

# Applications of Artificial Intelligence for Retinopathy of Prematurity Screening

J. Peter Campbell, MD, MPH,<sup>a,\*</sup> Praveer Singh, PhD,<sup>b,\*</sup> Travis K. Redd, MD, MPH,<sup>a</sup> James M. Brown, PhD,<sup>c</sup> Parag K. Shah, DNB,<sup>d</sup> Prema Subramanian, DNB,<sup>d</sup> Renu Rajan, MD,<sup>e</sup> Nita Valikodath, MD,<sup>f</sup> Emily Cole, MD,<sup>g</sup> Susan Ostmo, MS,<sup>a</sup> R.V. Paul Chan, MD,<sup>f</sup> Narendran Venkatapathy, DNB,<sup>d</sup> Michael F. Chiang, MD,<sup>a,g</sup> Jayashree Kalpathy-Cramer, PhD<sup>b</sup>

abstract

**OBJECTIVES:** Childhood blindness from retinopathy of prematurity (ROP) is increasing as a result of improvements in neonatal care worldwide. We evaluate the effectiveness of artificial intelligence (AI)-based screening in an Indian ROP telemedicine program and whether differences in ROP severity between neonatal care units (NCUs) identified by using AI are related to differences in oxygen-titrating capability.

**METHODS:** External validation study of an existing AI-based quantitative severity scale for ROP on a data set of images from the Retinopathy of Prematurity Eradication Save Our Sight ROP telemedicine program in India. All images were assigned an ROP severity score (1–9) by using the Imaging and Informatics in Retinopathy of Prematurity Deep Learning system. We calculated the area under the receiver operating characteristic curve and sensitivity and specificity for treatment-requiring retinopathy of prematurity. Using multivariable linear regression, we evaluated the mean and median ROP severity in each NCU as a function of mean birth weight, gestational age, and the presence of oxygen blenders and pulse oxygenation monitors.

**RESULTS:** The area under the receiver operating characteristic curve for detection of treatment-requiring retinopathy of prematurity was 0.98, with 100% sensitivity and 78% specificity. We found higher median (interquartile range) ROP severity in NCUs without oxygen blenders and pulse oxygenation monitors, most apparent in bigger infants (>1500 g and 31 weeks' gestation: 2.7 [2.5–3.0] vs 3.1 [2.4–3.8];  $P = .007$ , with adjustment for birth weight and gestational age).

**CONCLUSIONS:** Integration of AI into ROP screening programs may lead to improved access to care for secondary prevention of ROP and may facilitate assessment of disease epidemiology and NCU resources.



<sup>a</sup>Department of Ophthalmology, Casey Eye Institute and <sup>d</sup>Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon; <sup>b</sup>Athinoula A. Martinos Center for Biomedical Imaging and Department of Radiology, Massachusetts General Hospital, Charlestown, Massachusetts; <sup>c</sup>Department of Computer Science, University of Lincoln, Lincoln, United Kingdom; <sup>d</sup>Pediatric Retina and Ocular Oncology Division, Aravind Eye Hospital, Coimbatore, India; <sup>e</sup>Department of Retina and Vitreous, Aravind Eye Hospital, Madurai, India; and <sup>f</sup>Department of Ophthalmology and Visual Sciences, Illinois Eye and Ear Infirmary and University of Illinois at Chicago, Chicago, Illinois

\*Contributed equally as co-first authors

Drs Campbell, Chiang, Kalpathy-Cramer, and Chan were involved in all aspects of the study, including conceptualizing and designing the study, analyzing the data, drafting the initial manuscript, and reviewing and revising the manuscript; Drs Shah, Subramanian, and Venkatapathy developed the original data set, contributed to analysis of data, and critically revised the manuscript; (Continued)

**WHAT'S KNOWN ON THIS SUBJECT:** Childhood blindness from retinopathy of prematurity (ROP) is increasing in many parts of the world as a result of improved neonatal survival after preterm birth in the setting of underresourced health systems with limited ability to monitor oxygen and provide screening.

**WHAT THIS STUDY ADDS:** Evaluation of an artificial intelligence system in an ROP telescreening program in India revealed high accuracy for detection of sight-threatening ROP, and by using a quantitative scale, it was found that units with oxygen-monitoring capability had lower ROP severity.

**To cite:** Campbell JP, Singh P, Redd TK, et al. Applications of Artificial Intelligence for Retinopathy of Prematurity Screening. *Pediatrics*. 2021;147(3):e2020016618

The incidence of retinopathy of prematurity (ROP) worldwide is increasing because health systems are improving in low- and middle-income countries (LMICs).<sup>1-3</sup> One hundred years ago, prematurely born infants died before ROP was recognized clinically. As modern neonatal care developed in the mid-20th century,<sup>4</sup> but before our understanding of the relationship between oxygen and ROP,<sup>4,5</sup> ROP emerged and quickly caused an epidemic of blindness in the United States and Europe.<sup>3</sup> After the implementation of stricter oxygen-monitoring protocols, the incidence of blindness from ROP fell dramatically before modern disease classification and before any clinical trials for ROP treatment.<sup>4,6,7</sup> Since then, the epidemiology of the disease has varied geographically and over time primarily on the basis of 2 factors: the survival of increasingly premature infants and the implementation of strict oxygen-monitoring protocols.<sup>3,8</sup>

Over the last few decades, the incidence of blindness from ROP has been rising rapidly in many LMICs because of the same epidemiological factors.<sup>1,3</sup> As neonatal care unit (NCU) capacity expands in countries such as India,<sup>1</sup> which leads the world in premature births,<sup>9</sup> these new NCUs must balance the primary goal of reducing mortality with the secondary goal of minimizing all of the consequent morbidities of premature birth, including ROP. Many NCUs lack the material resources, such as oxygen blenders and pulse oxygenation monitors, necessary for primary prevention of ROP.<sup>1</sup> This has 2 effects: (1) both the incidence and the severity of ROP are greater for a given degree of prematurity in LMICs and (2) a greater number of infants need to be screened because in the setting of unmonitored oxygen, even mildly premature infants remain at risk for blinding ROP, which is no longer true in the United States.<sup>10,11</sup> As a result, ~20 000 infants go blind

and tens of thousands more develop severe visual impairment every year, primarily in LMICs.<sup>12</sup>

Artificial intelligence (AI)-facilitated disease screening has been proposed for a number of ophthalmic diseases and would have the greatest impact in these regions where the disease burden far outweighs the existing capacity for screening.<sup>13</sup> ROP severity is diagnosed on the basis of 3 clinical features: how much of the retina is vascularized (zone), the degree of pathology at the vascular-avascular border (stage), and the degree of dilation and tortuosity of the posterior retinal vessels (plus disease).<sup>6,14,15</sup> Brown et al<sup>16</sup> reported the expert-level automated performance of an AI algorithm for the diagnosis of plus disease that was developed for a North American population of premature infants, and extensions of this work have revealed that the same algorithm (Imaging and Informatics in Retinopathy of Prematurity Deep Learning [i-ROP DL]) can be used to assign a quantitative severity score (1–9) on the basis of a single photograph that correlates with the full zone, stage, and plus disease classification.<sup>17-20</sup> Two retrospective evaluations of this system in a North American population revealed high sensitivity for detection of treatment-requiring retinopathy of prematurity (TR-ROP).<sup>19,21</sup> However, it is well recognized that AI algorithms that reveal high diagnostic accuracy in research data sets or one demographic population may not generalize to other populations, and there has been little real-world evaluation of AI-based screening in LMICs.<sup>17,22,23</sup>

Furthermore, the ability of AI to quantitatively assess ROP severity in individuals opens the door for quantification of ROP severity in groups of individuals, such as comparing rates of ROP between NCUs within a geographic region or over time. There are several potential

reasons that NCUs may have different levels of ROP: (1) differences in neonatal mortality may affect whether infants survive to the point of ROP screening, which generally occurs after 3 to 4 weeks of life; (2) differences in oxygen management between NCUs may lead to higher rates of ROP; and (3) tertiary referral centers may provide care for higher-risk patients on the basis birth weight and gestational age. In this study, we retrospectively evaluate the diagnostic performance of this AI-based ROP severity score in a real-world data set from an Indian ROP telescreening program. We further evaluate the hypothesis that this severity score could be used as a quantitative metric at the NCU level to assess differences in ROP severity and possibly differences in neonatal care in this population.

## METHODS

### Data Set

The data set was obtained from images collected through the Retinopathy of Prematurity Eradication Save Our Sight telemedicine program at the Aravind Eye Hospital in Coimbatore, India, between August 2015 and October 2017. Each NCU was given a unique study identification on the basis of the order it appeared in the database. Trained technicians who traveled weekly to each NCU used the Retcam Shuttle (Natus Medical Incorporated, Pleasanton, CA) to obtain fundus photographs for each infant who met Indian screening guidelines (born at  $\leq 34$  weeks and weighing 2000 g).<sup>11</sup> Images of each eye were obtained, including an anterior segment photograph and multiple views of the posterior retina (containing the optic nerve) as well as the retinal periphery. The data set also included demographic characteristics associated with ROP, including birth weight, gestational age, and

postmenstrual age at time of initial ROP examination.

### Consensus Diagnosis of Plus Disease

Each eye examination in the data set was classified as plus, pre-plus, or no plus by the original clinicians (P.K.S. or P.S.) via telemedicine and subsequently by a trained study coordinator (S.O.) who was masked to the clinical diagnosis. Disagreements were adjudicated by 2 additional ROP clinicians (J.P.C. and M.F.C.). All graders were masked to the deep learning results. Thus, each eye examination received a label of plus, pre-plus, or normal on the basis of a majority vote among 3 graders, which served as the ground truth for evaluation of the i-ROP DL system.

### Image Preprocessing

Because the i-ROP DL system is only used to evaluate images of the posterior retina, we developed a preprocessing step to exclude images that did not have an optic nerve present within the image (such as anterior segment images or far peripheral retinal images). We trained an optic disc segmenter using a U-Net<sup>24</sup> implemented in Keras with TensorFlow.<sup>25</sup> Our model was trained by using Adam optimization (with  $\beta_1 = .9$  and  $\beta_2 = .999$ ) for 200 epochs with a batch size of 8 and an initial learning rate of 0.05. Images were preprocessed by rescaling to size  $480 \times 480$  and applying various data augmentation techniques, such as horizontal and vertical flips, rotations, etc. From the original data set of 8567 images, 4383 nonposterior retinal images were excluded by this process. An additional 9 nonposterior pole images were excluded on manual review during the image grading process.

### Comparison of Clinical Diagnosis With Severity Score

All of the remaining images were then analyzed by the i-ROP DL system and assigned a plus disease classification (plus, pre-plus, or normal). Because

the output of the i-ROP DL system is at the image level, but clinical diagnosis is performed at the eye level, the mode classification was used for the deep learning label for each eye examination, with the higher classification chosen in the case of an even split. Each image was also assigned an AI severity score from 1 to 9 by using methods previously published.<sup>19,20</sup> We averaged the AI severity score for all posterior retinal images for each eye examination and compared the AI severity score to the consensus plus disease diagnosis using analysis of variance. We then calculated the area under the receiver operating characteristic (AUROC) curve, sensitivity, specificity, and positive predictive value (PPV) for the detection of plus disease (all infants diagnosed with plus disease require treatment) and any infant who was deemed to require treatment by the clinician.<sup>15</sup> In this article, we report the performance of the AI system for detection of TR-ROP, as determined by the clinician.

### Quantitative Evaluation of ROP Severity at the NCU Level

We identified all NCUs that had screened  $>5$  infants during the study period, which was an arbitrary number chosen to limit analysis of NCUs that were only recently added to the screening program. To generate a quantitative metric of ROP severity for each NCU, we averaged the individual-level severity scores from the first examination for all infants within that NCU during the study period. Per Indian ROP screening guidelines, the first examination occurs at 31 weeks' gestation or 4 weeks of life, whichever is later.<sup>11</sup> This produced a number, from 1 to 9, for each NCU in the data set, which we labeled "NCU-level ROP severity."

### Assessment of Oxygen-Management Capability

Ophthalmic technicians who were masked to the results of the AI output surveyed neonatal nurses and

documented the characteristics of the NCUs, including the number of beds, whether they were government funded or private, and whether each NCU had oxygen blenders and pulse oxygenation monitors. We split the cohort into 2 groups: those with oxygen blenders and pulse oxygenation monitors for every bed in the unit (both necessary for best practices regarding oxygen management) and those without one or the other or both. We then performed a multivariable linear regression of NCU-level ROP severity as a function of the mean NCU birth weight and gestational age, government versus private status, and oxygen-monitoring capability.

This study was conducted in accordance with Health Insurance Portability and Accountability Act guidelines, and institutional review board approval was obtained at both Oregon Health & Science University and the Aravind Eye Hospital under a waiver of consent for retrospective evaluation of clinical and imaging data obtained as part of routine clinical care. Statistical analysis was performed by using Stata MP 13 (Stata Corp, College Station, TX) and R (R Foundation, Vienna, Austria). A  $P$  value  $<.05$  was considered statistically significant.

## RESULTS

### Description of Data Set

After preprocessing with removal of nonposterior retinal images and duplicates, there were 4175 unique images from 1253 eye examinations of 363 infants from 32 NCUs for the eye examination level analysis. Two hundred three infants were male (56%). The mean  $\pm$  SD gestational age was  $31 \pm 4$  weeks. The mean  $\pm$  SD birth weight was  $1405 \pm 390$  g, and the mean  $\pm$  SD postmenstrual age was  $38 \pm 5$  weeks. The results of 762 examinations (61%) were classified as no plus disease, 436 (35%) were classified as pre-plus

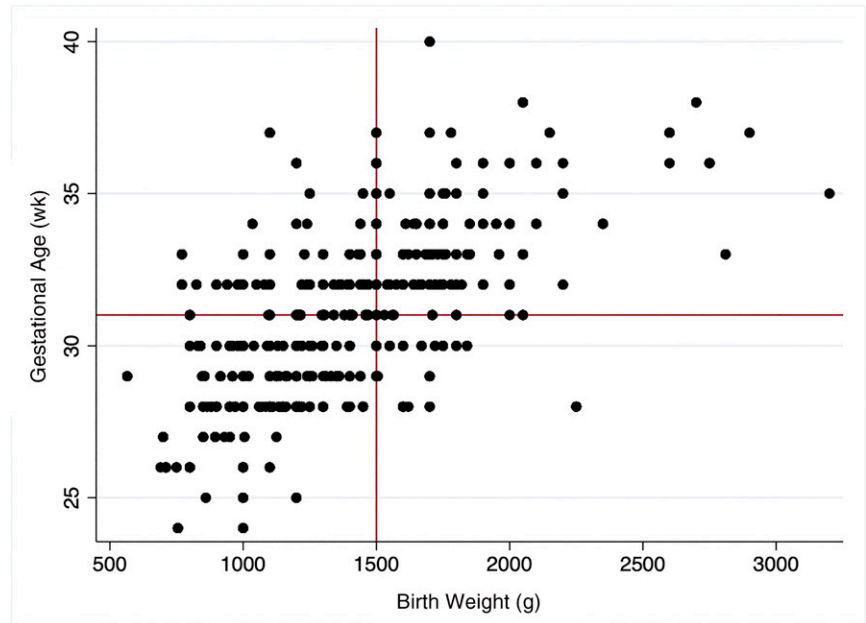
disease, and 55 (4%) were classified as plus disease by consensus diagnosis. All infants diagnosed with TR-ROP had a consensus diagnosis of plus disease in this data set. Four hundred thirteen of the 1253 eye examinations (33%) occurred in infants who would not meet ROP screening guidelines in the United States. Figure 1 reveals the distribution of infants in this data set by birth weight and gestational age.

### Comparison of Clinical Diagnosis With Severity Score

Figure 2A reveals the receiver operating characteristic curve, and Fig 2B reveals the confusion matrix for the output of the i-ROP DL system compared to the consensus diagnosis. The i-ROP DL system agreed with the 3-level consensus diagnosis of plus disease (no plus versus pre-plus versus plus) in 939 of 1253 (75%) examinations. Figure 2C reveals a box plot of the median AI severity score for each eye examination compared to the consensus diagnosis of normal, pre-plus, or plus disease. The median (interquartile range [IQR]) AI severity score of images with no plus disease according to the telemedicine grading was 1.8 (1.3–2.4), compared to 3.5 (2.4–4.3) for pre-plus disease and 6.2 (5.3–6.9) for plus disease ( $P < .001$ ). The AUROC was 0.98 for detection of TR-ROP by using the 1 to 9 severity score. Optimizing for sensitivity, the system performed optimally at 3.5, with 100% sensitivity and 78% specificity for detection of treatment-requiring disease (Youden's index operating point was 3.6, with 98.2% sensitivity and 79.7% specificity) and a PPV of 12% for treatment-requiring disease and 74% for pre-plus or worse disease.<sup>19,20</sup>

### Quantitative Evaluation of ROP Severity at the NCU Level

During the study period, 14 of the 32 NCUs screened >5 infants, resulting in 3928 images from 583 initial eye examinations (325 infants) for the NCU-level analysis. NCU-level ROP



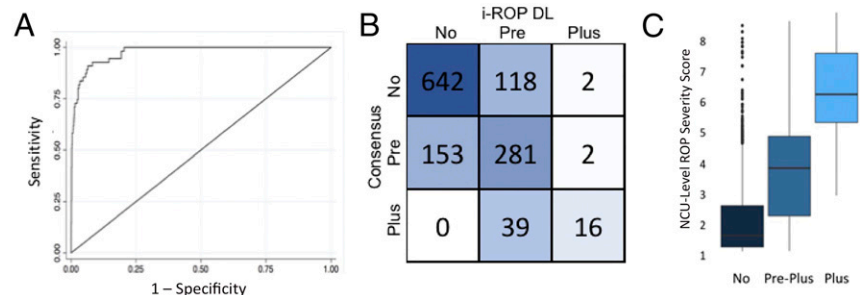
**FIGURE 1**

Scatterplot of birth weight and gestational age in the Retinopathy of Prematurity Eradication Save our Sight program population. Reference lines indicate cutoffs for screening guidelines in the United States (<31 weeks or 1500 g). Screening criteria in India are more liberal, which increases ROP screening burden because infants at a higher birth weight and gestational age remain at risk for disease.

severity scores at each hospital are shown in the Table 1. Only 3 of the 14 NCUs (21%) had both oxygen blenders and pulse oxygenation monitors at the time of the survey. NCUs with both oxygen blenders and pulse oxygenation monitors had lower NCU-level ROP severity, with a mean ROP severity of 3.2 (median 2.9; IQR 2.2–3.5), compared to 3.5 (median 3.2; IQR 2.3–4.0) in units without these resources ( $P = .04$  by analysis of variance). On the

multivariable linear regression, higher NCU-level ROP severity was associated with lower birth weight ( $P < .001$ ) and the absence of oxygen-management capability ( $P = .003$ ). Gestational age and government versus private NCU status were not associated with ROP severity.

The differences between NCUs were even more apparent when we compared bigger infants (birth weight >1500 g) in the NCUs with and



**FIGURE 2**

External validation of the i-ROP DL system for plus disease. A, AUROC for the classification of plus disease = 0.98. B, Confusion matrix for AI versus consensus diagnosis. C, Quantitative severity score (1–9) by consensus diagnosis of plus disease ( $P < .001$  for all comparisons).



**TABLE 1** Demographics for NCUs With >5 Screened Infants

ID	n	ROP Disease Severity, %			Mean Birth Weight, g (Range; SD)	Mean Gestational Age, wk (Range; SD)	Mean Postnatal Age, d (Range; SD)	Government Hospital	Oxygen-Titrating Capability	Mean NCU-Level Severity Score (SD)
		None	Pre-Plus	Plus Disease						
6	15	93	7	0	1439 (930–1710; 281)	31 (27–37; 3)	53 (20–125; 37)	Yes	Yes	2.9 (0.4)
7	15	73	27	0	1363 (825–2100; 379)	32 (28–37; 3)	33 (9–59; 15)	Yes	No	2.9 (0.7)
2	26	62	39	0	1419 (1000–2350; 348)	30 (25–34; 3)	44 (24–66; 13)	Yes	Yes	3.2 (0.9)
3	43	77	19	5	1432 (865–2600; 364)	31 (28–37; 3)	40 (11–74; 15)	Yes	No	3.2 (1.4)
14	26	69	23	8	1271 (750–1850; 330)	30 (26–34; 2)	52 (18–100; 19)	Yes	No	3.3 (1.2)
10	77	66	22	12	1104 (565–1800; 294)	29 (24–34; 3)	48 (7–121; 24)	No	Yes	3.3 (1.3)
19	18	89	0	11	1198 (800–1440; 187)	30 (26–34; 2)	57 (29–118; 26)	No	No	3.4 (1.8)
21	28	64	36	0	1649 (800–2810; 617)	32 (28–37; 3)	34 (17–63; 15)	No	No	3.4 (1.1)
18	141	79	17	4	1545 (800–3200; 346)	32 (25–37; 2)	37 (7–115; 20)	Yes	No	3.5 (1.3)
1	82	66	32	2	1479 (800–2250; 323)	31 (26–36; 2)	40 (12–113; 21)	Yes	No	3.6 (1.3)
25	36	69	19	11	1462 (850–2900; 406)	32 (28–37; 2)	35 (13–55; 10)	Yes	No	3.6 (1.5)
20	45	56	40	4	1288 (900–1850; 287)	31 (28–36; 2)	36 (16–88; 17)	No	No	3.9 (1.2)
11	19	58	21	21	1506 (900–2000; 371)	32 (28–34; 1)	42 (17–71; 16)	Yes	No	3.9 (1.7)
4	12	25	75	0	1491 (1000–2000; 327)	30 (26–34; 3)	39 (10–91; 28)	Yes	No	4.7 (1.6)

ID, study identification number.

without oxygen-management capabilities. Those with oxygen blenders and oxygen monitors for every infant had a mean severity of 2.7 (median 2.7; IQR 2.5–3.0), compared to 3.4 (median 3.1; IQR 2.4–3.8) in those without ( $P = .007$ ), despite those NCUs having a population of infants at significantly higher risk by birth weight ( $P = .02$ ) and gestational age ( $P = .04$ ).

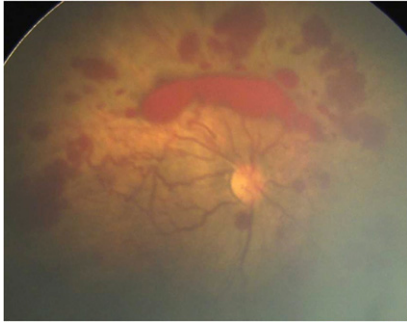
## DISCUSSION

In this study, we retrospectively evaluated an AI system for ROP diagnosis that was developed for a North American population of premature infants on a data set from an ROP telemedicine program in India. The key findings are the following: (1) at the individual eye examination level, the system revealed high diagnostic accuracy as a screening device for TR-ROP; and (2) at the population level, looking at individual NCUs, we found higher ROP severity in NCUs that did not have the resources to monitor and titrate oxygen. We consider these results to be proof of principle that AI may be used to improve the efficiency of ROP screening and also as an

epidemiological tool for monitoring NCU-level ROP severity across geography and time.

In this article, we demonstrate high diagnostic accuracy in an external data set in a real-world ROP screening population in India. Telescreening programs have proven to be an effective force multiplier for ROP screening across large geographic areas; however, because most screening examinations reveal no or mild disease and because even the most efficient systems take significant clinical time away from other patient care responsibilities for clinicians, there is a compelling argument for AI-based ROP screening.<sup>18,26,27</sup> As an autonomous ROP screening device, the system could provide automated real-time referral decisions and refer only positive cases for clinician review, reducing the screening burden by 60% to 80%.<sup>19</sup> In this population, the PPV for TR-ROP was only 12%; however, the PPV for pre-plus or worse disease was 74%. Future prospective evaluation is necessary to determine the cost-effectiveness of various operating points in diverse settings, both in low- and high-income countries.

We further evaluated the relationship between quantitative assessment of ROP severity in a NCU and oxygen-management capability as a rough measure of NCU quality. The relationship between ROP severity and NCU level of care is likely to be a U-shaped curve. At one end, units with high neonatal mortality may have no infants survive to be evaluated for ROP. As neonatal mortality improves, the subsequent risk of ROP increases, even for mildly preterm infants, increasing the population at risk and incidence of severe ROP. Figure 1 reveals the added screening burden caused by the more inclusive screening criteria. Figure 3 reveals an example of an infant born at NCU 11 who would not have met US screening guidelines but developed aggressive posterior retinopathy of prematurity (APROP) presumably related to oxygen exposure. Although there are rare cases of inherited retinal vasculopathies that may mimic APROP, at the population level, the observed relationship between oxygen and these cases of APROP is compelling.<sup>1,2</sup> As NCU quality further improves with strict oxygen monitoring and high-quality neonatal



**FIGURE 3**

Example of APROP at unit 11. This infant would not have met ROP screening criteria in the United States (32 weeks' gestational age and 1980 g), suggesting that this case might have been prevented with improved oxygen monitoring.

care, the population at risk for ROP shifts toward younger and smaller infants, which would reduce the screening burden, and this AI tool could be used for resource allocation to those NCUs that take care of the youngest infants who remain at risk for TR-ROP and APROP.<sup>8,28,29</sup>

There are several limitations to this study. First, it is possible that there are unmeasured variables, such as mortality, referral patterns, and loss to follow-up of discharged patients, that introduced selection bias to the population studied here. However, in general, higher mortality within an NCU and high rates of referral out of an NCU would tend to lower, rather than raise, apparent ROP severity. We believe that although the associations identified here have face validity in light of the ROP epidemic in India, they need further prospective validation, with careful assessment of mortality, referrals, and disease at any

time point (not just the first examination). Second, there are rare cases of TR-ROP without plus disease (zone I, stage 3, no plus); however, none were observed in this population, and nearly all such cases in previous publications would have had positive screen results at the proposed operating point.<sup>19</sup> The relationship between the AI severity score and zone I stage 3 eyes is worth further prospective evaluation in other populations. Fourth, the i-ROP DL system was developed by using images from a single camera system (Retcam; Natus Medical Incorporated), which is expensive and not universally available in LMICs. Further work is being done to evaluate AI on images from other, lower cost, camera systems, which will be important to scale this approach.

In this article, we demonstrate that AI may not only have implications for secondary prevention of ROP but also be a useful tool for monitoring disease epidemiology, which may have application both in high- and low-income countries. Improved primary prevention would be more impactful in terms of reducing the incidence of ROP and reducing the screening burden than any new therapeutic intervention for treatment of late-stage ROP. In addition, AI may also have a role in developing objective disease classification systems<sup>19,20</sup> and standardizing treatment thresholds<sup>30</sup> and may play a role in ROP education, especially in regions of the world where ROP is an emerging disease.<sup>31</sup> AI has been used for outbreak and

infectious disease surveillance via big data sources, such as social media and search metadata.<sup>32,33</sup> However, to our knowledge, the current study is the first application of an AI image-based disease classifier at the population level for epidemiological assessment of disease severity and may be a generalizable approach for other disease states. The final, and likely hardest, challenge for this and all AI technologies will be to incorporate these technologies into sustainable models and integrated health systems so that the potential benefits of these technologies may be seen.

#### ACKNOWLEDGMENT

We acknowledge Szu-Yeu Hu, MD, who developed the optic disc segmenter used in this analysis.

#### ABBREVIATIONS

AI: artificial intelligence  
 APROP: Aggressive posterior retinopathy of prematurity  
 AUROC: area under the receiver operating characteristic curve  
 IQR: interquartile range  
 i-ROP DL: Imaging and Informatics in Retinopathy of Prematurity Deep Learning  
 LMIC: low- and middle-income country  
 NCU: neonatal care unit  
 PPV: positive predictive value  
 ROP: retinopathy of prematurity  
 TR-ROP: treatment-requiring retinopathy of prematurity

Drs Brown and Singh contributed to the development of the artificial intelligence system evaluated in this article, performed portions of the data analysis, and reviewed and revised the manuscript; Drs Cole, Redd, Valikodath, and Rajan and Ms Ostro assisted with the data analysis and critically revised the manuscript; and all authors approved the final manuscript as submitted and agree to be accountable for all aspects of the work.

**DOI:** <https://doi.org/10.1542/peds.2020-016618>

Accepted for publication Oct 29, 2020

Address correspondence to J. Peter Campbell, MD, MPH, Department of Ophthalmology, Casey Eye Institute, Oregon Health & Science University, 545 SW Campus Dr, Portland, OR 97239. E-mail: campbelp@ohsu.edu

**FINANCIAL DISCLOSURE:** Dr Chan is on the Scientific Advisory Board for Phoenix Technology Group (Pleasanton, CA) and is a consultant for Novartis (Basel, Switzerland) and Alcon (Fort Worth, TX); Dr Chiang is a consultant for Novartis (Basel, Switzerland) and an equity owner of InTeleretina (Honolulu, HI); Drs Chiang, Campbell, Chan, and Kalpathy-Cramer receive research support from Genentech; Dr Chan receives research support from Regeneron; the other authors have indicated they have no financial relationships relevant to this article to disclose.

**FUNDING:** Supported by grants R01EY19474, K12EY027720, and P30EY10572 from the National Institutes of Health (Bethesda, MD), by grants SCH-1622679 and SCH-1622542 from the National Science Foundation (Arlington, VA), and by unrestricted departmental funding and a Career Development Award (to Dr Campbell) from Research to Prevent Blindness (New York, NY). The Retinopathy of Prematurity Eradication Save our Sight program was funded in part through a grant from the US Agency for International Development Child Blindness Prevention Program. None of the funding agencies had any 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. Funded by the National Institutes of Health (NIH).

**POTENTIAL CONFLICT OF INTEREST:** The artificial intelligence technology evaluated in this article was invented by Drs Chan, Campbell, Chiang, Brown, and Kalpathy-Cramer and is owned by Oregon Health & Science University, Massachusetts General Hospital, University of Illinois at Chicago, and Northeastern University. Related technology has been licensed for commercial development, which may result in royalties to Massachusetts General Hospital, Oregon Health & Science University, and Dr Kalpathy-Cramer. This potential conflict of interest has been reviewed and managed by Massachusetts General Hospital and Oregon Health & Science University.

**COMPANION PAPER:** A companion to this article can be found online at [www.pediatrics.org/cgi/doi/10.1542/peds.2020-034314](http://www.pediatrics.org/cgi/doi/10.1542/peds.2020-034314)

## REFERENCES

1. 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
2. Shah PK, Narendran V, Kalpana N, Gilbert C. Severe retinopathy of prematurity in big babies in India: history repeating itself? *Indian J Pediatr*. 2009;76(8):801–804
3. Gilbert C. Retinopathy of prematurity: a global perspective of the epidemics, population of babies at risk and implications for control. *Early Hum Dev*. 2008;84(2):77–82
4. Terry TL. Extreme prematurity and fibroblastic overgrowth of persistent vascular sheath behind each crystalline lens. I. Preliminary report. *Am J Ophthalmol*. 1942;25(2):203–204
5. Patz A, Hoek LE, De La Cruz E. Studies on the effect of high oxygen administration in retrolental fibroplasia. I. Nursery observations. *Am J Ophthalmol*. 1952;35(9):1248–1253
6. The Committee for the Classification of Retinopathy of Prematurity. An international classification of retinopathy of prematurity. *Arch Ophthalmol*. 1984;102(8):1130–1134
7. Cryotherapy for Retinopathy of Prematurity Cooperative Group. Multicenter trial of cryotherapy for retinopathy of prematurity. Preliminary results. *Arch Ophthalmol*. 1988;106(4):471–479
8. Quinn GE, Barr C, Bremer D, et al. Changes in course of retinopathy of prematurity from 1986 to 2013: comparison of three studies in the United States. *Ophthalmology*. 2016;123(7):1595–1600
9. Blencowe H, Cousens S, Chou D, et al.; Born Too Soon Preterm Birth Action Group. Born too soon: the global epidemiology of 15 million preterm births. *Reprod Health*. 2013;10(suppl 1):S2
10. Fierson WM; American Academy of Pediatrics Section on Ophthalmology; American Academy of Ophthalmology; American Association for Pediatric Ophthalmology and Strabismus; American Association of Certified Orthoptists. Screening examination of premature infants for retinopathy of prematurity. [published correction appears in *Pediatrics*. 2019;143(3):e20183810]. *Pediatrics*. 2018;142(6):e20183061
11. Shukla R, Murthy GVS, Gilbert C, Vidyadhar B, Mukpalkar S. Operational guidelines for ROP in India: a summary. *Indian J Ophthalmol*. 2020;68(suppl 1):S108–S114
12. Blencowe H, Lawn JE, Vazquez T, Fielder A, Gilbert C. Preterm-associated visual impairment and estimates of retinopathy of prematurity at regional and global levels for 2010. *Pediatr Res*. 2013;74(suppl 1):35–49
13. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol*. 2019;103(2):167–175
14. International Committee for the Classification of Retinopathy of Prematurity. The international classification of retinopathy of prematurity revisited. *Arch Ophthalmol*. 2005;123(7):991–999
15. Early Treatment For Retinopathy Of Prematurity Cooperative Group. Revised indications for the treatment of retinopathy of prematurity: results of the early treatment for retinopathy of prematurity randomized trial. *Arch Ophthalmol*. 2003;121(12):1684–1694
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
17. Bellefleur V, Lim ZW, Lim G, et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *Lancet Digit Health*. 2019;1(1):e35–e44

18. Ting DSW, Wu WC, Toth C. Deep learning for retinopathy of prematurity screening. *Br J Ophthalmol*. 2019; 103(5):577–579
19. Redd TK, Campbell JP, Brown JM, et al. Evaluation of a deep learning image assessment system for detecting severe retinopathy of prematurity. *Br J Ophthalmol*. 2019;103(5):580–584
20. 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
21. 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
22. Zhao M, Jiang Y. Great expectations and challenges of artificial intelligence in the screening of diabetic retinopathy. *Eye (Lond)*. 2020;34(3):418–419
23. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med*. 2018;15(11):e1002683
24. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. Navab N, Hornegger J, Wells W, Frangi A, eds *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. 2015;Vol 9351:234–241
25. Abadi M. TensorFlow: learning functions at scale. In: ICFP 2016: Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming, September 18–24, 2016; Nara, Japan
26. 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
27. Wang SK, Callaway NF, Wallenstein MB, Henderson MT, Leng T, Moshfeghi DM. SUNDROP: six years of screening for retinopathy of prematurity with telemedicine. *Can J Ophthalmol*. 2015; 50(2):101–106
28. 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
29. Shah PK, Subramanian P, Venkatapathy N, Chan RVP, Chiang MF, Campbell JP. Aggressive posterior retinopathy of prematurity in two cohorts of patients in South India: implications for primary, secondary, and tertiary prevention. *J AAPOS*. 2019;23(5):264.e1-264.e4
30. Choi RY, Brown JM, Kalpathy-Cramer J, et al.; Imaging and Informatics in Retinopathy of Prematurity Consortium. Variability in plus disease identified using a deep learning-based retinopathy of prematurity severity scale. *Ophthalmol Retina*. 2020;4(10): 1016–1021
31. Campbell JP, Swan R, Jonas K, et al. Implementation and evaluation of a tele-education system for the diagnosis of ophthalmic disease by international trainees. *AMIA Annu Symp Proc*. 2015;2015:366–375
32. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*. 2009; 457(7232):1012–1014
33. Sadilek A, Caty S, DiPrete L, et al. Machine-learned epidemiology: real-time detection of foodborne illness at scale. *NPJ Digit Med*. 2018;1:36