中山大學 中山眼科中心 ZHONGSHAN OPHTHALMIC CENTER, SUN YAT-SEN UNIVERSITY

Contents lists available at Lighthouse Scholar

EYE SCIENCE

journal homepage: journal.gzzoc. com

doi: 10.12419/es24082501
View this article at: https://dx.doi.org/10.12419/es24082501

· Review Article ·

Application and performance of artificial intelligence in screening retinopathy of prematurity from 2018 to 2024: a meta-analysis and systematic review

Rui Liu[#](刘睿), Guina Liu[#](刘桂娜), Xiaoshuang Jiang (蒋小爽), Fang Lu (陆方)

Department of Ophthalmology, West China Hospital, Sichuan University, Chengdu, China

HIGHLIGHTS

- This review revealed the significant impact of artificial intelligence (AI) in identifying ROP and differentiating PLUS disease. However, considerable heterogeneity was observed across the included studies. Further researches will be needed to address these disparities and optimize the use of AI to improve healthcare for ROP.
- The research was a comprehensive meta-analysis to appraise the performance of AI screening on ROP and PLUS disease from 2018 to 2024. Statistical analysis included data pooling, forest plot construction, heterogeneity testing, and meta-regression. QUADAS-2 and QUADAS-AI were used to evaluate the quality of included studies better.
- AI will shine in the future of screening and treating ROP.
- Sharing data to build public database platform is encouraged.
- More relevant researches will be needed in the future to enhance the robustness of AI screening gold standard.

Abstract: **Purpose:** Artificial intelligence (AI) significantly enhances the screening and diagnostic processes for retinopathy of prematurity (ROP). In this article,we focused on the application and performance of AI in detecting ROP and distinguishing plus disease (PLUS) in ROP. **Methods:** We searched PubMed, Embase, Medline, Web of Science, and Ovid for studies published from January2018 to July 2024. Studies evaluating the diagnostic performance of AI with expert ophthalmologists'judgment as a reference standard were included. The risk of bias was assessed using the QUADAS-2 tool and QUADAS-AI tool.Statistical analysis included data pooling, forest plot construction, heterogeneity testing, and meta-regression. **Results:** Fourteen of the 186 studies

Received date: 2024-07-25; **Revised date:** 2024-08-14; **Accepted date:** 2024-08-30; **Published online:** 2024-09-15

# These authors have contributed equally tothis work.

Corresponding author: Fang Lu, E-mail: lufang@wchscu.cn

Full Text

were included.The pooled sensitivity, specificity and the area under the curve (AUC) of the AI diagnosing ROP were 0.95 (95% CI 0.93-0.96), 0.97 (95% CI 0.94-0.98) and 0.97 (95% CI 0.95-0.98), respectively.The pooled sensitivity, specificity and the AUC of the AI distinguishing PLUS were 0.92 (95% CI 0.80-0.97), 0.95 (95% CI 0.91-0.97) and 0.98 (95% CI 0.96-0.99), respectively. Cochran's Q test ($P< 0.01$) and Higgins I^2 heterogeneity index revealed considerable heterogeneity. The country of study, number of centers, data source and the number of doctors were responsible for the heterogeneity. For ROP diagnosing, researches conducted in China using private data in single center with less than 3 doctors showed higher sensitivity and specificity. For PLUS distinguishing, researches in multiple centers with less than 3 doctors showed higher sensitivity. **Conclusions:** This study revealed the powerful role of AI in diagnosing ROP and distinguishing PLUS. However, significant heterogeneity was noted among all included studies, indicating challenges in the application of AI for ROP diagnosis in real-world settings. More studies are needed to address these disparities, aiming to fully harness AI's potential in augmenting medical care for ROP.

Keywords: retinopathy of prematurity; plus disease; artificial intelligence; diagnosis; meta-analysis; systematic review

Cite this article as: Liu R, Liu GN, Jiang XS, Lu F. Application and performance of artificial intelligence in screening retinopathy of prematurity from 2018 to 2024: a meta-analysis and systematic review. Eye Science, 2024, 1(3): 206-223. doi: 10.12419/es24082501

INTRODUCTION

Retinopathy of prematurity (ROP) is a major cause of potentially preventable blindness among preterm infants.[1] The incidence of ROP is increasing due to rapid improvements in rescue technology in neonatology worldwide.[1-2] Global age-standardized rates of blindness and vision loss attributable to ROP have dramatically increased from 1990 to 2019. In 2019, there were 2,169,800 (95% CI 1,714.5 to 2,670.1 thousand) cases of blindness and vision loss due to ROP worldwide, compared to 1,870,100 cases in 1990.[3] Inadequate management of oxygen delivery is likely a key factor in the increasing incidence of ROP-related morbidity among preterm infants.[4] Plus disease (PLUS), characterized by venous dilation and arteriolar tortuosity in the posterior retinal vessels, serves as a critical indicator for identifying cases of ROP that require treatment.[5] Early screening, timely diagnosis, and advanced treatment are effective strategies for preventing vision loss related to ROP.[1-3]

The clinical diagnosis of ROP relies solely on the appearance of retinal vessels, assessed through dilated ophthalmoscopic examination by retinal doctors, making it highly subjective.[6] Research has highlighted the imbalance between the limited number of experienced ophthalmologists and the large number of preterm infants needing ROP screening and treatment, particularly in developing countries.[7-8] In China, uneven development of the pediatric care system, inadequately trained pediatricians and unmet demand for pediatric care are major challenges. There are approximately 4 pediatrics per 10 000 children.[9] With the huge population in China, the workload of pediatric ophthalmologists for screening ROP is especially excessive.[10] Moreover, personalized screening and accurate diagnosis are crucial for each newborn's condition varies.[1,11-13] Therefore, developing efficient and accurate diagnostic tools is vital.

Since the emergence of the MYCIN system, artificial

intelligence (AI) has begun to play a significant role in areas such as medical diagnosis, treatment planning, and drug discovery.[14-15] In recent years, AI-based automatic screening systems have been rapidly developed, with the advantages to save time and reduce subjectivity.[16] AI has demonstrated significant capabilities in diagnosing ocular diseases, like age-related macular degeneration (AMD),[17-18] glaucoma,[19] and diabetic retinopathy (DR).[20] Similarly, AI has been utilized in research related to ROP. According to Brown et al., increasingnumber of studies focused on AI-based screening of ROP, which might become a valuable tool in the future.[6]

In April 2018, FDA authorized the first AI diagnostic system IDx-DR for diagnosing DR using color fundus photographs with a sensitivity of 87.4% and specificity of 89.5% for 900 patients with diabetes at ten primary care sites.[20] This approval marked a significant milestone in the application of AI for diagnosing retinal diseases.

Previous meta-analysis research has either concentrated on binary screening for ROP or on the identification of PLUS. However, a comprehensive evaluation of the potential of AI in screening ROP and PLUS would be essential for guiding clinical practice. Considering the brilliant improvement of AI in retinal disease diagnostics in 2018 and the promising future of AI applications in ROP, we collected studies from 2018 to 2024, performing a first meta-analysis study to comprehensively assess the performance of AI, aiming to objectively appraise the current diagnostic performance of AI for ROP and PLUS at the same time.

METHODS

Protocol

This meta-analysis was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analysis,[21] with a standardized review and data extraction protocol. The study protocol was registered on the PROSPERO platform under entry number CRD42024564204.

Search strategy and selection criterion

We searched PubMed, Embase, Medline, Web of Science and Ovid for studies published between January, 2018, and July, 2024. The full search strategy for each database is available in Appendix 1. Manual searches of bibliographies and citations from included studies were also completed to identify additional potentially missed articles.

Only studies aiming to identify the presence of AI in ROP were identified. We accepted standard-of-care diagnosis, expert opinion or consensus as adequate reference standards to classify the disease. We excluded studies that did not test the diagnostic performance or just investigated the accuracy of image segmentation.

Inclusion and exclusion criteria

The inclusion criteria were (1) studies using AI in ROP diagnosis or PLUS classification; (2) studies using clinical diagnosis as the reference standard; (3) original scientific articles; (4) sufficient data for reconstructing 2×2 tables for diagnostic accuracy.

The extraction criteria were (1) duplication of publications; (2) non-original studies, including editorials, letters to the editor, review articles and case reports; (3) non-English articles; (4) studies without sufficient information for reconstructing a 2×2 table.

Data extraction

We extracted the following data from the included studies using a standardized form: (1)true positives, false negatives, true negatives, and false positives; (2) study

characteristics, including the first author, publication year, country, camera, reference standard, model, algorithm evaluation, screening criteria, source of data (public database or private dataset, data from hospital was defined as private dataset), number of centers, number of doctors, experience year of doctors,gestational age (GA), birth weight (BW), gender (M/F), dataset (validation dataset), classification, outcome, sensitivity and specificity, accuracy, and AUROC.

Quality Assessment

The methodological quality of the studies was assessed using the QUADAS-2[22] and QUADAS-AI.[23] Each study was rated in the following domains: patient selection, index test, reference standard, and flow and timing. Each domain was assessed based on the risk of bias and the first three domains, including applicability.

Statistical analysis

We created 2×2 tables to calculate the pooled sensitivity, specificity, and corresponding 95% confidence intervals (CIs) using a bivariate random effects model. Moreover, we calculated the diagnostic odds ratio (DOR), a single indicator combining the sensitivity and specificity, was selected for its capacity to demonstrate the overall accuracy of diagnostic test across all threshold settings and less affected by the prevalence of disease among included samples. The positive likelihood ratio (LR^+) and negative likelihood ratio (LR^-) would provide information on the probability of disease increases or decreases with a positive or negative test result, respectively. The results are graphically shown in the forest plots. We constructed hierarchical summary receiver operating characteristic (HSROC) curves. Furthermore, we calculated the area under the curve (AUC). We performed Deeks' funnel plot asymmetry test to evaluate the possible presence of publication bias, with $P < 0.1$ indicating the possibility of publication bias.[24] The heterogeneity of the included studies was evaluated using the inconsistency index (I^2) and Q statistic of the chi-square test.[25]

Heterogeneity was further explored through meta-regression by adding the following covariates to the bivariate model: (1) country (China *vs.* other countries), (2) number of centers (<1 *vs.*≥ 1), (3) data source (public database *vs.* private data), and (4) number of doctors (≤3 *vs.* >3).

Statistical analyses were performed using STATA 17.0 and RevMan 5.3. Statistical significance was set at $P < 0.05$.

RESULTS

Selection and data extraction

A total of 186 studies were identified and 147 were excluded according to the exclusion criteria. Thirty-nine full-text articles were assessed for eligibility, and 14 studies were finally included in the meta-analysis.[6,26-38] Twenty-five studies were excluded for various reasons, mainly including no 2×2 table available and AI used for other purposes not diagnosing ROP or PLUS (Figure 1).

Data characteristics and demographics

These characteristics were summarized in Table 1. All researches were conducted retrospectively, from 2006 to 2018. Most studies conducted in China, while the rest 5 studies were from other counties like India, America, New Zealand, Japan and the UK. All studies emphasized the expert judgement as reference standards. Nine studies provided information on their screening criteria, while five did not. Twelve studies reported their algorithm evaluation, while two studies did not. The data was

Table 1 Characteristics of 14 included studies

First Author	Publication Year	Country	Camera	Reference standard	Model	Algorithm Evaluation	Screening Criteria	Source of data	Number of centers	Number of doctors
Brown[6]	2018	America	RetCam	Clinical diagnosis	CNN: U-Net and Inception V1	the 5-fold cross validation	NA	i-ROP	8	3
Wang[26]	2018	China	RetCam3	ICROP, CRYO-ROP, and ETROP	Id-Net; Gr-Net	NR	NA	Hospital	1	4
Hu[27]	2019	China	RetCam3	ICROP	CNNs: VGG-16, Inception-V2; ResNet-50	select the best model	NA	Hospital	1	3
Tan[28]	2019	New Zealand	RetCam	ETROP	ROP. AI	NA	< 1250 g birth weight or < 30 weeks gestational age	ART-ROP	4	NA
Zhang[29]	2019	China	Retcam 2/3	Clinical diagnosis	DNN: AlexNet, VGG 16, GoogLeNet	select the best model	1) birth weight <2,000 g and 2) preterm infants with birth weight 2,000 g but having severe systemic disorders (according to pediatricians' assessment).	Hospital	1	5
Huang[30]	2020	China and Japan	RetCam	ICROP	DNN: VGG19*, VGG16, InceptionV3, DenseNet, and MobileNet	select the best module and then 5-fold cross validation	born within 37 weeks of gestation and/or had to weigh ≤ 1500 g at birth	Hospital	2	3
Mao[31]	2020	China	RetCam	Clinical diagnosis	U-Net, DenseNet	Select the best model based on the 5-fold cross-validation	NA	Hospital	1	1
Tong[32]	2020	China	RetCam	Clinical diagnosis	ResNet; Faster-RCNN	10-fold cross-validation	NA	Hospital	1	13
Huang[33]	2021	China	RetCam	Clinical diagnosis	CNN	5-fold cross-validation	infants with a BW of 1500–2000 g or a GA above 32 weeks with any unstable clinical condition	Hospital	3	3
Lei[34]	2021	China	RetCam2 or 3	ICROP	CASA, Grade CAM, Res-Net 50	Select the best model	Birth weight ≤2000 g and gestational age≤36.5 weeks	ROP Group	1	5
Ramachandran[35]	2021	India	RetCam3	ICROP	U-COSFIRE; Darknet-53	Select the best model	BW<2000 g or GW<34w	KIDROP	1	3
Li[36]	2022	China	RetCam3	Clinical diagnosis	Retina U-Nets; Dense Net	Select the best model based on the 5-fold cross-validation	BW<2000 g and GW<37w	Hospital	1	3
Attallah[37]	2023	China	Retcam2/3	Clinical diagnosis	ResNet-50; DarkNet-53; MobileNet	5-fold cross-validation	<2000 g in weight at birth and 2000 g premature neonates who have significant systemic diseases at birth.	Hospital	30	5
Wagner[38]	2023	UK	RetCam Version 2	Clinical diagnosis	Bespokeand CFDL models	Select the best model	BW<1501 g or GW≤32w	Hospital	1	4

Table 1 (continued)

First Author	Experience year of doctors	GA/w	Birth Weight/kg	Gender (M/F)	Dataset (Validation dataset)	Classification	Validation Dataset	Outcome	Sensitivity/ Specificity	Accuracy	AUROC
Brown[6]	2 ophthalmologists and 1 coordinator	NA	NA	NA	5 511 (100)	cases	54Normal, 31pre-plus, 15PLUS	Normal *vs.* pre and plus	0.93/0.94	0.91	0.98*
								Normal and pre *vs.* PLUS	1/0.94	-	
Wang[26]	NA	NA	NA	93/78	2 226 (298)	cases	149 Normal, 149 ROP, 52 minor ROP, 52 severe ROP	ROP *vs.* no-ROP	0.97/0.99	NA	NA
					(520) 104			Minor *vs.* Severe ROP#	0.88/0.92	NA	NA
Hu[27]	1 chief physician and 2 doctors 5+ years	32 (25-41)	1.994 (0.7-4.25)	NA	2 068 (300)	images	150 ROP, 150 no ROP	ROP *vs.* no-ROP	0.96/0.98	0.97	0.99
					466 (100)		50 mild, 50 severe	Mild *vs.* Severe ROP×	0.82/0.86	0.84	0.92
Tan[28]	NA	NA	NA	NA	3 487 (116)	images	33 PLUS, 26 pre-plus; 57 normal	PLUS *vs.* not-PLUS	0.94/0.81	0.86	0.98
								Pre plus *vs.* normal	0.81/0.81	0.81	-
Zhang[29]	2 chief physicians, 2 attending physicians, 1 resident	32.0 (25,36.2)	1.50 (0.78-2.00	10075/ 7726	19 543 (17 801)	images	8,090 ROP, 9,711 without ROP	ROP *vs.* no ROP	0.941/ 0.993	0.9	0.998
Huang[30]	10 years of experience working	NA	NA	NA	267 (101)	cases	59 ROP, 42 no ROP	ROP *vs.* no ROP	0.97/0.95	0.96	0.97
					254 (85)		63 mild ROP, 22 severe ROP	Mild *vs.* severe$	0.99/0.99	0.99	0.99
Mao[31]	NA	31.0 ± 2.0	1.583 3 ± 0.401 6	NA	5 711 (450)	images	305 normal, 104 pre-plus, 41 PLUS	PLUS *vs.* not PLUS	0.95/0.98	-	0.93
								Preplus *vs.* not preplus	0.92/0.97	-	0.99
Tong[32]	Junior (11), 10 years (2)	NA	NA	NA	36 231 (9 772)	images	519 Grading', 261 PLUS, 8,992 normal	Grading' *vs.* others	0.78/0.93	0.90	-
								PLUS *vs.* not PLUS	0.71/0.91	0.90	-
Huang[33]	At least 3years	NA	NA	NA	1975 (244)	images	94 no-ROP, 44 Stage 1, 106 Stage 2	ROP *vs.* no ROP	0.96/0.96	0.92	0.96
								Stage 1 *vs.* others	0.92/0.95	-	0.93
								Stage 2 *vs.* others	0.90/0.99	-	0.92
Lei[34]	Two are chief physicians, two are attending physicians, one is junior ophthalmologist	NA	NA	NA	22961 (5160)	images	3,078 ROP, 2,082 no ROP	ROP *vs.* no ROP	0.95/0.99	0.99	0.99
Ramachandran[35]	NA	No-PLUS: 32.4 ± 1.1 PLUS: 30.9 ± 1.8	No-PLUS: 1,350 ± 240 PLUS: 1.925 ± 0.774	NA	289(161)	images	94 normal, 67 plus	PLUS *vs.* no PLUS	0.99/0.98	0.97	0.99

Table 1 (continued)

First Author	Experience year of doctors	GA/w	Birth Weight/kg	Gender (M/F)	Dataset (Validation dataset)	Classification	Validation Dataset	Outcome	Sensitivity/ Specificity	Accuracy	AUROC
Li[36]	NA	30.43 ± 5.80	1.442 03 ± 0.517 03	NA	18,827 (3,680)	images	2,893 no ROP, 378 stage I, 262 stageII, 147 stage III	Stage I *vs.* others	0.90/0.98	0.98	0.9663
								Stage II *vs.* others	0.93/0.99		
								Stage III *vs.* otherts	0.92/0.99		
								Normal *vs.* others	0.96/0.96		
Attallah[37]	2 chief physicians, 2 attending physicians, 1 resident	31.9 (24-36.4) 32.0 (25-36.2)	1.49 (0.63-2.00) 1.50 (0.78-2.00)	10,075/ 7,726 988/754	17,801 (1,742)	images	155 ROP, 1,587 no ROP	ROP *vs.* no ROP	0.90/0.97	0.94	0.98
Wagner[38]	3 years	NA	NA	NA	6,141 (200)	Images	111 no ROP, 43 pre plus, 46 PLUS	ROP *vs.* no ROP	0.973/0.900	-	0.986
								Pre plus *vs.* others	0.860/0.860	-	0.927
								PLUS *vs.* others	0.522/0.981	-	0.974

ROP Retinopathy of Prematurity

*ICROP*International Committee for the Classification of Retinopathy of Prematurity

*CYROP*Cryotherapy for Retinopathy of Prematurity

ETROP Early Treatment ROP

ART-ROP Auckland Regional Telemedicine ROP

CNN convolutional neural network

DNN deep convolutional neural network

AUROC area under the receiver operating characteristic curve

#Minor ROP was defined as patients who required reexamination after two weeks, and severe ROP was defined as patients who required immediate treatment by a clinical ophthalmologist.

× Mild is defined as stage 1 or 2, and severe is defined as stage 3-5.

$Mild is defined as stage 1 and 2, and severe is defined as stage 3.

*Different models were tested; the most accurate and most accurate results are quoted.

☆AP-ROP, Aggressive Posterior Retinopathy of Prematurity

obtained from private datasets or public database like the i-ROP, the KID ROP, the ROP Group and the ART-ROP. Among 14 studies, nine were single-center studies, while five were multicenter studies, from at least 2 centers to 30 centers at most. For the labeling process, most studies reported the number of doctors, but one study did not mention it. Over half of the studies provided informationon the experience of the doctors. Only six studies provided information on GA and BW of included premature infants. Three studies classified the material as “cases”, while the others classified the material as “images”. Moreover, eleven studies reported AI in diagnosing the presence of ROP, six studies reported AI in diagnosing the presence of PLUS and three studies reported AI distinguishing severe ROP which was considered treatments in clinical practice.

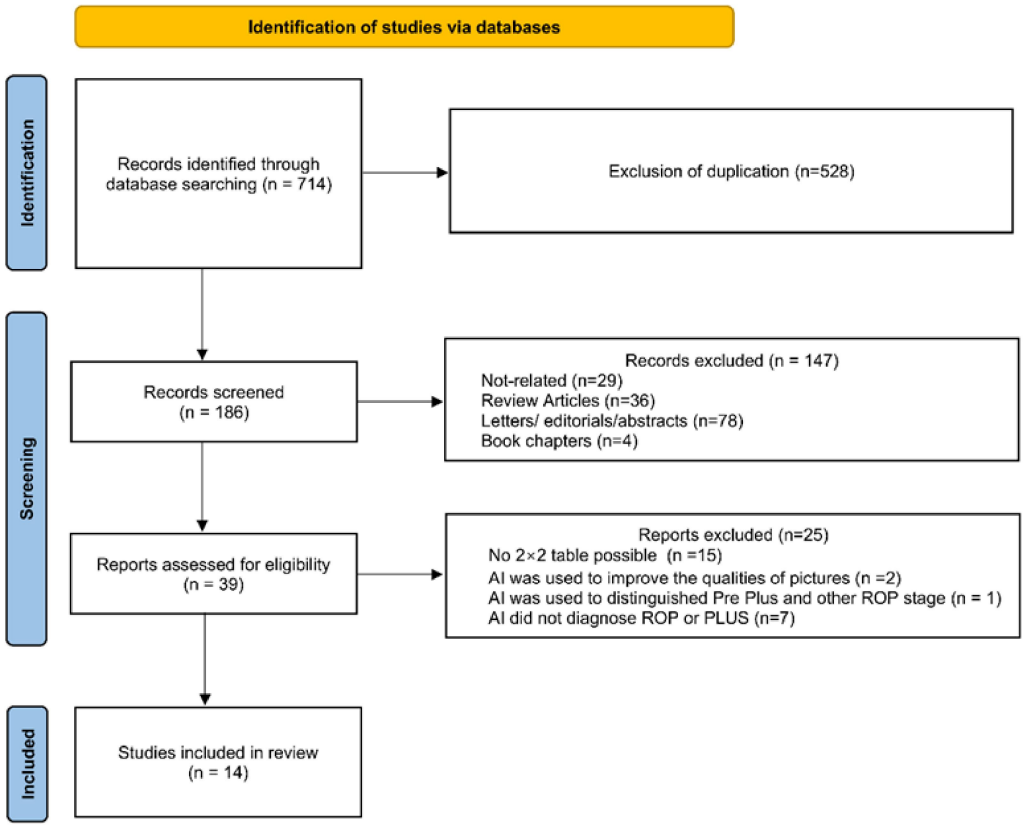

Figure 1 PRISMA flow chart of article selection process.

Quality Assessment

Figure 2 showed the quality variables of the 14 included studies. All studies had a low risk of bias in index tests and reference standard. However, the exclusion of low-definition images without ensuring a consecutive or random sample might bring bias in the patient selection process. Consequently, studies that excluded low-quality images were considered to have high risk of bias, while those did not emphasize the exclusion of such images had unclear risks. In addition, it would be optimal for the results of the index test and the reference standard to be collected simultaneously in order to prevent misclassification, due to the progress of diseases.[22] In the case of ROP, a progressive disease, it was necessary to have clear time intervals between the selection of images and their validation. However, all included studies did not emphasize, leading to an unclear risk of bias in the section of flow and timing.

Diagnostic accuracy

For 11 studies using AI to diagnose ROP, the pooled sensitivity and specificity were 0.95 (95% CI 0.93-0.96) and 0.97 (95% CI 0.94-0.98), respectively (Figure 3a). The AUC was 0.97 (95% CI 0.95-0.98) (Figure 4a). The DOR of AI for diagnosing ROP was 611 (95% CI 300-1,244). The LR^+ was 31.7 (95% CI 16.7-59.5), and the LR^- was 0.05 (95% CI 0.04-0.07) (Table 2). There was considerable among-study heterogeneity according to Cohran's Q test ($P < 0.01$) and the I^2 heterogeneity index (Figure 3a). Deeks' funnel plots revealed major publication bias in AI diagnosing ROP disease, with statistical significance (P=0.02) (Figure 5a).

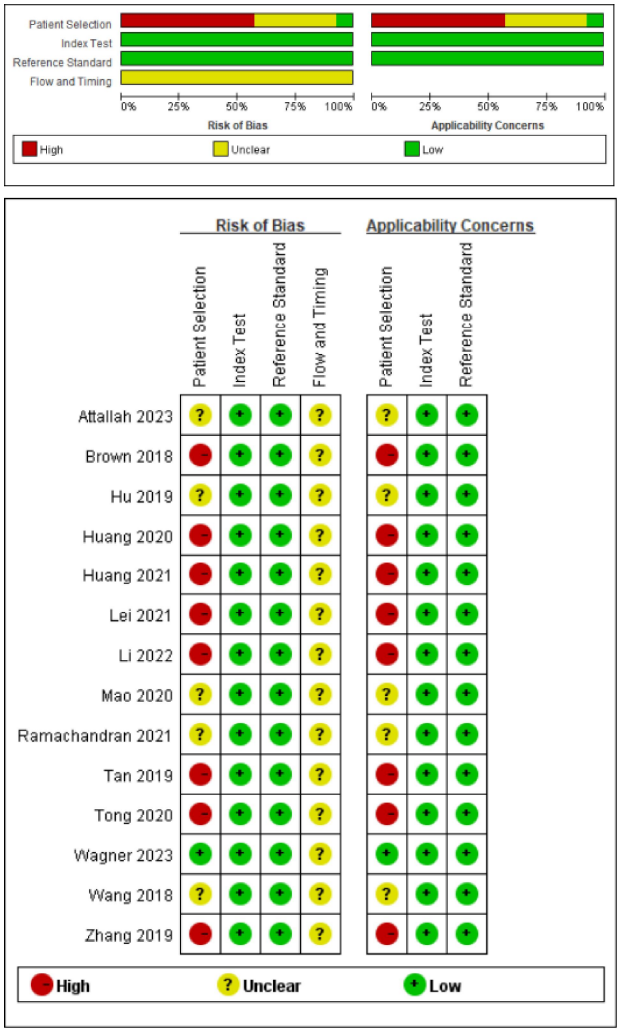

Figure 2 Risk of bias and applicability concerns graph and quality assessment of diagnostic accuracy studies-2 (QUADAS-2) and quality assessment of diagnostic accuracy studies-AI (QUADAS-AI) criteria for the 14 included studies.

In 9 studies using AI to distinguish PLUS, the pooled sensitivity and specificity were 0.92 (95% CI 0.80-0.97) and 0.95 (95% CI 0.91-0.97), respectively (Figure 3b). The AUC was 0.98 (95% CI 0.96-0.99) (Figure 4b). The DOR of PLUS was 218 (95% CI 58-815). The LR^+ was 18.5 (95% CI 9.9-34.8) and the LR^- was 0.09 (95% CI 0.03-0.22) (Table 2). There was considerable among-study heterogeneity according to Cohran's Q test ($P < 0.01$) and the I^2 heterogeneity index (Figure 3b). Deeks' funnel plots revealed no major publication bias in AI diagnosing PLUS($P = 0.07$) (Figure 5b).

Table 2 Sensitivity, specificity, LR^+, LR^-, and DORs of AI detection in the ROP and PLUS cohorts

	AI detection in ROP	AI detection in PLUS
Sensitivity (95%CI)	0.95 (0.93, 0.96)	0.92 (0.80, 0.97)
Specificity (95%CI)	0.97 (0.94, 0.98)	0.95 (0.91, 0.97)
LR^+ (95%CI)	31.7 (16.7, 59.9)	18.5 (9.9, 34.8)
LR^- (95%CI)	0.05 (0.04, 0.07)	0.09 (0.03,0.22)
DOR (95%CI)	611 (300, 1,244)	218 (58, 815)

Abbreviations: LR^+, likelihood ratio positive; LR^-, likelihood ratio negative; DOR, diagnostic odds ratio; CI, confidence interval.

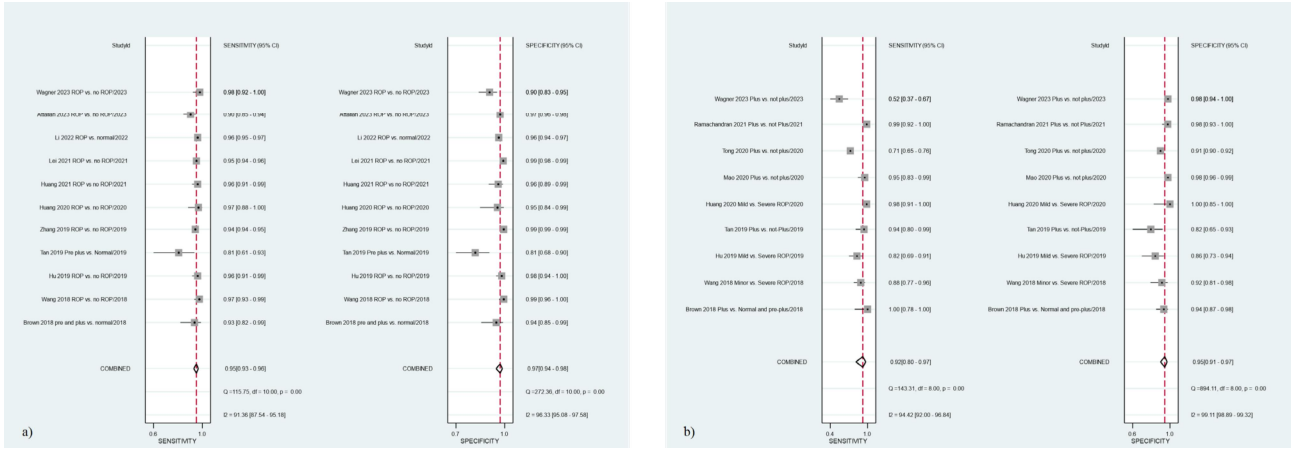

Figure 3 Coupled forest plots of the pooled sensitivity and specificity of AI detection in ROP patients (a) and in the PLUS cohort (b).

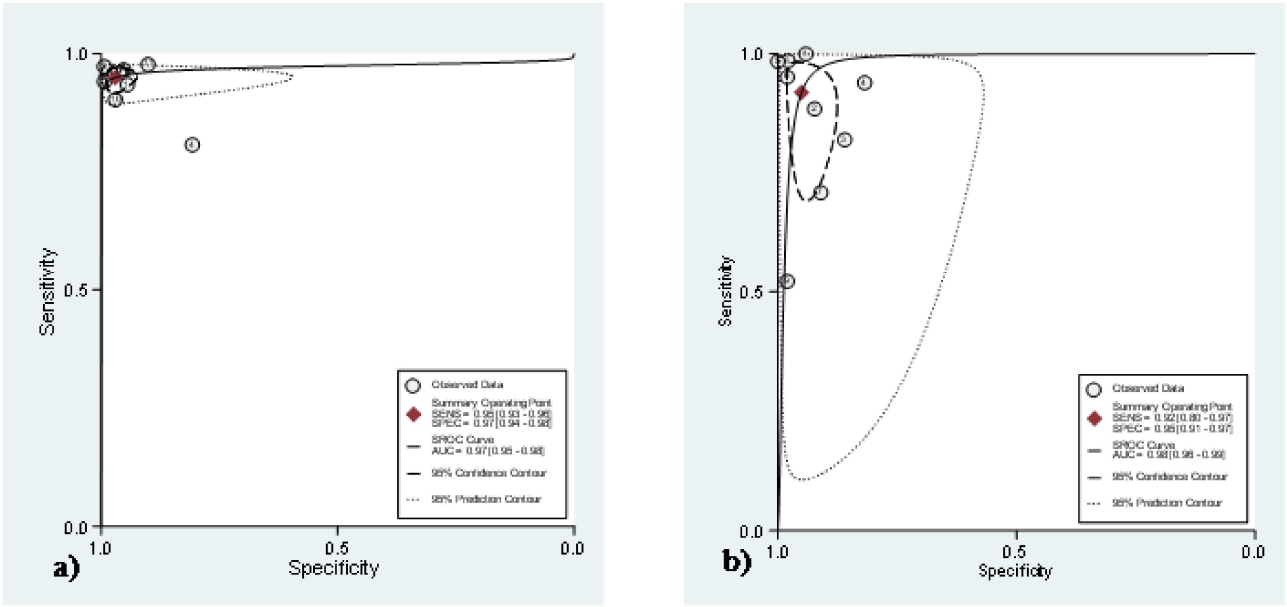

Figure 4 Hierarchical summary receiver operating characteristic (HSROC) curve of the diagnostic performance of AI detection in patients with ROP (a) and in patients with PLUS (b).

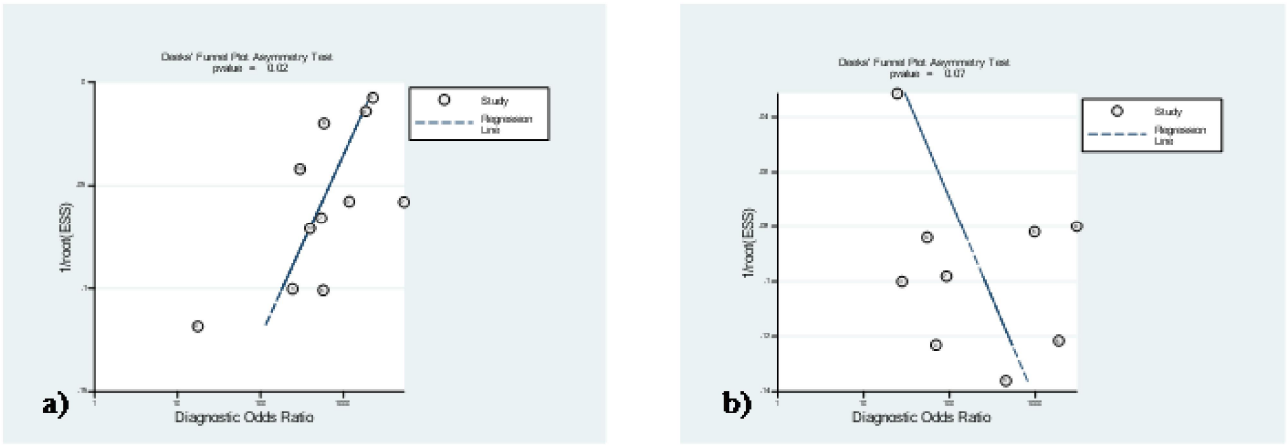

Figure 5 Deeks' funnel plot used to evaluate the potential publication bias of AI detection in ROP patients (a) and in patients with PLUS (b).

Meta-regression

Meta-regression was used to explore the causes of heterogeneity among the patients who received an AI-based diagnosis of ROP and those who was diagnosed as PLUS (Table 3). Study heterogeneity was independently associated with country, number of centers, data source and number of doctors responsible for the initial screening.

Among the 11 studies on the AI diagnosis of ROP, 8 studies conducted in China had a greater sensitivity (0.95, 95% CI 0.95-0.96 *vs.* 0.94, 95% CI 0.90-0.98, $P < 0.01$) than 3 studies conducted in other countries. Multicenter studies had lower sensitivity (0.93, 95% CI 0.91-0.96 *vs.* 0.95 95% CI 0.95-0.96, $P < 0.01$) and specificity (0.94, 95% CI 0.90-0.99 *vs.* 0.98 95% CI 0.97-0.99, $P < 0.01$) than single-center studies. Studies whose data

Table 3 Results of the meta-regression analysis of the AI for the detection of ROP and the PLUS

AI detection	Covariates	Category	Studies (n)	Meta analytic summary estimates			
				Sensitivity (95% CI)	*P*	Specificity (95% CI)	*P*
ROP	Country	China	8	0.95 (0.95-0.96)	< 0.01	0.98 (0.97-0.99)	0.60
		Other country	3	0.94 (0.90-0.98)		0.89 (0.82-0.97)	
	Centers	>1	5	0.93 (0.91-0.96)	< 0.01	0.94 (0.90-0.99)	< 0.01
		=1	6	0.95 (0.95-0.96)		0.98 (0.97-0.99)	
	Data source	Hospitals	8	0.95 (0.94-0.97)	< 0.01	0.97 (0.96-0.99)	0.59
		Database	3	0.93 (0.89-0.97)		0.95 (0.90-1.00)	
	Doctors	≥3	5	0.94 (0.92-0.97)	< 0.01	0.96 (0.92-0.99)	0.01
		<3	6	0.95 (0.94-0.97)		0.98 (0.96-1.00)	
PLUS	Country	China	5	0.91 (0.80-1.00)	0.85	0.95 (0.90-0.99)	0.06
		Other country	4	0.93 (0.83-1.00)		0.95 (0.91-1.00)	
	Centers	>1	3	0.98 (0.95-1.00)	0.03	0.93 (0.86-1.00)	0.05
		=1	6	0.86 (0.75-0.98)		0.96 (0.93-0.98)	
	Data source	Hospitals	6	0.86 (0.75-0.98)	0.06	0.96 (0.92-0.99)	0.40
		Database	3	0.98 (0.95-1.00)		0.94 (0.87-1.00)	
	Doctors	≥3	3	0.73 (0.53-0.93)	< 0.01	0.95 (0.90-1.00)	0.11
		<3	6	0.96 (0.90-1.00)		0.95 (0.92-0.99)	

were obtained from hospitals had a greater sensitivity (0.95, 95% CI 0.94-0.97 *vs.* 0.93 95% CI 0.89-0.97, $P < 0.01$) than those whose data were obtained from public database. The sensitivity (0.94, 95% CI 0.92-0.97 *vs.* 0.95, 95% CI 0.94-0.97, $P < 0.01$) and specificity (0.96, 95% CI 0.92-0.99 *vs.* 0.98, 95% CI 0.96-1.00, $P < 0.01$) of screening by more than three doctors were lower than those of screening by 3 or fewer doctors.

Among the 9 studies in which AI was used to distinguish PLUS, multicenter studies had greater sensitivity (0.98, 95% CI 0.95-1.00 *vs.* 0.86 95% CI 0.75-0.98, P=0.03).The screening of more than three doctors had a lower sensitivity than that of screening by 3 or fewer doctors (0.73, 95% CI 0.53-0.93 *vs.* 0.96 95% CI 0.90-1.00, $P < 0.01$).

DISCUSSION

AI has been widely used in medical diagnostic research, but few has been actually used in practice. In this meta-analysis, we evaluated the performance of AI detection in the diagnosis of ROP and PLUS, which was the first study to comprehensively assess the performance of AI in ROP and PLUS at the same time.The results demonstrated that the AI system achieved high sensitivity

and specificity in identifying both ROP and PLUS.

For diagnosing ROP, high sensitivity was emphasized to ensure that no patients at potential risk were overlooked. We achieved a strong sensitivity of 0.95 (95% CI 0.93-0.96) across 11 studies, with values ranging from 0.81 to 0.98. The AUC of AI for diagnosing of ROP was 0.97 (95% CI 0.95-0.98), showing outstanding diagnostic efficiency.[39] Moreover, the value of LR^- was 0.05, indicating that infants diagnosed without ROP by AI had a low risk of developing ROP. For diagnosing PLUS, high specificity was critical to accurately identify those needing potential therapy to prevent the adverse outcomes of ROP. We got a great specificity value of 0.95 (95% CI 0.91-0.97) from 9 researches, from 0.82 to 1.00. The AUC of AI for the diagnosis of PLUS also performed outstandingly, with a value of 0.98 (95% CI 0.96-0.99). The value of LR^+ was 18.5, indicating that patients diagnosed with PLUS by AI are highly likely to actually have PLUS and require close attention for therapy. Consequently, the included studies suggested that AI detection for ROP and PLUS was effective and there still existed potential space to improve sensitivity in ROP and specificity in PLUS, with the highest sensitivity reaching 0.98[38] in ROP and the highest specificity reaching 1.00 in PLUS.[6]

Multiple factors contribute to heterogeneity, including patients' selection, time interval, publication bias, and so on.Excluding low-quality fundus photographs or those taken from peripheral angles may artificially inflate the sensitivity reported in AI diagnoses.[40-41] Additionally, variations in AI algorithms across studies cause discrepancies in how disease parameters like vessel tortuosity, direction, or ridge position are assessed, leading to inconsistencies even within the same study using different AI tools.[6,26,34,37] The choice of time interval between image captures is another critical factor affecting results. Researchers might choose higher-resolution images to obtain more favorable results, thereby introducing a selection bias based on the timing of the photographs.[42-43] Similarly, most included studies presented a high risk or provided unclear information regarding the time intervals, yet selecting appropriate intervals was crucial since ROP progresses rapidly.If the time interval was not suitable, AI diagnoses using earlier fundus photos might significantly differ from specialist diagnoses based on current fundus examinations with an ophthalmoscope.[44-45] Therefore, the bias of timing and flow was essential in diagnostic models. A synchronous diagnosis would be conducted using AI and standard reference in the future. Moreover, publication bias contributed to the heterogeneity. This phenomenon may be linked to enterprise support.Publication bias can lead to overestimated or underestimated effects, potentially resulting in inappropriate therapies in clinical practice.[46-47] Given the factors contributing to heterogeneity, emphasis was placed on appropriate patient selection and study design, as well as improving the adaptability of AI to varying qualities of fundus photos and disease parameters.[48-49] Overall, risk bias may bring high heterogeneity which finally reduce diagnostic effectiveness of AI in ROP and PLUS.

Meta-regression analysis was employed to explore study heterogeneity based on country, number of centers, data sources, and number of doctors.Significant differences were observed in the sensitivity of ROP diagnosis among different countries, varying numbers of centers, and data sources.[50-52] However, variability in sensitivity across different studies may result from the varying number of research centers and the diversity of data sources. For the number of research centers,AI applications in single-center studies typically encountered a less heterogeneous population and

simplify data collection, thereby reducing confounding factors.[53-54] In contrast, multi-center studies exhibited significant variations among patients, image data, and instruments.[37-38] Therefore, binary diagnosis of ROP in a single-center setting tended to achieve higher sensitivity. For the data source,private data demonstrated greater sensitivity compared to data from public databases. Private data, sourced exclusively from a single hospital and utilizing the same equipment, ensured consistent data quality. This consistency enabled AI to effectively focus on training and recognition, thus enhancing its sensitivity. In contrast, public data posed more challenges due to variable quality and lack of timely updates; however, it can be useful for training AI on robustness in diagnosis, given the diverse patient populations and varied photo quality.[55-56] Finally, to enhance the exploration of AI models for diagnosing ROP, timely data sharing from various research centers was encouraged.

The performance of AI in diagnosing PLUS has also been investigated. Unlike the diagnosis of ROP, the identification of PLUS was more effective in multi-center studies than in single-center studies. Several reasons might explain this.Firstly, AI identification currently relies primarily on vascular segmentation. The characteristics of vascular curvature in PLUS can be easily identified.[57-58] Therefore, despite the significant variation in image data quality from different sources, AI can more effectively extract and identify the characteristics of vascular curvature.[59] Secondly, AI can quantify vasodilation and tortuosity in PLUS, overcoming differences in image data quality between multi-center studies and the subjective diagnosis of doctors.[6,32,54] Thirdly, PLUS primarily affected the quadrants around the optic disk, where the identification range was more concentrated, reducing the demand for information from peripheral blood vessels. [31,60] Moreover, the optic disc was served as a point of reference for positioning recognition,[61-62] aiding AI in identifying features, thereby improving the training and increasing the applicability of AI. Therefore, it can be seen that the difficulties brought by multi-center research to the previous binary diagnosis of ROP performed better in the recognition of PLUS. Moreover, the number of doctors responsible for the initial screening also introduced bias. In our study, both in ROP diagnosis and PLUS identification, AI demonstrated lower sensitivity when more than three doctors were responsible for the initial screening.This result highlighted the fragility of the current gold standard, which heavily relies on the experience and diagnostic consistency of the involved practitioners.[63-64] To mitigate potential biases, establishing a more robust gold standard for AI training is crucial. Implementing this could significantly enhance the effectiveness and reliability of AI applications in clinical practice.[63]

Although AI presents promising potential for clinical applications, several significant challenges remain. AI models trained on homogeneous datasets may struggle to generalize across diverse populations, particularly in low- and middle-income regions with distinct ROP phenotypes.[65] Additionally, the lack of transparency in AI decision-making processes, often referred to as "black box" algorithms, raises concerns regarding explainability, which is critical in fields like ophthalmology, where diagnostic accuracy is paramount.[66] Furthermore, determining liability when AI assists in clinical decision-making remains unclear. Over-reliance on AI could also hinder the development of essential clinical skills.[67] Thus, optimizing AI's integration into clinical practice should be the focus of future research.

This study has several limitations. Firstly, the included studies exhibited considerable heterogeneity. Although we conducted a meta-regression analysis, the exploration of factors contributing to heterogeneity might

have been insufficient.Secondly, none of the included studies adequately emphasized the time intervals between assessments, leading to a high risk of bias in the flow and timing domain. Thirdly, some studies did not provide information on patient selection or photo inclusion criteria, leading to a high risk of bias.Finally, many studies lacked detailed information on patient characteristics, which is crucial for diagnosing ROP in clinical practice.

In conclusion, AI demonstrated excellent performance in diagnosing ROP and PLUS. Given the current shortage of pediatric ophthalmologists, AI could serve as a valuable tool for ROP screening.However, heterogeneity poses a significant challenge to the use of AI in clinical practice.It is recommended that more well-designed studies be conducted to enhance the generalizability of AI in diagnosing ROP and PLUS.

Correction notice

None

Acknowledgement

None

Author Contributions

(I) Conception and design: Rui Liu, Guina Liu

(II) Administrative support: Fang Lu, Jiang Xiaoshuang

(III) Provision of study materials or patients: Rui Liu

(IV) Collection and assembly of data: Rui Liu, Guina Liu

(V) Data analysis and interpretation: Ruiyang Li, Wenben Chen, and Jialing Chen

(VI) Manuscript writing: All authors

(VII) Final approval of manuscript: All authors

Funding

This work was supported by Sichuan University West China Hospital 2024 Plateau Medicine Center '1.3•5 Project (Project NosGYYX24011).

Conflict of Interests

None of the authors has any conflicts of interest to disclose. All authors have declared in the completed the ICMJE uniform disclosure form.

Patient consent for publication

None

Ethical Statement

None

Provenance and Peer Review

None

Data Sharing Statement

None

References

1. Sabri K, Ells AL, Lee EY, et al. Retinopathy of prematurity: a global perspective and recent developments. Pediatrics. 2022, 150(3): e2021053924. DOI: 10.1542/peds.2021-053924.
2. Bao Y, Ming WK, Mou ZW, et al. Current application of digital diagnosing systems for retinopathy of prematurity. Comput Methods Programs Biomed, 2021, 200:105871. DOI: 10.1016/j.cmpb.2020.105871.
3. 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 Pediatr Open. 2024,

8(1):e002267. DOI:10.1136/bmjpo-2023-002267 .

4. Darlow BA, Husain S. Primary prevention of ROP and the oxygen saturation targeting trials. Semin Perinatol. 2019, 43(6): 333-340. DOI: 10.1053/j.semperi.2019.05.004.
5. Ghergherehchi L, Kim SJ, Campbell JP, et al. Plus disease in retinopathy of prematurity: more than meets the ICROP? Asia Pac J Ophthalmol. 2018, 7(3): 152-155.DOI: 10.22608/APO.201863.
6. Brown JM, CampbellJP, Beers A, et al. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. JAMA Ophthalmol. 2018, 136(7): 803-810. DOI: 10.1001/jamaophthalmol.2018.1934.
7. Barrero-Castillero A, Corwin BK, VanderVeen DK, etal. Workforce shortage for retinopathy of prematurity care and emerging role of telehealth and artificial intelligence. Pediatr Clin North Am. 2020, 67(4): 725-733. DOI: 10.1016/j.pcl.2020.04.012.
8. Vartanian RJ, Besirli CG, Barks JD, etal. Trends in the screening and treatment of retinopathy of prematurity. Pediatrics. 2017, 139(1): e20161978. DOI: 10.1542/peds.2016-1978.
9. Zhang Y, Huang L, Zhou X, et al. Characteristics and workload of pediatricians in China. Pediatrics. 2019, 144(1): e20183532. DOI: 10.1542/peds.2018-3532.
10. Yang Q, Zhou X, Ni Y, et al. Optimised retinopathy of prematurity screening guideline in China based on a 5-year cohort study. Br J Ophthalmol. 2021, 105(6): 819-823. DOI: 10.1136/bjophthalmol-2020-316401.
11. Schumacher AC, Ball ML, Arnold AW, et al. Oculocardiac reflex during ROP exams. Clin Ophthalmol Auckl N Z. 2020, 14: 4263-4269. DOI: 10.2147/OPTH.S288043.
12. Belda S, Pallás CR, De la Cruz J, et al. Screening for retinopathy of prematurity: is it painful? Biol Neonate. 2004, 86(3): 195-200. DOI: 10.1159/000079542.
13. Hered RW, Gyland EA. The retinopathy of prematurity screening examination: ensuring a safe and efficient examination while minimizing infant discomfort. Neonatal Netw. 2010, 29(3): 143-151. DOI: 10.1891/0730-0832.29.3.143.
14. Amsterdam D. Perspective: limiting antimicrobial resistance with artificial intelligence/machine learning. BME Front. 2023, 4: 0033. DOI: 10.34133/bmef.0033.
15. Jiménez-Luna J, Grisoni F, Weskamp N, et al. Artificial intelligence in drug discovery: recent advances and future perspectives. Expert Opin Drug Discov. 2021, 16(9): 949-959.DOI: 10.1080/17460441.2021.1909567.
16. Razzak, M. I., Naz, S. & Zaib, A. in Classification in BioApps: Automation of Decision Making (eds Nilanjan Dey, Amira S. Ashour, & Surekha Borra) 323-350 (Springer International Publishing, 2018).
17. Li JO, Liu H, Ting DSJ, et al.Digital technology, tele-medicine and artificial intelligence in ophthalmology: A global perspective. Prog Retin Eye Res. 2021, 82: 100900. DOI: 10.1016/j.preteyeres.2020.100900.
18. Li Z, Wang L, Wu X, et al. Artificial intelligence in ophthalmology: the path to the real-world clinic. Cell Rep Med. 2023, 4(7): 101095. DOI: 10.1016/j.xcrm.2023.101095.
19. Perepelkina T, Fulton AB. Artificial intelligence (AI) applications for age-related macular degeneration (AMD) and other retinal dystrophies. Semin Ophthalmol,2021, 36(4): 304-309.DOI: 10.1080/08820538.2021.1896756.
20. Abràmoff MD, Lavin PT, Birch M, et al. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med. 2018, 1: 39. DOI: 10.1038/s41746-018-0040-6.
21. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ. 2009, 339: b2700. DOI: 10.1136/bmj.b2700.
22. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of

diagnostic accuracy studies. Ann Intern Med. 2011, 155(8): 529-536. DOI: 10.7326/0003-4819-155-8-201110180-00009.

23. Sounderajah V, Ashrafian H, Rose S, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. Nat Med. 2021, 27: 1663-1665. DOI: 10.1038/s41591-021-01517-0.
24. Deeks JJ, MacAskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol. 2005, 58(9): 882-893. DOI: 10.1016/j.jclinepi.2005.01.016.
25. Higgins JPT, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. BMJ. 2003, 327(7414): 557-560. DOI: 10.1136/bmj.327.7414.557.
26. Wang J, Ju R, Chen Y, et al. Automated retinopathy of prematurity screening using deep neural networks. EBio Medicine. 2018, 35: 361-368. DOI: 10.1016/j.ebiom.2018.08.033.
27. Hu J, Chen Y, Zhong J, et al. Automated analysis for retinopathy of prematurity by deep neural networks. IEEE Trans Med Imaging. 2019, 38(1): 269-279. DOI: 10.1109/TMI.2018.2863562.
28. Tan Z, Simkin S, Lai C, et al. Deep learning algorithm for automated diagnosis of retinopathy of prematurity plus disease. Transl Vis Sci Technol. 2019, 8(6): 23. DOI: 10.1167/tvst.8.6.23.
29. 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. 1042, 7: 10232-10241. DOI: 10.1109/ACCESS.2018.2881042.
30. Huang, Y, Vadloori, S, Chu, H, et al. Deep Learning Models for Automated Diagnosis of Retinopathy of Prematurity in Preterm Infants. ELECTRONICS. 2020, 9(9): 1444. DOI: 10.3390/electronics9091444.
31. Mao J, Luo Y, Liu L, et al. Automated diagnosis and quantitative analysis of plus disease in retinopathy of prematurity based on deep convolutional neural networks. Acta Ophthalmol. 2020, 98(3): e339-e345. DOI: 10.1111/aos.14264.
32. Tong Y, Lu W, Deng QQ, et al. Automated identification of retinopathy of prematurity by image-based deep learning. Eye Vis. 2020, 7: 40. DOI: 10.1186/s40662-020-00206-2.
33. Huang YP, Basanta H, Kang EYC, et al. Automated detection of early-stage ROP using a deep convolutional neural network. BrJ Ophthalmol. 2021, 105(8): 1099-1103. DOI: 10.1136/bjophthalmol-2020-316526.
34. Lei, B, Zeng, X, Huang, S, et al.Automated detection of retinopathy of prematurity by deep attention network. Multimedia Tools and Applications. 2021, 80(30): 36341-36360, DOI:10.1007/s11042-021-11208-0 .
35. Ramachandran S, Niyas P, Vinekar A, et al. A deep learning framework for the detection of Plus disease in retinal fundus images of preterm infants. Biocybern Biomed Eng. 2021, 41(2): 362-375. DOI: 10.1016/j.bbe.2021.02.005.
36. 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.
37. Attallah O. Gab ROP: Gabor wavelets-based CAD for retinopathy of prematurity diagnosis *via* convolutional neural networks. Diagnostics, 2023,13(2): 171. DOI: 10.3390/diagnostics13020171.
38. Wagner SK, Liefers B, Radia M, et al. Development and international validation of custom-engineered and code-free deep-learning models for detection of plus disease in retinopathy of prematurity: a retrospective study. Lancet Digit Health. 2023, 5(6): e340-e349. DOI: 10.1016/S2589-7500(23)00050-X.
39. Mandrekar J. Receiver operating characteristic curve in diagnostic test assessment. J THORAC ONCOL. 2010, 5 (9): 1315-6. DOI: 10.1097/JTO.0b013e3181ec173d.
40. Chuter B, Huynh J, Bowd C, et al. Deep learning identifies high-quality fundus photographs and increases accuracy in automated primary open angle glaucoma detection.

Transl Vis Sci Technol. 2024, 13(1): 23. DOI: 10.1167/tvst.13.1.23.

41. Shi C, Lee J, Wang G, et al. Assessment of image quality on color fundus retinal images using the automatic retinal image analysis. Sci Rep. 2022, 12: 10455. DOI: 10.1038/s41598-022-13919-2.
42. Kwon MR, Chang Y, Park B, et al. Performance analysis of screening mammography in Asian women under 40 years. Breast Cancer. 2023, 30(2): 241-248. DOI: 10.1007/s12282-022-01414-5.
43. Vijayakumar K, Rajinikanth V, Kirubakaran MK. Automatic detection of breast cancer in ultrasound images using Mayfly algorithm optimized handcrafted features. J Xray Sci Technol. 2022, 30(4): 751-766. DOI: 10.3233/XST-221136.
44. Fierson WM, American Academy of Pediatrics Section on Ophthalmology, American Academy of Ophthalmology, et al. Screening examination of premature infants for retinopathy of prematurity. Pediatrics. 2018, 142(6): e20183061. DOI: 10.1542/peds.2018-3061.
45. Parrozzani R, Nacci EB, Bini S, et al. Severe retinopathy of prematurity is associated with early post-natal low platelet count. Sci Rep. 2021, 11(1): 891. DOI: 10.1038/s41598-020-79535-0.
46. Joober R, Schmitz N, Annable L, et al. Publication bias: what are the challenges and can they be overcome? J Psychiatry Neurosci. 2012, 37(3): 149-152. DOI: 10.1503/jpn.120065.
47. Lin L, Chu H. Quantifying publication bias in meta-analysis. Biometrics. 2018, 74(3): 785-794. DOI: 10.1111/biom.12817.
48. Lin D, Xiong J, Liu C, et al. Application of Comprehensive Artificial intelligence Retinal Expert (CARE) system: a national real-world evidence study. Lancet Digit Health. 2021, 3(8): e486-e495. DOI: 10.1016/S2589-7500(21)00086-8.
49. Jin K, Ye J. Artificial intelligence and deep learning in ophthalmology: current status and future perspectives. Adv Ophthalmol Pract Res. 2022, 2(3): 100078. DOI: 10.1016/j.aopr.2022.100078.
50. Chen Y, Feng J, Li F, et al. Analysisofchangesincharacteristicsofsevereretinopathyofprematuritypatientsafterscreening guidelineswereissuedin China. Retina. 2015, 35(8): 1674-1679. DOI: 10.1097/iae.0000000000000512.
51. Zin A, Gole GA. Retinopathy of prematurity-incidence today. Clin Perinatol. 2013, 40(2): 185-200. DOI: 10.1016/j.clp.2013.02.001.
52. Sun H, Dong Y, Liu Y, et al. Using ROP Score and CHOP ROP for early prediction of retinopathy of prematurity in a Chinese population. Ital J Pediatr. 2021, 47(1): 39. DOI: 10.1186/s13052-021-00991-z.
53. Bleker J, Yakar D, van Noort B, et al. Single-center versus multi-center biparametric MRI radiomics approach for clinically significant peripheral zone prostate cancer. Insights Imaging. 2021, 12(1): 150. DOI: 10.1186/s13244-021-01099-y.
54. Bellomo R, Warrillow SJ, Reade MC. Why we should be wary of single-center trials. Crit Care Med. 2009, 37(12): 3114-3119. DOI: 10.1097/CCM.0b013e3181bc7bd5.
55. de Kok JWTM, de la Hoz MÁA, de Jong Y, et al. A guide to sharing open healthcare data under the General Data Protection Regulation. Sci Data. 2023, 10(1): 404. DOI: 10.1038/s41597-023-02256-2.
56. Chen H, Yu P, Hailey D, et al. Identification of the essential components of quality in the data collection process for public health information systems. Health Informatics J. 2020, 26(1): 664-682. DOI: 10.1177/1460458219848622.
57. Nisha KL, SreelekhaG, Sathidevi PS, et al. A computer-aided diagnosis system for plus disease in retinopathy of prematurity with structure adaptive segmentation and vessel based features. Comput Med Imaging Graph. 2019, 74: 72-94. DOI: 10.1016/j.compmedimag.2019.04.003.
58. Yildiz VM, TianP, Yildiz I, et al. Plus disease in retinopathy of prematurity: convolutional neural network

performance using a combined neural network and feature extraction approach. Transl Vis Sci Technol. 2020, 9(2): 10. DOI: 10.1167/tvst.9.2.10.

59. Campbell JP, Ataer-CansizogluE, Bolon-Canedo V, et al. Expert diagnosis of plus disease in retinopathy of prematurity from computer-based image analysis. JAMA Ophthalmol. 2016, 134(6): 651-657. DOI: 10.1001/jamaophthalmol.2016.0611.
60. Keck KM, Kalpathy-Cramer J, Ataer-Cansizoglu E, et al. Plus disease diagnosis in retinopathy of prematurity. Retina. 2013, 33(8): 1700-1707. DOI: 10.1097/iae.0b013e3182845c39.
61. Rogers DL, Bremer DL, Fellows RR, et al. Comparison of strategies for grading retinal images of premature infants for referral warranted retinopathy of prematurity. J AAPOS. 2017, 21(2): 141-145. DOI: 10.1016/j.jaapos.2017.01.001.
62. Sharafi SM, Ebrahimiadib N, Roohipourmoallai R, et al. Automated diagnosis of plus disease in retinopathy of prematurity using quantification of vessels characteristics. Sci Rep. 2024, 14(1): 6375. DOI:10.1038/s41598-024-57072-4 (2024).
63. Kurvers RHJM, Herzog SM, Hertwig R, et al. Boosting medical diagnostics by pooling independent judgments. Proc Natl Acad Sci U S A. 2016, 113(31): 8777-8782. DOI: 10.1073/pnas.1601827113.
64. Kämmer JE, Hautz WE, Herzog SM, et al. The potential of collective intelligence in emergency medicine: pooling medical students' independent decisions improves diagnostic performance. Med Decis Making. 2017, 37(6): 715-724. DOI: 10.1177/0272989X17696998.
65. Ting DSW, PasqualeLR,Peng L, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol. 2019, 103(2): 167-175. DOI: 10.1136/bjophthalmol-2018-313173.
66. Parmar UPS, Surico PL,Singh RB, et al. Artificial intelligence (AI) for early diagnosis of retinal diseases. Medicina. 2024, 60(4): 527. DOI: 10.3390/medicina60040527.
67. Bottomley D, Thaldar D. Liability for harm caused by AI in healthcare: an overview of the core legal concepts. Front Pharmacol. 2023, 14: 1297353. DOI: 10.3389/fphar.2023.1297353.

Appendix 1

We show the search strategy for a. PubMed b. Embase c. Medline d. Web of Science e. Ovid

a. PubMed

(retinopathy of prematurity[Title/Abstract]) AND (convolutional neural networks[Title/Abstract])

(retinopathy of prematurity[Title/Abstract]) AND (machine learning[Title/Abstract])

(artificial intelligence[Title/Abstract] OR AI[Title/Abstract] OR deep learning[Title/Abstract]) AND (retinopathy of prematurity[Title/Abstract] OR ROP [Title/Abstract])

b. Embase

(retinopathy of prematurity) AND (convolutional neural networks)

(retinopathy of prematurity) AND (machine learning)

(artificial intelligence OR AIOR deep learning) AND (retinopathy of prematurity OR ROP)

c. Medline

retinopathy of prematurity or ROP AND artificial intelligence or ai or a.i. or machine learning or deep learning or convolutional neural networks or CNNs

d. Web of Science

(TS=(retinopathy of prematurity OR ROP)) AND TS=(artificial intelligence OR AI OR deep learning OR convolutional neural networks OR CNN OR machine learning))

e. Ovid

(Retinopathy of prematurity OR ROP) AND (artificial intelligence OR AI OR deep learning OR convolutional neural networks OR CNN OR machine learning)