1/93.1

Bivariate binomial-normal random-effects estimates; confidence intervals for likelihood ratios and the diagnostic odds ratio were obtained from 20,000 random draws based on the model covariance matrix. I^2 values come from univariate random-effects logistic models for sensitivity and specificity. CI, confidence interval; DOR, diagnostic odds ratio; k, number of datasets; LR+ and LR-, positive and negative likelihood ratios; ROP, retinopathy of prematurity.

The difference in accuracy between index tests did not reach statistical significance ($p=0.08$). Across all 12 datasets, the pooled sensitivity was 93.1% (95% CI 86.0–96.8) and the pooled specificity 89.6% (95% CI 77.1–95.7), with a diagnostic odds ratio of 117 (95% CI

34–394) (Table 2). The between-study variance on the logit scale was 0.67 for sensitivity and 2.55 for specificity, and the univariate I^2 was 11.4% and 98.2%, respectively. In the receiver operating characteristic space, the summary points of both index tests lay in the

upper left corner, but the confidence region of AI extended further to the right and the prediction region

of the pooled estimate was wide, as illustrated in Figure 4.

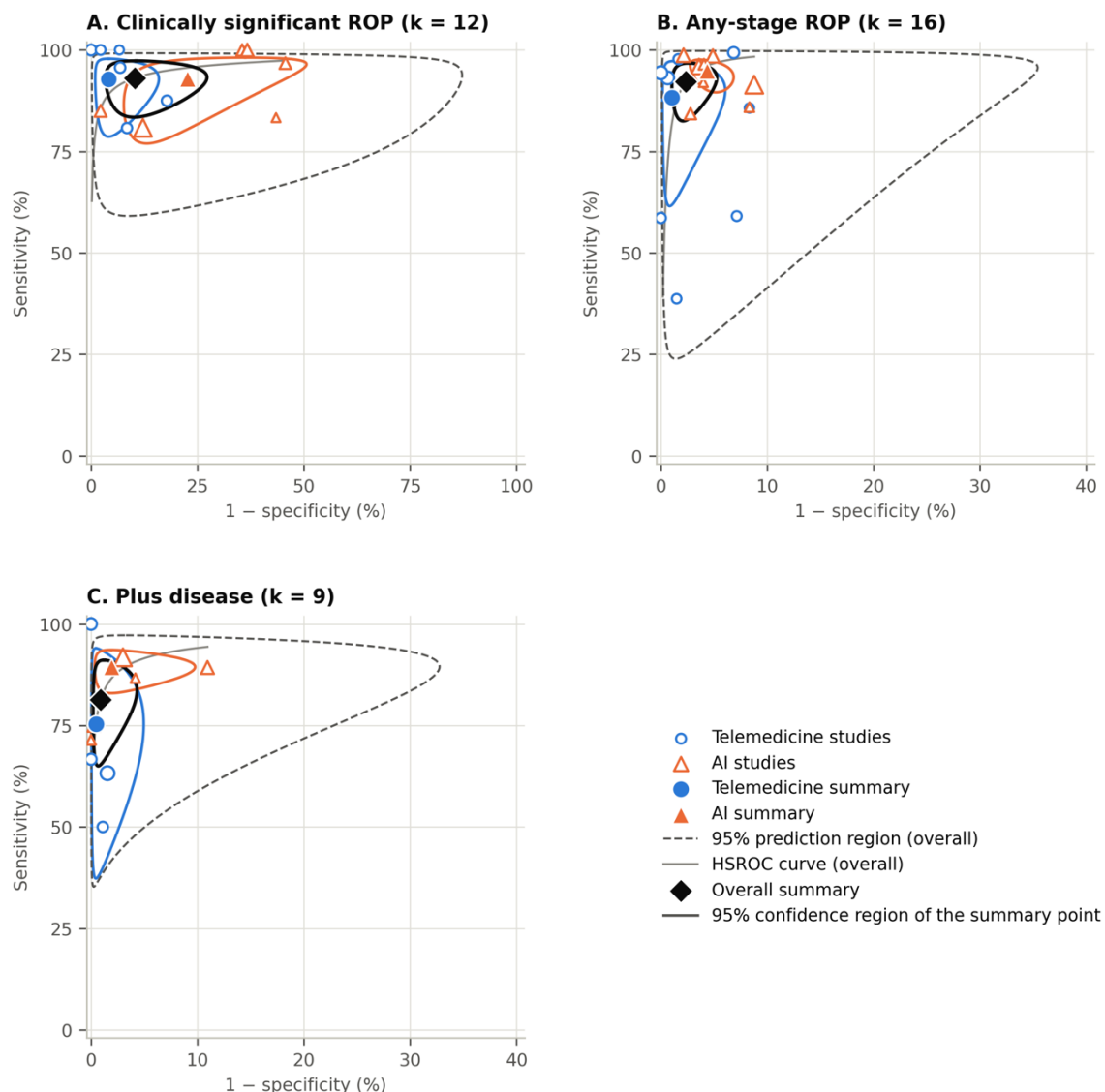


Figure 4. Summary receiver operating characteristic plots for (A) clinically significant retinopathy of prematurity, (B) any-stage retinopathy of prematurity and (C) plus disease. Open symbols are individual datasets (size proportional to the number of units); filled symbols are summary points for telemedicine (circles), artificial intelligence (triangles) and all datasets (diamond), with their 95% confidence regions. The dashed line is the 95% prediction region and the grey line the hierarchical summary receiver operating characteristic curve for all datasets. The horizontal axis of panels B and C is truncated at 40%.

3.5. Accuracy for any-stage ROP and plus disease

For any-stage ROP, telemedicine had a sensitivity of 88.3% (95% CI 68.6–96.3) and a specificity of 99.0% (95% CI 95.7–99.8); AI had a sensitivity of 94.9% (95% CI 90.9–97.2) and a specificity of 95.6% (95% CI 93.7–97.0) ($p=0.08$). Across all datasets, the pooled sensitivity was 92.2% (95% CI 85.0–96.1) and the pooled specificity 97.6% (95% CI 95.5–98.8) (Table 2). Heterogeneity was high for both sensitivity ($I^2=93.2\%$; between-study variance 2.06) and specificity ($I^2=92.2\%$; 1.51).

For plus disease, the sensitivity of telemedicine was 75.4% (95% CI 45.2–91.9) and that of AI 89.4% (95% CI 84.5–92.9), with specificities of 99.5% and 98.1% ($p=0.11$). Across all datasets, the pooled sensitivity was 81.4% (95% CI 68.8–89.6) and the pooled specificity 99.1% (95% CI 96.8–99.7) (Table 2). The summary points for both target conditions lay close to the vertical axis of the receiver operating characteristic space, reflecting high specificity (Figure 4).

3.6. Sources of heterogeneity

In meta-regression for clinically significant ROP, specificity was lower in low- or lower-middle-income

settings (77.9% (95% CI 66.0–86.5)) than in upper-middle-income settings (98.3% (95% CI 95.2–99.4); $p<0.001$), as detailed in Table 3. The threshold was also associated with accuracy ($p<0.001$): studies using a referral-warranted threshold had a sensitivity of 81.4% (95% CI 77.3–84.9) and a specificity of 82.7% (95% CI

50.4–95.8), compared with 96.3% (95% CI 92.5–98.2) and 92.2% (95% CI 79.2–97.3) for the treatment-requiring threshold. The reference standard, imaging device, person acquiring the images, unit of analysis and risk of bias were not associated with accuracy for this target condition.

Table 3. Meta-regression and sensitivity analyses by target condition.

Analysis	k	CS sensitivity	CS specificity	k	Any-stage sensitivity	Any-stage specificity	k	Plus sensitivity	Plus specificity
Primary analysis									
All datasets	12	93.1 (86.0–96.8)	89.6 (77.1–95.7)	16	92.2 (85.0–96.1)	97.6 (95.5–98.8)	9	81.4 (68.8–89.6)	99.1 (96.8–99.7)
Meta-regression (subgroup estimates)									
Index test		$p=0.08$			$p=0.08$		$p=0.11$		
Telemedicine	6	93.5 (82.8–97.7)	95.8 (87.9–98.6)	8	87.8 (73.0–95.1)	98.7 (96.9–99.5)	4	71.9 (54.4–84.5)	99.6 (97.7–99.9)
Artificial intelligence	6	92.7 (82.4–97.2)	77.3 (54.9–90.5)	8	95.1 (87.8–98.1)	95.8 (91.3–98.0)	5	88.1 (78.4–93.8)	98.0 (92.6–99.5)
Income group		$p<0.001$			$p=0.17$		$p=0.43$		
Low or lower-middle income	8	92.7 (84.8–96.6)	77.9 (66.0–86.5)	6	93.1 (79.9–97.9)	94.8 (87.7–97.9)	3	86.0 (66.0–95.1)	97.7 (87.5–99.6)
Upper-middle income	4	95.6 (81.8–99.0)	98.3 (95.2–99.4)	10	91.6 (81.4–96.5)	98.4 (96.7–99.2)	6	79.0 (62.6–89.4)	99.4 (97.3–99.9)
Threshold for clinically significant ROP		$p<0.001$			–		–		
Treatment-requiring	8	96.3 (92.5–98.2)	92.2 (79.2–97.3)	–	–	–	–	–	–
Referral-warranted	4	81.4 (77.3–84.9)	82.7 (50.4–95.8)	–	–	–	–	–	–
Reference standard		$p=0.22$			$p=0.03$		$p=0.21$		
Ophthalmoscopy or clinical diagnosis	9	95.1 (87.8–98.1)	89.6 (74.1–96.3)	7	90.4 (75.5–96.6)	99.1 (97.7–99.6)	7	76.9 (62.5–87.0)	99.4 (97.2–99.9)
Expert image grading	3	83.8 (60.7–94.5)	89.9 (58.1–98.3)	9	93.4 (84.3–97.4)	95.6 (91.9–97.7)	2	90.9 (76.7–96.8)	97.0 (78.6–99.7)
Imaging device		$p=0.28$			–		$p=0.03$		
Wide-field contact camera	8	94.7 (87.3–97.9)	91.3 (77.2–97.0)	–	–	–	6	89.3 (84.6–92.7)	98.7 (95.1–99.6)
Portable non-contact or smartphone	4	86.8 (67.9–95.4)	85.0 (53.3–96.5)	–	–	–	3	64.2 (53.2–73.9)	99.6 (96.5–100.0)
Image acquisition		$p=0.43$			$p=0.98$		$p=0.16$		
Non-physician staff	8	90.7 (81.3–95.6)	87.8 (69.8–95.7)	7	91.7 (78.2–97.1)	97.7 (93.9–99.1)	2	58.1 (24.9–85.3)	99.8 (95.6–100.0)
Physician or not reported	4	95.7 (86.8–98.7)	92.6 (70.8–98.5)	9	92.6 (82.5–97.1)	97.6 (94.4–99.0)	7	85.3 (74.2–92.1)	98.7 (95.6–99.6)
Unit of analysis		$p=0.90$			$p=0.97$		–		
Infant	5	93.9 (80.0–98.4)	87.3 (61.5–96.7)	4	90.9 (70.0–97.7)	97.7 (92.2–99.4)	–	–	–
Eye, examination or image	7	92.8 (84.0–96.9)	91.0 (75.0–97.2)	12	92.6 (84.2–96.7)	97.6 (95.0–98.9)	–	–	–
Overall risk of bias		$p=0.57$			$p=0.34$		$p=0.88$		
Low or unclear	8	93.3 (85.0–97.2)	92.3 (79.9–97.3)	6	89.1 (71.4–96.4)	98.6 (96.1–99.5)	4	84.0 (64.7–93.7)	99.1 (95.1–99.9)
High	4	92.5 (75.7–98.0)	81.2 (47.8–95.3)	10	93.6 (85.5–97.4)	96.7 (93.1–98.5)	5	79.4 (62.0–90.1)	99.0 (95.1–99.8)
Sensitivity analyses									
Excluding studies at high risk of bias	8	93.4 (84.9–97.2)	92.8 (77.5–98.0)	6	89.6 (64.7–97.6)	98.7 (96.5–99.5)	4	86.1 (59.5–96.3)	99.4 (92.7–99.9)
Ophthalmoscopy or clinical reference standard only	9	96.2 (87.0–99.0)	89.8 (73.2–96.6)	7	90.7 (71.7–97.4)	99.3 (96.6–99.9)	7	78.0 (61.3–88.8)	99.5 (96.8–99.9)

Analysis	<i>k</i>	CS sensitivity	CS specificity	<i>k</i>	Any-stage sensitivity	Any-stage specificity	<i>k</i>	Plus sensitivity	Plus specificity
Infant-level data only	5	95.4 (67.9–99.5)	86.9 (66.8–95.6)	4	91.0 (71.4–97.6)	98.3 (83.0–99.9)	–	–	–
Excluding image-level data	–	–	–	12	90.6 (79.4–96.1)	98.1 (95.4–99.2)	6	79.0 (59.8–90.5)	99.4 (96.3–99.9)
Second reader where available	12	93.2 (86.2–96.8)	87.6 (74.4–94.5)	16	92.3 (85.2–96.1)	97.5 (95.4–98.7)	9	83.6 (73.0–90.6)	98.5 (95.6–99.5)
Treatment threshold for all studies	12	96.7 (91.9–98.7)	88.6 (75.3–95.2)	–	–	–	–	–	–
Excluding the prognostic model (Coyner 2022)	11	91.8 (84.4–95.8)	91.1 (79.0–96.6)	–	–	–	–	–	–
Excluding case-control or class-balanced designs	–	–	–	15	92.2 (84.4–96.3)	97.7 (95.4–98.8)	–	–	–
Excluding studies with unreported masking of readers	–	–	–	14	91.5 (82.7–96.1)	97.3 (94.6–98.7)	–	–	–
Leave-one-out analysis (range)	11	91.8–94.3	85.9–91.3	15	90.9–93.5	96.9–97.9	8	77.3–84.2	98.7–99.3

Values are sensitivity and specificity in % (95% CI) from bivariate random-effects models. Each covariate row gives the *p* value for that covariate under the sensitivity column of each target condition, and the indented rows below it give the estimates for each level. Subgroup estimates come from bivariate meta-regression with the covariate as a fixed effect on logit sensitivity and logit specificity; *p* values are likelihood-ratio tests of the covariate (2 degrees of freedom). A dash indicates that the covariate or scenario did not apply or that all datasets fell into one level (the imaging-device covariate for any-stage ROP; the treatment threshold applies only to clinically significant ROP). In the scenarios that left the dataset unchanged for a target condition, no estimate is shown. CS, clinically significant ROP; *k*, number of datasets; ROP, retinopathy of prematurity.

Both associations remained below the Bonferroni threshold of 0.0063. For any-stage ROP, specificity was higher against ophthalmoscopy or a clinical reference standard (99.1% (95% CI 97.7–99.6)) than against expert image grading (95.6% (95% CI 91.9–97.7); *p*=0.03). For plus disease, sensitivity was lower with portable non-contact or smartphone cameras (64.2% (95% CI 53.2–73.9)) than with wide-field contact cameras (89.3% (95% CI 84.6–92.7); *p*=0.03) (Table 3). These two associations did not remain significant after correction for multiple testing and were based on small subgroups. The model for the threshold covariate required the uncorrelated form of the bivariate model.

3.7. Sensitivity analyses and small-study effects

The pooled estimates for clinically significant ROP were stable in sensitivity analyses (Table 3). After exclusion

of studies at high risk of bias, sensitivity was 93.4% (95% CI 84.9–97.2) and specificity 92.8% (95% CI 77.5–98.0); after exclusion of the prognostic model, 91.8% (95% CI 84.4–95.8) and 91.1% (95% CI 79.0–96.6); and when the treatment threshold was applied to all studies, 96.7% (95% CI 91.9–98.7) and 88.6% (95% CI 75.3–95.2). Omitting one dataset at a time gave sensitivities of 91.8–94.3% and specificities of 85.9–91.3%. When the analysis was restricted to infant-level data, sensitivity was 95.4% (95% CI 67.9–99.5) and specificity 86.9% (95% CI 66.8–95.6).

Deeks' funnel plots showed no significant asymmetry for clinically significant ROP (*p*=0.11) or any-stage ROP (*p*=0.15), as shown in Figure 5. The test was not performed for plus disease because fewer than ten datasets were available.

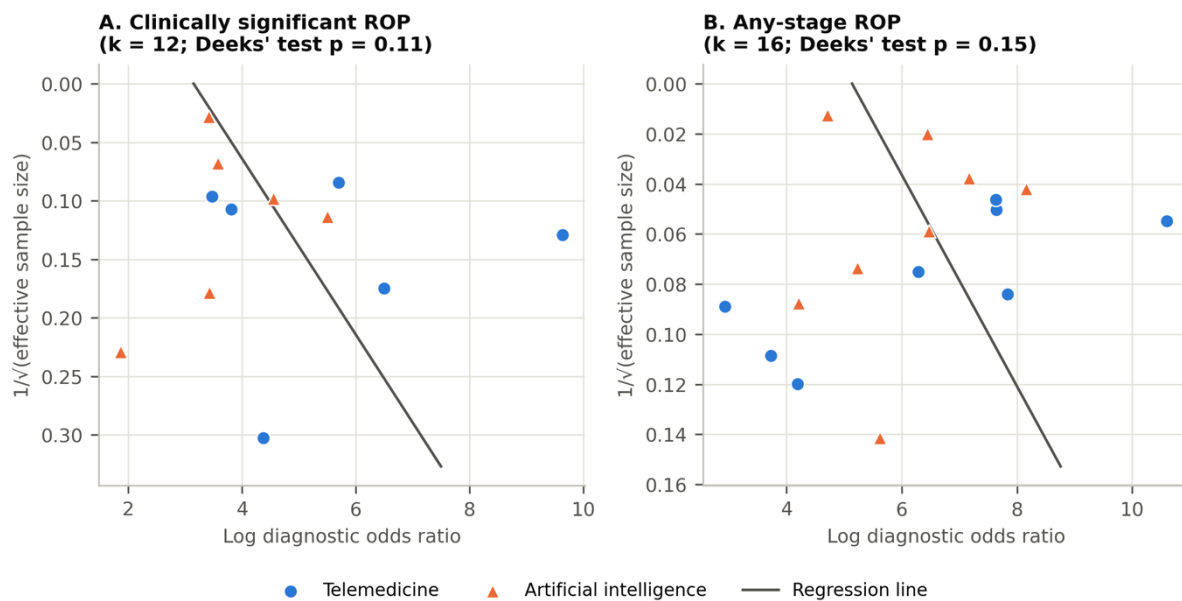


Figure 5. Deeks' funnel plots for (A) clinically significant and (B) any-stage retinopathy of prematurity. The log diagnostic odds ratio of each dataset is plotted against the inverse square root of its effective sample size; the line is the regression line of Deeks' test. k , number of datasets.

3.8. Economic evaluations

The 13 economic evaluations came from Iran, Brazil, Thailand, Ghana, Mexico, the Philippines, Rwanda, India and Guatemala, as detailed in Table 4. They fell into three groups: five cost-utility analyses that

compared telemedicine with ophthalmoscopy-based or no screening, three programme-level models that evaluated screening and treatment regardless of the screening modality, and five costing studies, three of which had no comparator.

Table 4. Economic evaluations of retinopathy of prematurity screening in low- and middle-income countries.

Study	Country (income group)	Design	Strategies compared	Main result	Source
Alizadeh 2026 ⁶⁸	Iran (upper-middle)	CUA; Markov model; societal; lifetime; 5.8%	Nurse-led RetCam telemedicine vs direct examination by a local ophthalmologist vs universal referral to a tertiary centre	Telemedicine vs direct examination US\$288.81/QALY; universal referral dominated; threshold US\$7,047/QALY; telemedicine cost-effective in 95.2% of iterations of the probabilistic sensitivity analysis	Full text
Mohammadi 2021 ⁶⁹	Iran (upper-middle)	CUA; decision tree; societal; lifetime; 5.8% (QALYs)	National tele-ROP network (nurse imaging, central reading centre) vs no organised screening	US\$161.4/QALY; threshold US\$4,616/QALY; one-way range US\$98.59–245.6/QALY	Full text
Neves 2021 ⁷⁰	Brazil (upper-middle)	CUA; decision tree; public payer; lifetime; 5%	Wide-field imaging by nurse technicians with remote expert reading plus BIO vs BIO alone	US\$1,746.99/QALY; cost-effective in all iterations of the probabilistic sensitivity analysis at about US\$1,800/QALY	Full text
Wongwai 2015 ⁷¹	Thailand (upper-middle)	CUA; decision tree; provider and societal; lifetime; 3%	Store-and-forward RetCam telemedicine vs indirect ophthalmoscopy by general ophthalmologists	THB 17,396/QALY; THB 259 per case of blindness prevented; ICER fell to zero at ≥650 infants screened per year	Full text
Shah 2026 ⁷²	Ghana (lower-middle)	CUA; decision-analytic Markov model; health care sector; 85 years; 3%	No screening vs BIO vs telemedicine with non-physician graders vs hybrid telemedicine (10% physician grading)	Non-physician telemedicine vs no screening US\$27–73/QALY; hybrid vs non-physician telemedicine US\$1,331–3,566/QALY; BIO dominated; threshold US\$2,238/QALY	Full text
Rothschild 2016 ⁷³	Mexico (upper-middle)	CEA, CUA and CBA (EcROP); societal; lifetime; 3%	Ideal national screening and treatment programme vs current coverage (52%)	Cost-saving; incremental cost –US\$2,203 and +0.067 QALY per child; net monetary benefit US\$5,556 per child	Full text
Naguib 2019 ⁷⁴	Philippines (lower-middle)	CEA, CUA and CBA (EcROP); societal; NR; NR	Ideal national screening and treatment programme vs current practice	Net monetary benefit US\$2,627 per child and US\$64.3 million per year nationally	Abstract

Study	Country (income group)	Design	Strategies compared	Main result	Source
Alder 2025 ⁷⁵	Rwanda (low)	Decision-analytic model; societal; NR; NR	National screening with anti-VEGF treatment vs status quo	US\$738 per infant; 257 cases of blindness averted per birth cohort; cost-saving (about US\$270,000)	Abstract
Zin 2014 ⁷⁶	Brazil (upper-middle)	Cost and budget-impact analysis; public payer; 1 year; NA	BIO screening with shared laser treatment at 52%, 80% and 100% coverage	US\$18 per infant examined; US\$80 per at-risk infant; national budget US\$0.56–1.07 million per year	Full text
Kelkar 2017 ⁷⁷	India (lower-middle)	Programme cost analysis; provider; 3 years; none	Mobile BIO screening and laser team serving five neonatal units (no comparator)	US\$198.9 per infant screened; US\$4,137.4 per case of visual disability addressed	Full text
Kelkar 2017b ⁷⁸	India (lower-middle)	Programme cost analysis; provider; 3 years; none	Mobile BIO screening van serving five neonatal units (no comparator)	US\$310 per infant screened; US\$6,500 per case of visual disability addressed	Full text
Padhi 2024 ⁷⁹	India (lower-middle)	Cost-benefit description; programme; 9 months; none	Real-time videoconference telescreening by allied ophthalmic personnel (no costed comparator)	Programme cost US\$30,125; net benefit about US\$1.34 million; 477 specialist hours saved	Full text
Del Busto Wilhelm 2023 ⁸⁰	Guatemala (upper-middle)	Comparative cost analysis; NR; lifetime; NR	Screening and timely treatment vs lifetime management of severe visual impairment	US\$1,223 vs US\$214,666 per child (0.57%)	Abstract

Design column: type of evaluation; model; perspective; time horizon; annual discount rate. Results are shown in the currency reported by each study (price years: Alizadeh 2026 and Shah 2026, 2025 US dollars; Neves 2021, 2019 prices; Wongwai 2015, 2013 prices; Zin 2014, 2012 US dollars; not reported in the others). BIO, binocular indirect ophthalmoscopy; CBA, cost-benefit analysis; CEA, cost-effectiveness analysis; CUA, cost-utility analysis; EcROP, Economic Model of Retinopathy of Prematurity; ICER, incremental cost-effectiveness ratio; NA, not applicable; NR, not reported; QALY, quality-adjusted life year; THB, Thai baht; VEGF, vascular endothelial growth factor.

In the cost-utility analyses, nurse-led RetCam telemedicine cost US\$288.81 per quality-adjusted life year (QALY) gained compared with direct examination in Iran (2025 US dollars), and a national tele-ROP network cost US\$161.4 per QALY compared with no organised screening.^{68,69} Wide-field imaging with remote reading combined with ophthalmoscopy cost US\$1,746.99 per QALY in Brazil (2019 prices converted to US dollars), store-and-forward telemedicine cost 17,396 Thai baht (THB) per QALY in Thailand (2013 prices), and telemedicine with non-physician graders cost US\$27–73 per QALY compared with no screening in Ghana (2025 US dollars), where ophthalmoscopy was dominated.^{70–72} Each ratio was below the willingness-to-pay threshold used in its own study.

Programme-level models from Mexico, the Philippines and Rwanda found that expanding screening and treatment was cost-saving or produced a large net monetary benefit, ranging from US\$2,627 to US\$5,556 per child.^{73–75} Costing studies from Brazil, India and Guatemala reported costs from US\$18 per infant examined to US\$310 per infant screened, far below the lifetime cost of managing severe visual impairment.^{76–80} No economic evaluation of AI-based ROP screening from a low- or middle-income country was identified.

Against the CHEERS 2022 items, the perspective was stated or clearly implied in nine evaluations and not stated in four costing studies. Costs and outcomes were discounted in five of the nine evaluations with a lifetime horizon and outcomes only in one, and discounting was not reported in three. Deterministic or probabilistic

sensitivity analyses were reported in eight evaluations, including probabilistic analyses in at least five. Three evaluations were available only as abstracts.

The models used accuracy inputs that were at least as favourable as the pooled estimates of this review. The Iranian and Thai analyses assumed a sensitivity of 100% and a specificity of 97.8% for RetCam telemedicine, and the Ghanaian analysis assumed a sensitivity of 99% and a specificity of 95% for non-physician graders; in that model, ophthalmoscopy became the preferred strategy when grader sensitivity fell below 94%.^{68,71,72} The Iranian national model used accuracy data from a study conducted in the United States, and the Brazilian result was driven mainly by the higher coverage of the telemedicine strategy (90% versus 52%) rather than by accuracy.^{69,70}

3.9. Certainty of the evidence

At a pre-test probability of 10%, telemedicine would detect 93 of 100 infants with clinically significant ROP per 1000 screened, miss 7 (3–17) and refer 37 (12–112) without the condition, as detailed in Table 5. AI would miss 7 (2–19) but refer 204 (87–400) infants unnecessarily. The certainty of the evidence was low for the sensitivity of both index tests and for the specificity of telemedicine, and very low for the specificity of AI. Put differently, about 72% of infants with a positive telemedicine grade and about 31% of those with a positive AI result would have clinically significant ROP at this pre-test probability.

Table 5. Summary of findings for clinically significant retinopathy of prematurity per 1000 infants screened.

Outcome	Prevalence 5%	Prevalence 10%	Prevalence 20%	Certainty (GRADE)	Reasons for downgrading
Telemedicine: sensitivity 92.8%, specificity 95.8% (6 datasets; 927 units)					
True positives	46 (41–49)	93 (83–97)	186 (165–195)	Low	Risk of bias (2 of 6 datasets at high risk); indirectness (mixed units and thresholds, small samples)
False negatives	4 (1–9)	7 (3–17)	14 (5–35)		
True negatives	910 (832–938)	863 (788–888)	767 (701–789)	Low	Risk of bias; inconsistency ($I^2=71.0\%$)
False positives	40 (12–118)	37 (12–112)	33 (11–99)		
Artificial intelligence: sensitivity 93.0%, specificity 77.3% (6 datasets; 7,346 units)					
True positives	46 (41–49)	93 (81–98)	186 (163–195)	Low	Risk of bias (2 of 6 datasets at high risk); indirectness (retrospective external validation, one prognostic model)
False negatives	4 (1–9)	7 (2–19)	14 (5–37)		
True negatives	735 (528–858)	696 (500–813)	619 (445–722)	Very low	Risk of bias; very serious inconsistency ($I^2=99.1\%$); imprecision
False positives	215 (92–422)	204 (87–400)	181 (78–355)		

Numbers per 1000 infants screened (95% CI) at pre-test probabilities of clinically significant ROP of 5%, 10% and 20%; the median prevalence in the included datasets was 9.1% (range 2.0–24.0%). False negatives are infants with clinically significant ROP who would be missed; false positives are infants referred unnecessarily. Certainty started at high for cross-sectional accuracy studies and was downgraded by one level for each serious concern. CI, confidence interval; ROP, retinopathy of prematurity.

4. Discussion

4.1. Principal findings

This meta-analysis of twenty-four studies from nine low- and middle-income countries found that telemedicine grading detected clinically significant ROP with a sensitivity of 92.8% and a specificity of 95.8%, and that AI reached a similar sensitivity (93.0%) with a lower and more variable specificity (77.3%). Accuracy for any-stage ROP was high, whereas the sensitivity for plus disease was lower (81.4%). Telemedicine was cost-effective in every cost-utility analysis conducted in these settings, but the certainty of the accuracy evidence was low to very low, mainly because of risk of bias and inconsistency.

4.2. Telemedicine

The sensitivity of telemedicine for clinically significant ROP was similar to that reported by large programmes in which technicians or nurses captured images and experts graded them remotely.^{20,21} In the KIDROP programme, the few false negatives were infants whom technicians had scheduled for early review rather than immediate referral, and in Nepal ungradable smartphone images were counted as screen positive, which protected sensitivity at the cost of specificity.^{20,52} The specificity of telemedicine was higher than that of AI. With a positive likelihood ratio of 22.3, a positive grade raised a pre-test probability of 10% to a post-test probability of about 72%. This is relevant for neonatal units in countries such as Indonesia, where the number of paediatric ophthalmologists trained in ROP is limited

and every unnecessary transfer of a preterm infant carries risk and cost.^{5,6}

The type of camera did not explain heterogeneity for clinically significant ROP ($p=0.28$), although the sensitivity with portable non-contact or smartphone cameras (86.8% (95% CI 67.9–95.4)) was lower than with wide-field contact cameras (94.7% (95% CI 87.3–97.9)). Portable devices are cheaper and easier to move between small neonatal units, and a target product profile has defined minimum requirements for low-cost imaging devices in resource-limited settings.³⁴ The present estimates suggest that such devices should be validated against ophthalmoscopy at the treatment threshold before they replace wide-field cameras.

A telemedicine sensitivity close to 93% means that some infants with clinically significant disease would not be identified at a single session. In practice, ROP screening is repeated until the retina is mature, so an infant who is missed at one session may be detected at the next, provided that follow-up is completed. Programmes should therefore combine telemedicine with reliable recall systems and with ophthalmoscopy for infants whose images are ungradable or whose disease is close to the treatment threshold.

4.3. Artificial intelligence

The lower specificity of AI appears to reflect, at least in part, the operating points chosen by the developers; the review did not synthesise areas under the curve, so this interpretation could not be tested directly. The i-ROP deep learning system was applied without retraining to images from Nepal and Mongolia and, at the threshold set in the United States, had a specificity of 64.5% and

54.3%, respectively.⁵⁸ The updated system, which was retrained, given an autonomous threshold and a preprocessing pipeline for different fields of view and image qualities, reached a specificity of 87.8% in an Indian telemedicine programme.⁶⁶ The prognostic model tested in India also had a low specificity by design, because it aimed to identify all infants who would later need treatment.⁵⁹ Differences in pigmentation, camera type and field of view between training and deployment populations are known to affect image-based algorithms, and a vascular severity score derived from vessel maps was proposed to reduce this dependence.^{28,30,31}

AI may also have a role beyond classification. Vascular severity scores derived by AI were used to monitor the epidemiology of ROP across a large Indian telemedicine programme and to compare disease phenotypes between India and the United States, which suggests uses in the surveillance and quality assurance of screening programmes that do not depend on a single referral threshold.^{27,28}

These findings suggest that AI could serve as a first reader in a telemedicine workflow, removing examinations with a very low probability of disease from the grading queue while human graders review the rest. However, none of the included studies evaluated this workflow prospectively, and the AI data came from retrospective and sometimes image-level datasets. Local calibration, that is, choosing the algorithm's threshold on images from the programme in which it will be used so that sensitivity is preserved at an acceptable referral rate, and prospective evaluation of the complete workflow are therefore required before AI is adopted.

4.4. Heterogeneity, setting and plus disease

Income group was the covariate most strongly associated with specificity. This finding should be interpreted cautiously because the index test and income group were partly confounded: most AI datasets for clinically significant ROP came from low- or lower-middle-income countries, whereas the telemedicine studies with the highest specificity came from Turkey and Thailand. The threshold also explained heterogeneity, as expected, because referral-warranted disease includes milder stages that are harder to distinguish from normal findings. After correction for multiple testing, only the income group and the threshold remained associated with accuracy.

The income-group finding also bears on the definition of the setting. Specificity was 98.3% (95% CI 95.2–99.4) in upper-middle-income settings but 77.9% (95% CI 66.0–86.5) in low- or lower-middle-income settings, and only the Nepalese data came from a low-income

country. The pooled estimates therefore describe mainly middle-income settings with established neonatal units, and they should be applied to low-income settings with caution.

The sensitivity of 81.4% for plus disease has clinical consequences, because plus disease is a key criterion for treatment-requiring ROP and its detection determines the timing of treatment.¹⁷ About one in five units with plus disease would not be identified by a single grading, and the sensitivity was only 64.2% (95% CI 53.2–73.9) with portable or smartphone cameras. The smaller field of view of these devices captures fewer vessels for the assessment of dilatation and tortuosity, and agreement between graders on plus features depends on image characteristics.^{31,32} Programmes that rely on portable cameras should therefore refer infants with pre-plus or borderline findings for ophthalmoscopy and use wide-field images where plus disease is suspected.

No Indonesian accuracy or economic study was identified. Several features of the Indonesian system affect the transfer of these results: screening includes infants born before 34 weeks, the recorded incidence of ROP is probably underestimated because of the high mortality of very preterm infants, trained paediatric ophthalmologists are few, and national health insurance has made screening and treatment accessible to more infants since 2014.^{5–7} Aggressive ROP has been treated with intravitreal bevacizumab at a national referral hospital, which underlines the importance of early detection in district neonatal units.¹⁶ These conditions resemble those of the Thai and Indian programmes included in this review more than those of high-income countries, but local validation of any imaging and grading pathway would still be needed.

4.5. Economic evidence

The economic evidence favoured telemedicine where it was evaluated. The five cost-utility analyses reported ratios well below their national thresholds, and in Ghana ophthalmoscopy was dominated by telemedicine strategies.⁷² The programme-level models evaluated screening and treatment regardless of modality; they showed that the benefit of screening came largely from averted lifelong blindness and lost productivity, so that national programmes were cost-saving from a societal perspective.^{73,75} These models support investment in screening coverage but do not by themselves favour telemedicine over ophthalmoscopy.

The accuracy inputs of the telemedicine models were more favourable than the pooled estimates. The pooled telemedicine sensitivity found here (92.8% (95% CI 82.6–97.3)), which mostly reflected grading by ophthalmologists, lies close to the 94% threshold

below which ophthalmoscopy became preferable in the Ghanaian model.⁷² The economic conclusions are therefore sensitive to grader performance and should be re-examined with the pooled accuracy estimates and with the costs of training and quality assurance. No evaluation of AI from these settings was found; the United States model and the diabetic retinopathy literature indicate that the cost-effectiveness of AI depends on its specificity, because false positives generate referrals and examinations.^{35,38} The specificity of AI estimated here should be used as an input when AI is modelled for low- and middle-income countries.

4.6. Implications for programmes, training and communities

The findings have practical implications for the organisation of screening. For community and neonatal health services, a tiered model, in which trained nurses or technicians acquire images in each neonatal unit, an algorithm or a trained grader triages them, and remote experts review positive and ungradable examinations, would concentrate the scarce ophthalmologist time on the infants who need examination or treatment. Such a model requires quality assurance of image acquisition, regular audit of false negatives and a guaranteed pathway to treatment, and the target product profile for low-cost imaging devices provides a framework for selecting cameras.^{20,34} In countries with national health insurance, such as Indonesia, the cost per QALY reported in neighbouring upper-middle-income settings suggests that financing telemedicine screening is likely to represent good value.^{5,71}

Training and certification of the non-physician workforce are central to this model. In the KIDROP programme, technicians who were accredited after structured training triaged infants in real time, and programmes in Guatemala and rural Tamil Nadu and Andhra Pradesh relied on trained technicians to reach district and rural neonatal units.^{20–22} Surveys of neonatal nurses identified gaps in knowledge of ROP and practical challenges in telemedicine-based screening, and surveys of ophthalmology fellows showed variable exposure to ROP care.^{18,19} Curricula for nurses, technicians and general physicians should therefore cover image acquisition, recognition of ungradable images, the screening schedule and communication with parents about the importance of follow-up, because missed follow-up visits undermine the sensitivity of any screening strategy.

4.7. Strengths and limitations

The review had several strengths. It was restricted to data collected in low- and middle-income countries, it analysed three clinically distinct target conditions with bivariate models that account for the correlation

between sensitivity and specificity, and every two-by-two table was checked against the published estimates, with inconsistent reports excluded rather than corrected. The accuracy estimates were also presented alongside the economic evidence from the same settings, which allows programme planners to weigh accuracy and cost together.

Several limitations should be considered. Most studies were at high or unclear risk of bias, and AI studies often used retrospective, curated image sets with image-level splits. The unit of analysis differed between studies; the largest datasets were at the examination or image level, and treating correlated eyes, examinations and images as independent overstates precision and gives these datasets more weight. The infant-level sensitivity analysis gave similar estimates, but with wider confidence intervals. The reference standard and threshold varied between studies, several subgroups contained three or four datasets, and meta-regression could not separate the effects of income group and index test. Deeks' test had limited power with 12 and 16 datasets. More than a third of the reports sought could not be retrieved, most of them AI development studies in engineering journals, and eight reports were excluded because their data were internally inconsistent; if these studies differed systematically, the pooled estimates could be biased in either direction. Embase, Web of Science and Chinese databases were not searched. Some economic evaluations were available only as abstracts, and currencies and price years differed, which precluded quantitative pooling. Finally, studies from high-income countries were not included, so direct comparisons with those settings were not possible.

5. Conclusion

In low- and middle-income countries, telemedicine grading of retinal images detected clinically significant ROP with high sensitivity and specificity, and it was cost-effective in all available cost-utility analyses, although the certainty of the accuracy evidence was low. AI achieved comparable sensitivity but more false-positive results, and it has not yet been evaluated economically in these settings. The sensitivity for plus disease was lower, particularly with portable cameras. Telemedicine programmes with trained and certified non-physician imagers should be expanded with reliable follow-up systems, and the use of AI as a triage step within graded programmes should be evaluated prospectively after local calibration.

Declarations

Declaration of generative AI and AI-assisted technologies in the writing process

The authors declare that no generative artificial intelligence (AI) tools were used in the preparation of this manuscript.

Author contributions (CRediT)

I.W.E.S.: conceptualisation, supervision, writing – review and editing. N.M.A.S.: methodology, validation, writing – review and editing. I.B.A.: data curation, formal analysis, writing – original draft. C.P.: investigation, data curation, visualisation. All authors approved the final manuscript.

Conflict of interest

The authors declare no conflict of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Ethical approval and informed consent

Not applicable. This review analysed published data only and did not require ethical approval.

Consent for publication

Not applicable.

Data availability

The extracted two-by-two tables and the risk-of-bias judgements are provided in the Supplementary Material; the analysis code is available from the corresponding author on reasonable request.

Acknowledgements

The authors have no acknowledgements to declare.

References

- Wong ES, Choy RW, Zhang Y, et al. Global and regional trends in retinopathy of prematurity. *JAMA Ophthalmol.* 2026;144(2):139-47. doi:10.1001/jamaophthalmol.2025.5103
- Wang S, Liu J, Zhang X, et al. Global, regional and national burden of retinopathy of prematurity among childhood and adolescent: a spatiotemporal analysis based on the Global Burden of Disease Study 2019. *BMJ Paediatr Open.* 2024;8(1):e002267. doi:10.1136/bmjpo-2023-002267
- Chen J, Zhu Y, Li L, et al. Visual impairment burden in retinopathy of prematurity: trends, inequalities, and improvement gaps. *Eur J Pediatr.* 2024;183(4):1891-900. doi:10.1007/s00431-024-05450-5
- Zepeda Romero LC, Barrera de Leon JC, Gutiérrez Padilla JA, et al. Retinopathy of prematurity remains the leading cause of childhood blindness in children attending schools for the blind in Guadalajara: a fifteen-year comparison. *Front Pediatr.* 2026;14:1892954. doi:10.3389/fped.2026.1892954
- Siswanto JE, Bos AF, Dijk PH, et al. Multicentre survey of retinopathy of prematurity in Indonesia. *BMJ Paediatr Open.* 2021;5(1):e000761. doi:10.1136/bmjpo-2020-000761
- Siswanto JE, Adisasmita AC, Ronoatmodjo S, et al. A risk scoring model to predict progression of retinopathy of prematurity for Indonesia. *PLoS One.* 2023;18(2):e0281284. doi:10.1371/journal.pone.0281284
- Youn YA, Kim SY, Cho SJ, et al. Variations in medical practice of retinopathy of prematurity among 8 Asian countries from an international survey. *Sci Rep.* 2023;13(1):15602. doi:10.1038/s41598-023-42432-3
- Das A, Subramaniam P, Shah PK, et al. Prevalence of unscreened premature infants presenting with advanced stages of retinopathy of prematurity in South India: an analysis of 220 eyes. *J Pediatr Ophthalmol Strabismus.* 2024;61(5):339-43. doi:10.3928/01913913-20240508-02
- Mhina C, Mtogo Y, Mafwiri M, et al. Prevalence and associated factors for retinopathy of prematurity at a tertiary hospital in Dar es Salaam, Tanzania. *Eye (Lond).* 2025;39(8):1476-80. doi:10.1038/s41433-025-03651-2
- Sherief ST, Taye K, Teshome T, et al. Retinopathy of prematurity among infants admitted to two neonatal intensive care units in Ethiopia. *BMJ Open Ophthalmol.* 2023;8(1):e001257. doi:10.1136/bmjophth-2023-001257
- Nguyen TTB, Bui VT, Pham VPT, et al. Retinopathy of prematurity: a study of incidence and risk factors in a tertiary hospital in Vietnam. *Clin Ophthalmol.* 2022;16:3361-7. doi:10.2147/OPTH.S386808
- Du J, Chen X, Wang Y, et al. Regional variations in retinopathy of prematurity incidence for preterm infants <32 weeks' gestation in China. *Public Health.* 2024;226:91-8. doi:10.1016/j.puhe.2023.10.033
- Malcolm J, Ngabo D, Nziyomaze E, et al. Retinopathy of prematurity screening and treatment practice patterns: an African multi-country survey. *Graefes Arch Clin Exp Ophthalmol.* 2026. doi:10.1007/s00417-026-07403-0
- van der Lecq T, Rhoda N, Jordaan E, et al. Screening for retinopathy of prematurity (ROP) in South Africa: data from a newly established prospective regional register. *BMJ Open Ophthalmol.* 2025;10(1):e002036. doi:10.1136/bmjophth-2024-002036
- van der Lecq T, Rhoda N, Jordaan E, et al. Screening for retinopathy of prematurity in South Africa: are those developing severe ROP screened on time? Data from a prospective register. *BMJ Open Ophthalmol.* 2025;10(1):e002239. doi:10.1136/bmjophth-2025-002239
- Yulia DE, Soeharto DA. One year follow-up of intravitreal bevacizumab injection in aggressive retinopathy of prematurity at Indonesian national referral hospital: case series. *Ann Med Surg (Lond).* 2022;84:104853. doi:10.1016/j.amsu.2022.104853
- Chiang MF, Quinn GE, Fielder AR, et al. International classification of retinopathy of prematurity, third edition. *Ophthalmology.* 2021;128(10):e51-68. doi:10.1016/j.ophtha.2021.05.031
- Daniel Y, Young BK, Besirli CG, et al. Trends in retinopathy of prematurity (ROP) training and future practice intentions among retina and pediatric ophthalmology fellows. *J AAPOS.* 2026;30(2):104799. doi:10.1016/j.jaapos.2026.104799
- Joshi A, Tiwari S, Subramanyam A. Knowledge, practices, and challenges among neonatal intensive care unit nurses in retinopathy of prematurity screening: a comparative study of on-site versus telemedicine-based screening approaches. *Indian J Ophthalmol.* 2026. doi:10.4103/IJO.IJO_2410_25

20. Vinekar A, Gilbert C, Dogra M, et al. The KIDROP model of combining strategies for providing retinopathy of prematurity screening in underserved areas in India using wide-field imaging, tele-medicine, non-physician graders and smart phone reporting. *Indian J Ophthalmol.* 2014;62(1):41-9. doi:10.4103/0301-4738.126178
21. Asturias AL, Gilbert C, Silva JC, et al. Implementation of telemedicine screening for retinopathy of prematurity in rural areas in Guatemala. *J AAPOS.* 2022;26(1):22.e1-22.e5. doi:10.1016/j.jaapos.2021.08.307
22. Vedantham V, Ratnagiri PK. Vision on wheels: technician-assisted retinopathy of prematurity screening in rural Tamil Nadu and Andhra Pradesh: 2018-2021 project report. *Indian J Ophthalmol.* 2024;72(5):718-21. doi:10.4103/IJO.IJO_1981_23
23. Guo Z, Ma N, Wu Y, et al. The safety and feasibility of the screening for retinopathy of prematurity assisted by telemedicine network during COVID-19 pandemic in Wuhan, China. *BMC Ophthalmol.* 2021;21(1):258. doi:10.1186/s12886-021-02018-x
24. Gul N, Sethi S, Ahmad I, et al. Exploring retinopathy of prematurity: insights from the first USAID-funded screening program in KPK, Pakistan. *Pak J Med Sci.* 2025;41(7):2073-9. doi:10.12669/pjms.41.7.11021
25. Tawfik S, Mansour A, Selim NL, et al. Analysis of a two-year independent screening effort for retinopathy of prematurity in rural Egypt. *BMC Ophthalmol.* 2021;21(1):445. doi:10.1186/s12886-021-02193-x
26. Naraval ES, Santos-Rayos JY, Navarette MFS, et al. Retinopathy of prematurity computerized screening and monitoring system in a tertiary hospital in the Philippines. *Acta Med Philipp.* 2026;60(6):44-50. PMID: 42040879.
27. deCampos-Stairiker MA, Coyner AS, Gupta A, et al. Epidemiologic evaluation of retinopathy of prematurity severity in a large telemedicine program in India using artificial intelligence. *Ophthalmology.* 2023;130(8):837-43. doi:10.1016/j.ophtha.2023.03.026
28. Campbell JP, Nayak S, Ostmo S, et al. Application of a quantitative vascular severity score in retinopathy of prematurity in the United States and India: new insights into disease epidemiology and pathophysiology. *Am J Ophthalmol.* 2026;290:123-36. doi:10.1016/j.ajo.2026.06.008
29. Coyner AS, Young BK, Ostmo SR, et al. Use of an artificial intelligence-generated vascular severity score improved plus disease diagnosis in retinopathy of prematurity. *Ophthalmology.* 2024;131(11):1290-6. doi:10.1016/j.ophtha.2024.06.006
30. Coyner AS, Singh P, Brown JM, et al. Association of biomarker-based artificial intelligence with risk of racial bias in retinal images. *JAMA Ophthalmol.* 2023;141(6):543-52. doi:10.1001/jamaophthalmol.2023.1310
31. Sharafi SM, Amanollahi M, Ebrahimiadib N, et al. Does field of view matter? Assessing its impact on plus disease diagnosis in retinopathy of prematurity. *Int J Retina Vitreous.* 2025;11(1):127. doi:10.1186/s40942-025-00750-w
32. Catania F, Barjol A, Gallardo M, et al. Impact of image enhancement on grading agreement of plus disease features in retinopathy of prematurity. *Graefes Arch Clin Exp Ophthalmol.* 2026. doi:10.1007/s00417-026-07361-7
33. Luo Z, Ding X, Hou N, et al. A deep-learning-based collaborative edge-cloud telemedicine system for retinopathy of prematurity. *Sensors (Basel).* 2022;23(1):276. doi:10.3390/s23010276
34. Kirby RP, Malik ANJ, Palamouni KM, et al. Improved screening of retinopathy of prematurity (ROP): development of a target product profile (TPP) for resource-limited settings. *BMJ Open Ophthalmol.* 2023;8(1):e001197. doi:10.1136/bmjophth-2022-001197
35. Morrison SL, Dukhovny D, Chan RVP, et al. Cost-effectiveness of artificial intelligence-based retinopathy of prematurity screening. *JAMA Ophthalmol.* 2022;140(4):401-9. doi:10.1001/jamaophthalmol.2022.0223
36. Purohit N, Gupta PC, Buttan S, et al. Optimizing diabetic retinopathy screening at primary health centres in India: a cost-effectiveness analysis. *Pharmacoecon Open.* 2025;9(4):627-38. doi:10.1007/s41669-025-00572-4
37. Chen Y, Song F, Zhao Z, et al. Acceptability, applicability, and cost-utility of artificial-intelligence-powered low-cost portable fundus camera for diabetic retinopathy screening in primary health care settings. *Diabetes Res Clin Pract.* 2025;223:112161. doi:10.1016/j.diabres.2025.112161
38. Wang Y, Liu C, Hu W, et al. Economic evaluation for medical artificial intelligence: accuracy vs. cost-effectiveness in a diabetic retinopathy screening case. *NPJ Digit Med.* 2024;7(1):43. doi:10.1038/s41746-024-01032-9
39. Hu W, Joseph S, Li R, et al. Population impact and cost-effectiveness of artificial intelligence-based diabetic retinopathy screening in people living with diabetes in Australia: a cost effectiveness analysis. *EclinicalMedicine.* 2024;67:102387. doi:10.1016/j.eclinm.2023.102387
40. Liu H, Li R, Zhang Y, et al. Economic evaluation of combined population-based screening for multiple blindness-causing eye diseases in China: a cost-effectiveness analysis. *Lancet Glob Health.* 2023;11(3):e456-65. doi:10.1016/S2214-109X(22)00554-X
41. Smith SR, Nakibuuka V, Frick KD, et al. Oxygen blenders for prevention of retinopathy of prematurity are highly cost-effective in Uganda. *Front Med (Lausanne).* 2026;13:1715127. doi:10.3389/fmed.2026.1715127
42. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ.* 2021;372:n71. doi:10.1136/bmj.n71
43. Deeks JJ, Bossuyt PM, Leeflang MM, et al., editors. *Cochrane handbook for systematic reviews of diagnostic test accuracy. Version 2.0.* Chichester: Wiley; 2023. doi:10.1002/9781119756194
44. Husereau D, Drummond M, Augustovski F, et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations. *BMJ.* 2022;376:e067975. doi:10.1136/bmj-2021-067975
45. Shah PK, Narendran V, Saravanan VR, et al. Screening for retinopathy of prematurity--a comparison between binocular indirect ophthalmoscopy and RetCam 120. *Indian J Ophthalmol.* 2006;54(1):35-8. doi:10.4103/0301-4738.21612
46. Wu H, Dong Y, Mu X, et al. Digital retinal camera for screening of retinopathy of prematurity [in Chinese]. *Zhongguo Dang Dai Er Ke Za Zhi.* 2010;12(10):774-6.
47. Sekeroglu MA, Hekimoglu E, Sekeroglu HT, et al. Alternative methods for the screening of retinopathy of prematurity: binocular indirect ophthalmoscopy vs wide-field digital retinal imaging. *Eye (Lond).* 2013;27(9):1053-7. doi:10.1038/eye.2013.128
48. Lin HB, Li WL, Zhuo X, et al. Effects of two screening models for eye diseases in premature infants [in Chinese]. *Int Eye Sci.* 2018;18(3):596-8. doi:10.3980/j.issn.1672-5123.2018.3.48

49. Prakalapakorn SG, Freedman SF, Hutchinson AK, et al. Real-world simulation of an alternative retinopathy of prematurity screening system in Thailand: a pilot study. *J Pediatr Ophthalmol Strabismus*. 2018;55(4):245-53. doi:10.3928/01913913-20180327-04
50. Wongwai P, Suwannaraj S, Asawaphureekorn S. Diagnostic accuracy of a digital fundus photographic system for detection of retinopathy of prematurity requiring treatment (ROP-RT). *PLoS One*. 2018;13(7):e0201544. doi:10.1371/journal.pone.0201544
51. Tian RY, Zhang GM, Xie H, et al. Application of standard azimuth fundus camera technique in remote screening of retinopathy of prematurity [in Chinese]. *Recent Adv Ophthalmol*. 2019;39(6):568-70. doi:10.13389/j.cnki.rao.2019.0131
52. Adhikari S, Bajimaya S, Ranjekar EP. Validation of smartphone-based screening for retinopathy of prematurity in a low-resource setting. *J AAPOS*. 2021;25(5):274.e1-274.e5. doi:10.1016/j.jaapos.2021.04.010
53. Lanzelotte V, Vieira AA, Monteiro Fonseca AB, et al. ROP screening with the Pictor Plus camera: a telemedicine solution for developing countries. *J AAPOS*. 2022;26(5):244.e1-244.e6. doi:10.1016/j.jaapos.2022.05.019
54. Young BK, Cole ED, Shah PK, et al. Efficacy of smartphone-based telescreening for retinopathy of prematurity with and without artificial intelligence in India. *JAMA Ophthalmol*. 2023;141(6):582-8. doi:10.1001/jamaophthalmol.2023.1466
55. Wang J, Ju R, Chen Y, et al. Automated retinopathy of prematurity screening using deep neural networks. *EBioMedicine*. 2018;35:361-8. doi:10.1016/j.ebiom.2018.08.033
56. Zhang Y, Wang L, Wu Z, et al. Development of an automated screening system for retinopathy of prematurity using a deep neural network for wide-angle retinal images. *IEEE Access*. 2019;7:10232-41. doi:10.1109/ACCESS.2018.2881042
57. Wang J, Ji J, Zhang M, et al. Automated explainable multidimensional deep learning platform of retinal images for retinopathy of prematurity screening. *JAMA Netw Open*. 2021;4(5):e218758. doi:10.1001/jamanetworkopen.2021.8758
58. Cole E, Valikodath NG, Al-Khaled T, et al. Evaluation of an artificial intelligence system for retinopathy of prematurity screening in Nepal and Mongolia. *Ophthalmol Sci*. 2022;2(4):100165. doi:10.1016/j.xops.2022.100165
59. Coyner AS, Oh MA, Shah PK, et al. External validation of a retinopathy of prematurity screening model using artificial intelligence in 3 low- and middle-income populations. *JAMA Ophthalmol*. 2022;140(8):791-8. doi:10.1001/jamaophthalmol.2022.2135
60. Li P, Liu J. Early diagnosis and quantitative analysis of stages in retinopathy of prematurity based on deep convolutional neural networks. *Transl Vis Sci Technol*. 2022;11(5):17. doi:10.1167/tvst.11.5.17
61. Peng Y, Chen Z, Zhu W, et al. ADS-Net: attention-awareness and deep supervision based network for automatic detection of retinopathy of prematurity. *Biomed Opt Express*. 2022;13(8):4087-101. doi:10.1364/BOE.461411
62. Liu Y, Du Y, Wang X, et al. An artificial intelligence system for screening and recommending the treatment modalities for retinopathy of prematurity. *Asia Pac J Ophthalmol (Phila)*. 2023;12(5):468-76. doi:10.1097/APO.0000000000000638
63. Raja Sankari VM, Umapathy S, Alasmari S, et al. Automated detection of retinopathy of prematurity using quantum machine learning and deep learning techniques. *IEEE Access*. 2023;11:94306-21. doi:10.1109/ACCESS.2023.3311346
64. Rao DP, Savoy FM, Tan JZE, et al. Development and validation of an artificial intelligence based screening tool for detection of retinopathy of prematurity in a South Indian population. *Front Pediatr*. 2023;11:1197237. doi:10.3389/fped.2023.1197237
65. Subramaniam A, Orge F, Douglass M, et al. Image harmonization and deep learning automated classification of plus disease in retinopathy of prematurity. *J Med Imaging (Bellingham)*. 2023;10(6):061107. doi:10.1117/1.JMI.10.6.061107
66. Coyner AS, Murickian T, Oh MA, et al. Multinational external validation of autonomous retinopathy of prematurity screening. *JAMA Ophthalmol*. 2024;142(4):327-35. doi:10.1001/jamaophthalmol.2024.0045
67. Agrawal R, Kulkarni S, Deshpande M, et al. Classification of stages 1,2,3 and preplus, plus disease of ROP using MultiCNN_LSTM classifier. *MethodsX*. 2025;14:103182. doi:10.1016/j.mex.2025.103182
68. Alizadeh P, Manaberi M, Farahani AD, et al. Cost-utility analysis of retinopathy of prematurity screening strategies of universal referral, RetCam telemedicine, and direct examination in Iran: a Markov model approach. *BMC Health Serv Res*. 2026;26(1):1073. doi:10.1186/s12913-026-14872-w
69. Mohammadi SF, Rahban A, Darabeigi S, et al. Cost-effectiveness analysis of tele-retinopathy of prematurity screening in Iran. *Int J Ophthalmol*. 2021;14(4):560-6. doi:10.18240/ijo.2021.04.13
70. Neves LM, Haefeli LM, Zin AA, et al. Cost-utility analysis of wide-field imaging as an auxiliary technology for retinopathy of prematurity care in Brazil. *Front Pediatr*. 2021;9:757258. doi:10.3389/fped.2021.757258
71. Wongwai P, Kingkaew P, Asawaphureekorn S, et al. A store-and-forward telemedicine for retinopathy of prematurity screen: is it cost-effective in Thailand? *Asian Biomed*. 2015;9(5):665-73. doi:10.5372/1905-7415.0905.438
72. Shah SV, Brant A, James MR, et al. Cost-effectiveness of telemedicine-based screening programs for retinopathy of prematurity in low- and middle-income countries. *Ophthalmol Sci*. 2026;101395 [Epub ahead of print]. doi:10.1016/j.xops.2026.101395
73. Rothschild MI, Russ R, Brennan KA, et al. The economic model of retinopathy of prematurity (EcROP) screening and treatment: Mexico and the United States. *Am J Ophthalmol*. 2016;168:110-21. doi:10.1016/j.ajo.2016.04.014
74. Naguib MM, Soares RR, Anzures R, et al. Regionally specific economic impact of screening and treating retinopathy of prematurity in middle-income societies in the Philippines. *J Pediatr Ophthalmol Strabismus*. 2019;56(6):388-96. doi:10.3928/01913913-20190925-02
75. Alder CJ, Mutangana F, Phillips V, et al. Cost-effectiveness of addressing retinopathy of prematurity in Rwanda. *Ophthalmic Epidemiol*. 2025;32(3):341-9. doi:10.1080/09286586.2024.2372803
76. Zin AA, Magluta C, Pinto MF, et al. Retinopathy of prematurity screening and treatment cost in Brazil. *Rev Panam Salud Publica*. 2014;36(1):37-43.
77. Kelkar J, Kelkar A, Sharma S, et al. A mobile team for screening of retinopathy of prematurity in India: cost-

- effectiveness, outcomes, and impact assessment. Taiwan J Ophthalmol. 2017;7(3):155-9. doi:10.4103/tjo.tjo_48_17
78. Kelkar J, Agashe S, Kelkar A, et al. Mobile unit for retinopathy of prematurity screening and management at urban neonatal intensive care units: outcomes and impact assessment. Oman J Ophthalmol. 2017;10(1):13-6. doi:10.4103/0974-620X.200684
79. Padhi TR, Bhunia S, Das T, et al. Outcome of real-time telescreening for retinopathy of prematurity using videoconferencing in a community setting in Eastern India. Indian J Ophthalmol. 2024;72(5):697-703. doi:10.4103/IJO.IJO_2024_23
80. Del Busto Wilhelm E, Zimmermann Paiz MA, Ordóñez Rivas AM, et al. Cost of screening and timely treatment versus the comprehensive management of the severe visual impairment due to retinopathy of premature [in Spanish]. Andes Pediatr. 2023;94(3):333-8. doi:10.32641/andespediatr.v94i3.4530