JAMA Ophthalmology | Original Investigation

Multinational External Validation of Autonomous Retinopathy of Prematurity Screening

Aaron S. Coyner, PhD; Tom Murickan, MD; Minn A. Oh, PhD; Benjamin K. Young, MD; Susan R. Ostmo, MS; Praveer Singh, PhD; R. V. Paul Chan, MD, MSc; Darius M. Moshfeghi, MD; Parag K. Shah, MD; Narendran Venkatapathy, MD; Michael F. Chiang, MD, MA; Jayashree Kalpathy-Cramer, PhD; J. Peter Campbell, MD, MPH

IMPORTANCE Retinopathy of prematurity (ROP) is a leading cause of blindness in children, with significant disparities in outcomes between high-income and low-income countries, due in part to insufficient access to ROP screening.

OBJECTIVE To evaluate how well autonomous artificial intelligence (AI)–based ROP screening can detect more-than-mild ROP (mtmROP) and type 1 ROP.

DESIGN, SETTING, AND PARTICIPANTS This diagnostic study evaluated the performance of an AI algorithm, trained and calibrated using 2530 examinations from 843 infants in the Imaging and Informatics in Retinopathy of Prematurity (i-ROP) study, on 2 external datasets (6245 examinations from 1545 infants in the Stanford University Network for Diagnosis of ROP [SUNDROP] and 5635 examinations from 2699 infants in the Aravind Eye Care Systems [AECS] telemedicine programs). Data were taken from 11 and 48 neonatal care units in the US and India, respectively. Data were collected from January 2012 to July 2021, and data were analyzed from July to December 2023.

EXPOSURES An imaging processing pipeline was created using deep learning to autonomously identify mtmROP and type 1 ROP in eye examinations performed via telemedicine.

MAIN OUTCOMES AND MEASURES The area under the receiver operating characteristics curve (AUROC) as well as sensitivity and specificity for detection of mtmROP and type 1 ROP at the eye examination and patient levels.

RESULTS The prevalence of mtmROP and type 1 ROP were 5.9% (91 of 1545) and 1.2% (18 of 1545), respectively, in the SUNDROP dataset and 6.2% (168 of 2699) and 2.5% (68 of 2699) in the AECS dataset. Examination-level AUROCs for mtmROP and type 1 ROP were 0.896 and 0.985, respectively, in the SUNDROP dataset and 0.920 and 0.982 in the AECS dataset. At the cross-sectional examination level, mtmROP detection had high sensitivity (SUNDROP: mtmROP, 83.5%; 95% CI, 76.6-87.7; type 1 ROP, 82.2%; 95% CI, 81.2-83.1; AECS: mtmROP, 80.8%; 95% CI, 76.2-84.9; type 1 ROP, 87.8%; 95% CI, 86.8-88.7). At the patient level, all infants who developed type 1 ROP screened positive (SUNDROP: 100%; 95% CI, 81.4-100; AECS: 100%; 95% CI, 94.7-100) prior to diagnosis.

CONCLUSIONS AND RELEVANCE Where and when ROP telemedicine programs can be implemented, autonomous ROP screening may be an effective force multiplier for secondary prevention of ROP.

Invited Commentary

Supplemental content

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: J. Peter Campbell, MD, MPH, Casey Eye Institute, Oregon Health & Science University, 545 SW Campus Dr, Portland, OR 97239 (campbelp@ohsu.edu).

JAMA Ophthalmol. doi:10.1001/jamaophthalmol.2024.0045
Published online March 7, 2024.

Retinopathy of prematurity (ROP) is largely preventable yet remains a leading cause of childhood blindness.[1,2] Unequitable resource distribution and differences in the epidemiology of premature birth contribute to a higher at-risk population in regions where there are not enough ophthalmologists to screen for and treat blinding ROP. This, unfortunately, worsens as neonatal mortality improves.[3] Each year, an estimated 30 000 to 50 000 infants, predominantly from low-income and middle-income countries, experience ROP-related visual loss. Artificial intelligence (AI) is emerging as a key tool for disease screening, with several autonomous diabetic retinopathy (DR) screening systems already approved by the US Food and Drug Administration, which are expanding globally.[4-7] For ROP, autonomous screening could be implemented in one of several large ROP telemedicine programs in the US and India, where it may be even more impactful than DR screening since the population is captive—within neonatal intensive care units (NICUs)—and the time period for screening is finite.[8-11]

In autonomous deployment, the purpose of AI is to screen for and refer, rather than to diagnose, a disease. In DR, an output of more-than-mild DR that complements the Early Treatment of Diabetic Retinopathy Study (ETDRS) Diabetic Retinopathy Severity Score was developed.[12] In ROP, disease severity is described using 2 frameworks: the International Classification of ROP (ICROP) and the Early Treatment for ROP (ETROP) categories, of which there are 4: (1) no ROP, (2) less than type 2 ROP (ie, mild ROP), (3) type 2 ROP, and (4) type 1 ROP.[1,13,14] While it is currently recommended that all cases of type 1 ROP be treated, some cases of type 2 ROP are treated at clinicians' discretion.[13,15] Ultimately, the most important criterion determining the need for treatment is the presence of plus disease, which the most recent ICROP has defined as a spectrum of vascular abnormality from normal to pre-plus disease to plus disease.[1,13] Thus, one way to integrate these 2 paradigms into a simple binary heuristic (refer or not) would be to consider more-than-mild ROP (mtmROP), which we define as eyes with type 2 ROP or type 1 ROP or any eye with pre-plus disease.

Developed as a plus disease classifier, the i-ROP deep learning (DL) algorithm has since been used to assign a vascular severity score (VSS) to better reflect the spectrum of plus disease, which has been endorsed as an assistive software as medical device for ROP.[16-18] Previous work has also validated the concept that VSS may be a useful surrogate for overall ROP severity in an eye; however, many AI algorithms demonstrate efficacy in preliminary studies with curated datasets but fail, for a number of reasons, to demonstrate effectiveness in clinical practice.[19-25] Herein, we detail the optimization of the i-ROP DL, which involves retraining a more efficient model, using Monte Carlo dropout (MCD) and model ensembling for increased repeatability, and setting an autonomous mtmROP threshold through validation on the i-ROP consortium dataset. We also developed an image preprocessing pipeline for different fields of view (FOVs) and image qualities. Performance was assessed on 2 external datasets—the Stanford University Network for Diagnosis of ROP (SUNDROP) cohort in the US and the Aravind Eye Care Systems (AECS) cohort in India—with the aim to assess the model's effectiveness for autonomously detecting mtmROP and type 1 ROP.

Key Points

Question How does a fully autonomous artificial intelligence system perform in identifying more-than-mild retinopathy of prematurity (mtmROP) and type 1 ROP?

Findings In this diagnostic study, the performance of an artificial intelligence system, which was trained and calibrated using 2530 examinations from 843 infants in the i-ROP study, had more than 80% sensitivity and specificity for mtmROP and 100% sensitivity for type 1 ROP in 2 large external ROP programs (SUNDROP and AECS), with potential physician workload reductions of 80% in both populations.

Meaning While not available for clinical practice settings at this time, these results provide evidence that autonomous ROP screening may be effective in ROP telemedicine programs, without substantial risk of missing severe ROP.

Methods

The methods are described in 2 steps: optimization of the existing model using data from the i-ROP study and external validation of the optimized (locked) model using data from the SUNDROP and AECS cohorts. This diagnostic study used retrospective data and adhered to the Standards for Reporting of Diagnostic Accuracy (STARD) reporting guideline.[26] The i-ROP imaging study was approved by the institutional review boards at the coordinating center (Oregon Health & Science University) and at each of 8 North American study centers.[16] Informed written consent was obtained from guardians of all enrolled infants. Analysis of SUNDROP data was approved by the Stanford University School of Medicine Institutional Review Board and analysis of Retinopathy of Prematurity Eradication-Save Our Sight (ROPE-SOS) data was approved by the institutional review board at AECS, both under a waiver of consent for analysis of retrospective data. All institutions abided by the Declaration of Helsinki. Participants did not receive a stipend or any other incentive to participate.

Optimization of i-ROP DL Algorithm and Image Processing Pipeline

Data Collection and Partitioning

From January 2012 to July 2020, the i-ROP Consortium collected a dataset of serial retinal fundus images (RFIs) from infants who underwent routine ROP screenings (eTable 1 in Supplement 1). Images were acquired using the RetCam (Natus), and 5 standard FOVs (posterior, nasal, temporal, inferior, and superior) were captured. Bedside examinations were conducted alongside RFI analysis by 4 expert readers (S. R. O., R. V. P. C., M. F. C., and J. P. C.), who also assessed image quality. A reference standard diagnosis was formed from both methods. Images deemed not acceptable for diagnosis by the consensus of the 3 readers or showing stage 4 or 5 ROP were

excluded. Following collection, images were stratified by infant into a training dataset (635 of 843 infants [75.3%%]; 16 334 of 22 038 images [74.1%]) and test dataset (208 of 843 infants [24.7%]; 5704 of 22 038 images [25.9%]), with the training dataset further divided into 5 roughly equal cross-validation splits (eTable 2 in Supplement 1).

Optimization of VSS Network

Retinal blood vessels in RFIs were segmented into black-and-white vessel maps using a segmentation network, then used to train and validate EfficientNet-B0 models for detection of normal, pre-plus disease, and plus disease (eMethods and eFigure in Supplement 1). Training involved batch sizes of 16, early stopping at 50 epochs, a learning rate of 0.001, and a weighted random sampler, with models updating after each epoch with decreased cross-entropy loss of the validation dataset.

In total, 5 models were trained and combined to create a cross-validation ensemble, including MCD.[27,28] Whereas traditional dropout is only used during training, MCD activates dropout during inference so each forward pass of an image through a network traverses a slightly varied version of the trained model. Thus, every image was passed through each of the 5 models 5 times, resulting in 25 normal, pre-plus disease, and plus disease probabilities (P) per image, as illustrated in the eFigure in Supplement 1. These probabilities were averaged into single normal, pre-plus disease, and plus disease probabilities and converted into an image-level VSS:

$$\text{VSS} = \text{P(normal)} + 5 \times \text{P(pre-plus disease)} + 9 \times \text{P(plus disease)}$$

The VSS of all images captured from a single eye examination were averaged to create an eye-level VSS, and the greater of the 2 eye-level measurements was used as the examination-level VSS.

Calibration of Operating Point for Autonomous Use

To use the continuous VSS output for screening, we evaluated each 0.1-VSS increment for detection of mtmROP, optimizing for sensitivity and specificity of at least 80.0% with 95% CIs no lower than 75.0% and 100% sensitivity for type 1 ROP. The optimal cutoff was 3.1 or greater (sensitivity, 94.3%; 95% CI, 87.2-98.1; specificity, 80.8%; 95% CI, 77.2-84.1), with lower cutoffs providing higher sensitivity but lower specificity.

Assessment of Image Quality

For real-world datasets, it is essential to implement a workflow that can identify images of insufficient quality for analysis. For the i-ROP DL, 2 criteria are essential: images must show the posterior retina (where plus disease is diagnosed) and retinal vessels must be clearly visible for segmentation. Overly strict quality assurance risks insufficient images for analysis, while too lax assurance could lead to analyzing erroneous images (eg, anterior segment images). To meet the above-mentioned requirements, a workflow that detected the presence of optic nerves in images was implemented, which allowed further processing if present (eMethods and eFigure in Supplement 1). Both eyes were required to have at least 1 image of acceptable quality. If 1 or both eyes had no acceptable images, the examination was automatically referred for in-person examination (ie, labeled as mtmROP) and included as a positive examination in performance calculations. No examinations were excluded from analysis.

External Validation of Autonomous Screening Using an Optimized i-ROP DL System

Following training and calibration, all components of the system—assessment of image quality, vessel segmentation, VSS inference, and the mtmROP operating point—were locked for external validation.

Datasets

As part of the SUNDROP telemedicine program in the US, infants born February 2013 to July 2021 across 11 NICUs in northern California, Nevada, and Indiana were serially screened for ROP using the SUNDROP protocol and the American Academy of Pediatrics 2006 guidelines.[29,30] RFIs were captured by NICU nurses trained in using the RetCam (Natus). Five distinct FOVs were captured per eye, which were transferred to a clinician (D. M. M.) for ROP diagnosis (zone, stage, and plus disease).

The ROPE-SOS telemedicine program in India serially screened infants every week from March 2019 to December 2020, with trained technicians traveling to each of the 48 participating NICUs and using a RetCam Shuttle (Natus) or a Forus 3nethra (Forus) to capture RFIs of infants who aligned with Indian screening guidelines (born at 34 weeks' gestation or earlier and weighing 2000 g or less). RFIs of both eyes were acquired, which encompassed an anterior segment photograph and multiple FOVs. This dataset was stratified by camera manufacturer into 2 subsets: AECS-RetCam and AECS-3nethra.

Performance of the i-ROP DL System

We analyzed the performance, both at the examination and infant levels, in both datasets. Posttreatment examinations were excluded, and type 1 ROP was defined based on ETROP definitions. Area under the receiver operating characteristics curve (AUROC), sensitivity, and specificity for detection of mtmROP and type 1 ROP in the SUNDROP and AECS datasets were calculated cross-sectionally at the examination level. Sensitivity, at the infant level, for type 1 ROP was also evaluated. Finally, we evaluated the potential reduction in physician workload based on the fraction of examinations that could be read autonomously (screened negative for mtmROP) in both datasets. We assumed that eyes that accurately screened positive for mtmROP would be continuously evaluated by clinicians.

Statistical Analysis

Statistical analyses were performed using R version 4.3.0 (The R Foundation). Statistical differences in patient demographic characteristics stratified disease status as well as VSS and ROP category were determined using Kruskal-Wallis rank sum tests and post-hoc Dunn tests using the Benjamini-Hochberg procedure. 95% CIs were determined using the Clopper-Pearson method. ROC curves were compared using the DeLong test. All statistical tests were 2-sided, and *P* values were not adjusted for multiple comparisons.

Table 1. Demographic Characteristics of the Screened Population in the Stanford University Network for Diagnosis of ROP (SUNDROP) and Aravind Eye Care Systems (AECS) Cohorts

Characteristic	No. (%)			
	Total	No or mild ROP	mtmROP	Type 1 ROP
SUNDROP				
Birth weight, mean (SD), g	1311.5 (592)	1343.6 (591.5)	799.5 (286.2)	704.5 (181.3)
Gestational age, mean (SD), wk	29.3 (3.3)	29.6 (3.2)	25.8 (2.7)	24.8 (1.5)
Infants	1545 (100)	1454 (94.1)	91 (5.9)	18 (1.2)
Eyes	3089 (100)	2923 (94.6)	166 (5.4)	35 (1.1)
Examinations	6205 (100)	5892 (95.0)	313 (5.0)	19 (0.3)
Images	76 258 (100)	72671 (95.3)	3587 (4.7)	241 (0.3)
AECS (all)				
Birth weight, mean (SD), g	1734.5 (449.8)	1763.2 (443.8)	1301.1 (288)	1291.2 (280.6)
Gestational age, mean (SD), wk	33.5 (2.9)	33.7 (2.7)	30.0 (2.4)	30.0 (2.2)
Infants	2699 (100)	2531 (93.8)	168 (6.2)	68 (2.5)
Eyes	5635 (100)	5278 (93.7)	357 (6.3)	140 (2.5)
Examinations	5145 (100)	4811 (93.5)	334 (6.5)	71 (1.4)
Images	69 103 (100)	63604 (92.0)	5499 (8.0)	1224 (1.8)
AECS-RetCAM				
Birth weight, mean (SD), g	1766.1 (458.5)	1791.5 (452.0)	1293.6 (292.0)	1291.0 (258.4)
Gestational age, mean (SD), wk	33.8 (2.9)	34.0 (2.7)	30.1 (2.6)	30.2 (2.3)
Infants	1802 (100)	1710 (94.9)	92 (5.1)	33 (1.8)
Eyes	3747 (100)	3553 (94.8)	194 (5.2)	68 (1.8)
Examinations	2963 (100)	2795 (94.3)	168 (5.7)	34 (1.1)
Images	37 369 (100)	34880 (93.3)	2489 (6.7)	536 (1.4)
AECS-3nethra				
Birth weight, mean (SD), g	1657.5 (416.1)	1690.6 (412.4)	1316.9 (280.5)	1283.6 (296.5)
Gestational age, mean (SD), wk	32.9 (2.8)	33.2 (2.7)	30.1 (2.3)	29.7 (2.0)
Infants	1015 (100)	925 (91.1)	90 (8.9)	36 (3.5)
Eyes	2127 (100)	1935 (91.0)	192 (9.0)	74 (3.5)
Examinations	2182 (100)	2016 (92.4)	166 (7.6)	37 (1.7)
Images	31 734 (100)	28724 (90.5)	3010 (9.5)	688 (2.2)

Abbreviations: mtmROP, more-than-mild retinopathy of prematurity; ROP, retinopathy of prematurity.

Results

Demographic Characteristics of SUNDROP and AECS Datasets

The demographic characteristics of the SUNDROP and AECS external validation datasets are presented in **Table 1.** Consistent with known differences in the epidemiology of ROP between the US and India, the mean birth weight was 423 g (95% CI, 389-457; $P < .001$) lower and the mean gestational age was 4.2 weeks (95% CI, 4.0-4.4; $P < .001$) lower in SUNDROP compared with AECS.

Within all datasets, infants who developed mtmROP had lower birth weight and gestational age compared with infants with no or mild ROP. In SUNDROP, the mean birth weight was 544 g (95% CI, 475-613; $P < .001$) lower in those with mtmROP, and in AECS, the mean birth weight was 462 g (95% CI, 415-509) lower. In SUNDROP, the mean gestational age difference between infants with mtmROP and those with no or mild ROP was 3.9 weeks (95% CI, 3.3-4.5; $P < .001$) lower, and in AECS, it was 3.7 weeks (95% CI, 3.3-4.1) lower.

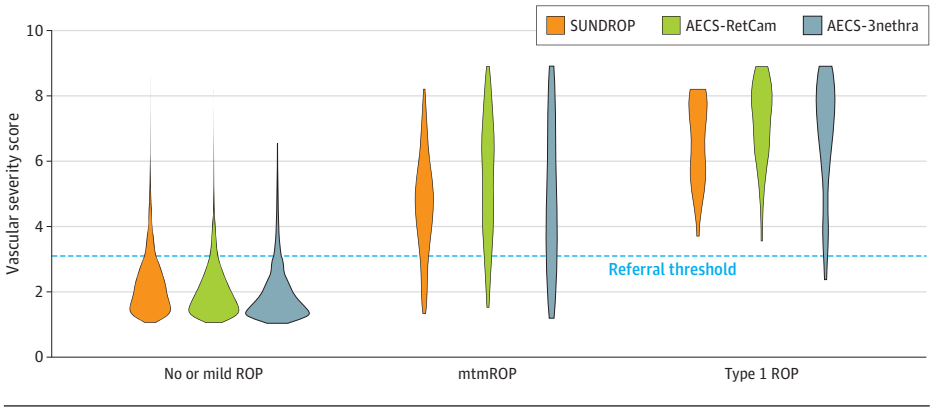
Performance of Autonomous ROP Screening at the Eye Examination and Infant Levels

Figure 1 displays the associations between the VSS, mtmROP, and type 1 ROP in both datasets. For each dataset, there were significant differences in VSS between ROP categories. Dunn tests determined that the differences for all datasets existed between no or mild ROP and mtmROP. The VSS was 2.4 points (95% CI, 2.2-2.6; $P < .001$) higher for eyes with mtmROP than those without in the SUNDROP dataset and 3.1 points (95% CI, 2.9-3.4; $P < .001$) higher in the AECS dataset. Compared with eyes with no or mild ROP, eyes with type 1 ROP had a VSS that was 4.3 points (95% CI, 3.7-5.0; $P < .001$) higher and 5.0 points (95% CI, 4.6-5.4; $P < .001$) higher in the SUNDROP and AECS datasets, respectively.

Eye Examination–Level Analysis

Cross-sectional examination-level AUROCs for detection of mtmROP and type 1 ROP is presented in **Table 2.** For both mtmROP and type 1 ROP, the AUROC was similar in both SUNDROP and AECS, with mean differences of 0.024 (95% CI, −0.006 to 0.054; $P = .12$) for mtmROP and 0.003 (95% CI,

Figure 1. Violin Plots of Vascular Severity Score vs Retinopathy of Prematurity (ROP) Screening Category



Violin plots demonstrate the association between the vascular severity score and no or mild ROP, more-than-mild ROP (mtmROP), including type 1 ROP, and type 1 ROP for the Stanford University Network for Diagnosis of ROP (SUNDROP), Aravind Eye Care Systems (AECS) captured via RetCam, and AECS captured via the Forus 3nethra camera datasets. AECS data were analyzed together as well as by camera.

Table 2. Examination-Level Diagnostic Performance for More-Than-Mild and Treatment-Requiring Retinopathy of Prematurity

	% (95% CI)			
Measure	SUNDROP	AECS (all)	AECS-RetCam	AECS-3nethra
mtmROP				
AUROC	0.896	0.920	0.947	0.898
Sensitivity	83.5 (76.6-87.7)	80.8 (76.2-84.9)	88.7 (82.9-93.1)	72.9 (65.5-79.5)
Specificity	82.2 (81.2-83.1)	87.8 (86.8-88.7)	86.9 (85.6-88.2)	89.0 (87.5-90.3)
Positive predictive value	18.1 (16.0-20.3)	31.5 (28.4-34.7)	29.0 (25.1-33.1)	35.3 (30.2-40.6)
Negative predictive value	99.1 (98.8-99.3)	98.5 (98.1-98.8)	99.2 (98.8-99.5)	97.6 (96.7-98.2)
Type 1 ROP				
AUROC	0.985	0.982	0.988	0.978
Sensitivity	100 (82.4-100)	98.6 (92.4-100)	100 (89.7-100)	97.3 (85.8-99.9)
Specificity	79.5 (78.4-80.5)	84.5 (83.5-85.5)	83.6 (82.2-84.9)	85.7 (84.1-87.1)
Positive predictive value	1.5 (0.9-2.3)	8.2 (6.4-10.2)	6.6 (4.6-9.1)	10.5 (7.5-14.2)
Negative predictive value	100 (99.9-100)	99.9 (99.9-100)	100 (99.8-100)	99.9 (99.7-100)

Abbreviations: AECS, Aravind Eye Care Systems; AUROC, area under the receiver operating characteristic curve; mtmROP, more-than-mild retinopathy of prematurity; ROP, retinopathy of prematurity; SUNDROP, Stanford University Network for Diagnosis of ROP.

−0.017 to 0.011; P = .65) for type 1 ROP. The performance on the Forus camera was slightly lower for mtmROP (mean difference, 0.049; 95% CI, 0.013-0.085; P = .007) but not for type 1 ROP (mean difference, 0.010; 95% CI, −0.005 to 0.028; P = .26). Confusion matrices for mtmROP and type 1 ROP show the number of true positives, false positives, true negatives, and false negatives (**Table 3**). In SUNDROP, 196 of 6205 examinations (3.2%; 196 total, including 19 with mtmROP and 0 with type 1 ROP) were automatically referred due to 1 or both eyes having no images of acceptable quality; in AECS, 152 of 5145 examinations (3.0%; 152 total, including 14 with mtmROP and 8 with type 1 ROP) were automatically referred. If implemented fully autonomously, the algorithm could reduce the number of telemedicine examinations requiring physician time by approximately 80% (4915 of 6205 [79.2%], 2449 of 2963 [82.7%], and 1839 of 2182 [84.3%] for the SUNDROP, AECS-RetCam, and AECS-3nethra datasets, respectively).

Patient-Level Analysis

No infants developed type 1 ROP prior to screening positive, but 1 screened negative at the examination where type 1 ROP was diagnosed (Table 3). However, the patient had already screened positive in prior weeks and thus would have been caught with continued follow-up. Specifically, this infant correctly screened positive for mtmROP at postmenstrual age of 35 weeks (VSS of 6.0) and 36 weeks (VSS4.7) when the disease was more vascularly active but did not screen positive at 37 weeks (VSS of 2.4) when the clinician decided to treat (**Figure 2**). Notably, the examination at 37 weeks' postmenstrual age had better visualization of the peripheral pathology, albeit reduced dilation and tortuosity. This case underscores 4 key aspects: the role of FOV in ROP diagnosis, the subjectivity in diagnosing plus disease, the vascular phases of ROP, and the need for clinical safeguards for implementing autonomous ROP screening.

Discussion

In this external validation study of autonomous ROP screening using AI in 2 large patient groups in the US and India, no infants were diagnosed with type 1 ROP before screening positive by an autonomous AI system, and approximately 80% of examinations screened negative in both populations. These results suggest that AI could significantly aid ROP prevention efforts globally, including in low-income and middle-income countries. However, the

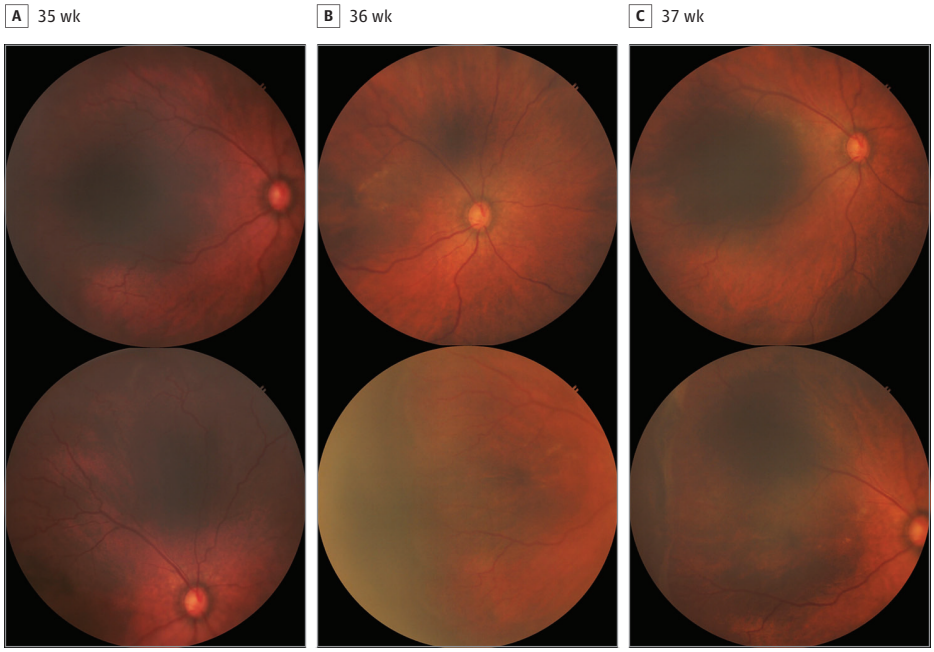
Table 3. Examination-Level Confusion Matrix for Clinical Diagnosis vs Autonomous Diagnosis

	Clinical diagnosis											
	SUNDROP				AECS-RetCam				AECS-3nethra			
Autonomous diagnosis	Less than mtmROP	mtmROP	Less than type 1 ROP	Type 1	Less than mtmROP	mtmROP	Less than type 1 ROP	Type 1	Less than mtmROP	mtmROP	Less than type 1 ROP	Type 1
Less than mtmROP	4869	46	4915	0	2430	19	2449	0	1794	45	1838	1[a]
mtmROP	1057	233	1271	19	365	149	480	34	222	121	307	36

Abbreviations: AECS, Aravind Eye Care Systems; mtmROP, more-than-mild retinopathy of prematurity; ROP, retinopathy of prematurity; SUNDROP, Stanford University Network for Diagnosis of ROP.

[a] See eFigure in Supplement 1 for details.

Figure 2. Examination With Missed Treatment-Requiring Type 1 Retinopathy of Prematurity (ROP) in the Aravind Dataset With the Forus 3nethra Camera



A posterior image and the best-available image of the temporal pathology in the right eye are shown for each week. This infant had 2 true-positive examinations for more-than-mild ROP at postmenstrual age 35 and 36 weeks (vascular severity scores of 6.0 and 4.7, respectively). At postmenstrual age 37 weeks, the vascular severity score had improved to 2.4, and the eye was below the referral threshold. However, the pathology was better visualized, and the physician decided to treat due to the appearance of the peripheral pathology, which was becoming fibrotic. With common-sense implementation, infants who screen positive for (and are confirmed to have) more-than-mild ROP would continue to be monitored until the disease progresses or regresses, and this infant would not have been missed. Thus, infant-level sensitivity for type 1 ROP was 100%. This case also nicely illustrates the pathophysiology of ROP that progresses through a vascularly active phase and then becomes cicatricial with disease regression, which has important implications for screening based solely using vascular severity score.

implementation challenges of this technology remain. Although telemedicine has proven effective in ROP screening, with successful large-scale programs in the US and India, the feasibility of deploying costly cameras varies (ie, it is more viable in densely populated areas where a camera serves multiple NICUs). Accordingly, UNICEF has highlighted the need for low-cost cameras as a global health priority.[31,32]

Currently, the most affordable option that provides sufficient FOV for ROP screening is the Forus 3nethra camera, developed and manufactured in India. It is encouraging that the overall performance for detecting type 1 ROP was similar (Table 2), which suggests that integrating and deploying this technology with the Forus is possible even though the algorithm was developed using RetCam data. Further prospective validation may better define whether performance on the 3nethra camera could be optimized with a refined algorithm or different operating point (post hoc sensitivity analysis available in eTable 3 in Supplement 1). Defining and solving implementation barriers to deploying ROP telemedicine programs ought to be a key next step to maximally use the potential benefits of AI.[32]

There are several potential advantages to incorporating autonomous AI-based screening into ROP telemedicine programs, besides the potential 80% workload reduction for telemedicine graders. Autonomous screening can provide real-time feedback to NICU teams and families, facilitating educational efforts aimed at maximizing follow-up rates up on discharge. Previous work has also demonstrated other clini-

cal benefits of using AI-based assessment of ROP severity, including longitudinal disease monitoring before and after treatment.[25,33-35] That is, for eyes that screen positive, the VSS can be used as an objective biomarker to aid clinicians in identifying infants' progression to type 1 ROP. The VSS concept has also been integrated into clinical risk models, which can further improve the specificity of disease screening and reduce the number of required examinations in low-risk neonates; this has been validated in the US, India, Nepal, and Mongolia.[19,20] Finally, data suggest that AI-based assessment of ROP severity at the NICU (rather than individual) level may identify NICUs with higher-than-expected ROP severity and be useful for assessment of interventions for primary prevention.[21,36-38]

Limitations

This study has limitations. The inherent obstacle in implementing autonomous ROP screening is the risk of missing type 1 ROP and the subsequent risk of visual loss. This is not a trivial risk, and therefore, discussion around when, if, and how autonomous ROP screening could be implemented—and with what safeguards—is essential. This must be balanced against the status quo: roughly 30 000 to 50 000 infants lose vision from ROP worldwide, primarily due to absent, late, or ineffective screening. In this analysis of 2 large ROP programs, no infants with type 1 ROP were missed; however, no retinal detachments (stage 4 or 5 ROP) were observed. As a result, we are unable to evaluate what would happen in the rare case a retinal detachment was present on the first examination. Regardless, clinical safeguards, such as the requirement that all first examinations are reviewed manually, could ensure that any pathology other than what the model was trained for (eg, stage 4 or 5 ROP, retinoblastoma) are captured. Another important consideration is how and when to discontinue screening, which will need to be evaluated at each health care system where this is deployed. Additionally, we recognize that adding pre-plus disease to the definition of mtmROP may be controversial, but preliminary results suggested there was no difference in AUROC for either outcome (mtmROP or type 1 ROP) in either dataset. The advantage of a VSS-based cutoff is that the operating point can be adjusted based on any outcome definition, be it the traditional definition of referral warranted, stage 3 ROP, type 2 ROP, or, as in this study, mtmROP. In all cases, it is important to consider appropriate follow-up intervals to minimize the risk that patients with mtmROP (by any definition) progress to the point of treatment before their next screening examination.

Conclusions

Although the current analysis has the advantage of being based on real-world telemedicine datasets, which have heterogeneity in image acquisition practices, clinical diagnosis, and patient demographic characteristics, it should be noted that this system is not available for clinical practice at this time. The major limitation to this work is that scaling these results into telemedicine programs requires investment in digital cameras that may be cost-prohibitive, and the algorithm would need to be validated and/or adapted to work with images from future cameras. As part of a comprehensive ROP program with optimization of primary prevention, and available ROP treatment, autonomous ROP screening may play a role in reducing the incidence of ROP-related blindness.

ARTICLE INFORMATION

Accepted for Publication: December 15, 2023.

Published Online: March 7, 2024. doi:10.1001/jamaophthalmol.2024.0045

Author Affiliations: Casey Eye Institute, Oregon Health & Science University, Portland (Coyner, Murickan, Oh, Young, Ostmo, Campbell); Ophthalmology, University of Colorado School of Medicine, Aurora (Singh, Kalpathy-Cramer); Illinois Eye and Ear Infirmary, University of Illinois at Chicago (Chan); Byers Eye Institute, Department of Ophthalmology, Stanford University School of Medicine, Palo Alto, California (Moshfeghi); Pediatric Retina and Ocular Oncology, Aravind Eye Hospital, Coimbatore, India (Shah, Venkatapathy); National Eye Institute, National Institutes of Health, Bethesda, Maryland (Chiang); National Library of Medicine, National Institutes of Health, Bethesda, Maryland (Chiang).

Author Contributions: Drs Coyner and Campbell had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: Coyner, Singh, Chiang, Kalpathy-Cramer, Campbell.
Acquisition, analysis, or interpretation of data: Coyner, Murickan, Oh, Young, Ostmo, Chan, Moshfeghi, Shah, Narendran, Kalpathy-Cramer, Campbell.
Drafting of the manuscript: Coyner, Murickan, Oh, Campbell.
Critical review of the manuscript for important intellectual content: Coyner, Murickan, Young, Ostmo, Singh, Chan, Moshfeghi, Shah, Narendran, Chiang, Kalpathy-Cramer, Campbell.
Statistical analysis: Coyner, Murickan, Oh.
Obtained funding: Kalpathy-Cramer, Campbell.
Administrative, technical, or material support: Young, Ostmo, Chan, Shah, Narendran, Kalpathy-Cramer, Campbell.
Supervision: Chan, Moshfeghi, Chiang, Kalpathy-Cramer, Campbell.

Conflict of Interest Disclosures: Dr Coyner reported grants from the National Eye Institute, Research to Prevent Blindness, Malcolm Marquis Innovation Fund, US Agency for International Development, and Seva Foundation during the conduct of the study as well as personal fees from Siloam Vision outside the submitted work. Dr Chan reported grants from the National Institutes of Health and Research to Prevent Blindness during the conduct of the study; personal fees from Alcon, Genentech, Ocular Therapeutix, and Regeneron; and owns equity in Siloam Vision outside the submitted work. Dr Moshfeghi reported grants from Genentech during the conduct of the study; grants from Research to Prevent Blindness and the National Eye Institute; personal fees from Akebia, Regeneron, Genentech, Alexion, Aspire Pharma, Affamed, Feliqs, Icon Clinical Research, Slack, University of Miami, and Vanotech Chengdu; owns equity in Visunex, Ainsly, Plenoptika, Pykus, DSENTZ, PROMISIGHT, and PR3VENT; is an unpaid advisor for Clinical Trials Research Group and Insite DME; and serves on the board of directors for DSENTZ, PROMISIGHT, and PR3VENT outside the submitted work. Dr Chiang reported grants from the National Institutes of Health, National Science Foundation, and Genentech as well as personal fees from Novartis during the conduct of the study. Dr Kalpathy-Cramer reported grants from GE Heathcare and Genentech outside the submitted work; has a patent for a retinopathy of prematurity deep learning algorithm pending; and consults for Siloam Vision. Dr Campbell reported grants from the National Eye Institute and Research to Prevent Blindness; personal fees from Boston AI during the conduct of the study; and grants from Genentech outside the submitted work; and owns equity in Siloam Vision. No other disclosures were reported.

Funding/Support: This work was supported by grants R01 EY019474, R01 EY031331, R21 EY031883, and P30 EY010572 from the National Institutes of Health; unrestricted departmental funding and a Career Development Award from Research to Prevent Blindness (Dr Campbell); the Malcolm Marquis Innovation Fund; and with support from the US Agency for International Development, the Seva Foundation, and the intramural research program of the National Eye Institute.

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Sharing Statement: See Supplement 2.

REFERENCES

1. Chiang MF, Quinn GE, Fielder AR, et al. International Classification of Retinopathy of Prematurity, Third Edition. *Ophthalmology*. 2021; 128(10):e51-e68. doi:10.1016/j.ophtha.2021.05.031

2. Sabri K, Ells AL, Lee EY, Dutta S, Vinekar A. Retinopathy of prematurity: a global perspective and recent developments. *Pediatrics*. 2022;150(3): e2021053924. doi:10.1542/peds.2021-053924

3. Blencowe H, Moxon S, Gilbert C. Update on blindness due to retinopathy of prematurity globally and in India. *Indian Pediatr*. 2016;53(suppl 2):S89-S92.

4. Digital Diagnostics. Digital Diagnostics and Orbis International announce study to help save sight in Bangladesh. Accessed September 20, 2023. https://www.digitaldiagnostics.com/newsroom/digital-diagnostics-and-orbis-international-announce-study-to-help-save-sight-in-bangladesh/

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

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

7. 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

8. 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): 41-49. doi:10.4103/0301-4738.126178

9. Vinekar A, Mangalesh S, Jayadev C, Gilbert C, Dogra M, Shetty B. Impact of expansion of telemedicine screening for retinopathy of prematurity in India. *Indian J Ophthalmol*. 2017;65 (5):390-395. doi:10.4103/ijo.IJO_211_17

10. Shah PK, Ramya A, Narendran V. Telemedicine for ROP. *Asia Pac J Ophthalmol (Phila)*. 2018;7(1): 52-55.

11. Shah PK, Narendran V, Kalpana N. Evolution of ROP screening at Aravind Eye Hospital, Coimbatore—lessons learnt and the way ahead. *Community Eye Health*. 2018;31(101):S23-S24.

12. Early Treatment Diabetic Retinopathy Study Research Group. Fundus photographic risk factors for progression of diabetic retinopathy. ETDRS report number 12. *Ophthalmology*. 1991;98(5) (suppl):823-833. doi:10.1016/S0161-6420(13)38014-2

13. Good WV; Early Treatment for Retinopathy of Prematurity Cooperative Group. Final results of the Early Treatment for Retinopathy of Prematurity (ETROP) randomized trial. *Trans Am Ophthalmol Soc*. 2004;102:233-248.

14. Good WV, Hardy RJ, Dobson V, et al; Early Treatment for Retinopathy of Prematurity Cooperative Group. The incidence and course of retinopathy of prematurity: findings from the early treatment for retinopathy of prematurity study. *Pediatrics*. 2005;116(1):15-23. doi:10.1542/peds.2004-1413

15. Gupta MP, Chan RVP, Anzures R, Ostmo S, Jonas K, Chiang MF; Imaging & Informatics in ROP Research Consortium. Practice patterns in retinopathy of prematurity treatment for disease milder than recommended by guidelines. *Am J Ophthalmol*. 2016;163:1-10. doi:10.1016/j.ajo.2015.12.005

16. Brown JM, Campbell JP, Beers A, et al; Imaging and Informatics in Retinopathy of Prematurity (i-ROP) Research Consortium. 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

17. Campbell JP, Kim SJ, Brown JM, et al; Imaging and Informatics in Retinopathy of Prematurity Consortium. Evaluation of a deep learning-derived quantitative retinopathy of prematurity severity scale. *Ophthalmology*. 2021;128(7):1070-1076. doi:10.1016/j.ophtha.2020.10.025

18. Campbell JP, Chiang MF, Chen JS, et al; Collaborative Community in Ophthalmic Imaging Executive Committee and the Collaborative Community in Ophthalmic Imaging Retinopathy of Prematurity Workgroup. Artificial intelligence for retinopathy of prematurity: validation of a vascular severity scale against international expert diagnosis. *Ophthalmology*. 2022;129(7):e69-e76. doi:10.1016/j.ophtha.2022.02.008

19. Coyner AS, Chen JS, Singh P, et al. Single-examination risk prediction of severe retinopathy of prematurity. *Pediatrics*. 2021;148(6): e2021051772. doi:10.1542/peds.2021-051772

20. 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):791-798. doi:10.1001/jamaophthalmol.2022.2135

21. Campbell JP, Singh P, Redd TK, et al. Applications of artificial intelligence for retinopathy of prematurity screening. *Pediatrics*. 2021;147(3): e2020016618. doi:10.1542/peds.2020-016618

22. Greenwald MF, Danford ID, Shahrawat M, et al. Evaluation of artificial intelligence-based telemedicine screening for retinopathy of prematurity. *J AAPOS*. 2020;24(3):160-162. doi:10.1016/j.jaapos.2020.01.014

23. 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

24. Redd TK, Campbell JP, Brown JM, et al; Imaging and Informatics in Retinopathy of Prematurity (i-ROP) Research Consortium. Evaluation of a deep learning image assessment system for detecting severe retinopathy of prematurity. *Br J Ophthalmol*. 2018;bjophthalmol-2018-313156.

25. Taylor S, Brown JM, Gupta K, et al; Imaging and Informatics in Retinopathy of Prematurity Consortium. Monitoring disease progression with a quantitative severity scale for retinopathy of prematurity using deep learning. *JAMA Ophthalmol*. 2019;137(9):1022-1028. doi:10.1001/jamaophthalmol.2019.2433

26. Bossuyt PM, Reitsma JB, Bruns DE, et al; STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015;351:h5527. doi:10.1136/bmj.h5527

27. Lemay A, Hoebel K, Bridge CP, et al. Improving the repeatability of deep learning models with Monte Carlo dropout. *NPJ Digit Med*. 2022;5(1):174. doi:10.1038/s41746-022-00709-3

28. Ahmed SR, Befano B, Lemay A, et al. Reproducible and clinically translatable deep neural networks for cancer screening. *Res Square*. Preprint posted online March 3, 2023. doi:10.21203/rs.3.rs-2526701/v1

29. Section on Ophthalmology American Academy of Pediatrics; American Academy of Ophthalmology; American Association for Pediatric Ophthalmology and Strabismus. Screening examination of premature infants for retinopathy of prematurity. *Pediatrics*. 2006;117(2):572-576. doi:10.1542/peds.2005-2749

30. Fijalkowski N, Zheng LL, Henderson MT, et al. Stanford University Network for Diagnosis of Retinopathy of Prematurity (SUNDROP): five years of screening with telemedicine. *Ophthalmic Surg Lasers Imaging Retina*. 2014;45(2):106-113. doi:10.3928/23258160-20140122-01

31. 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):582-588. doi:10.1001/jamaophthalmol.2023.1466

32. Kirby RP, Malik ANJ, Palamountain KM, Gilbert CE; TPP Survey and Meeting Participants Collaborator Group. 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

33. Bellsmith KN, Brown J, Kim SJ, et al. Aggressive posterior retinopathy of prematurity: clinical and quantitative Imaging features in a large North American cohort. *Ophthalmology*. 2020;127(8): 1105-1112. doi:10.1016/j.ophtha.2020.01.052

34. Gupta K, Campbell JP, Taylor S, et al; Imaging and Informatics in Retinopathy of Prematurity Consortium. A quantitative severity scale for retinopathy of prematurity using deep learning to monitor disease regression after treatment. *JAMA Ophthalmol*. 2019;137(9):1029-1036. doi:10.1001/jamaophthalmol.2019.2442

35. Eilts SK, Pfeil JM, Poschkamp B, et al; Comparing Alternative Ranibizumab Dosages for Safety and Efficacy in Retinopathy of Prematurity (CARE-ROP) Study Group. Assessment of retinopathy of prematurity regression and reactivation using an artificial intelligence-based vascular severity score. *JAMA Netw Open*. 2023;6(1): e2251512. doi:10.1001/jamanetworkopen.2022.51512

36. 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):837-843. doi:10.1016/j.ophtha.2023.03.026

37. Hanif A, Lu C, Chang K, et al; Imaging and Informatics in Retinopathy of Prematurity Consortium. Federated learning for multicenter collaboration in ophthalmology: implications for clinical diagnosis and disease epidemiology. *Ophthalmol Retina*. 2022;6(8):650-656. doi:10.1016/j.oret.2022.03.005

38. Lu C, Hanif A, Singh P, et al; Imaging and Informatics in Retinopathy of Prematurity Consortium Members of the Imaging and Informatics in Retinopathy of Prematurity research consortium are as follows. Federated learning for multicenter collaboration in ophthalmology: improving classification performance in retinopathy of prematurity. *Ophthalmol Retina*. 2022;6(8):657-663. doi:10.1016/j.oret.2022.02.015