Development and Validation of a Deep Learning Model to Predict the Occurrence and Severity of Retinopathy of Prematurity

Qiaowei Wu, MD; Yijun Hu, PhD; Zhenyao Mo, ME; Rong Wu, PhD; Xiayin Zhang, PhD; Yahan Yang, PhD; Baoli Liu, MD; Yu Xiao, MD; Xiaomin Zeng, MD; Zhanjie Lin, MD; Ying Fang, MD; Yijin Wang, MD; Xiaohui Lu, PhD; Yanping Song, MD; Wing W. Y. Ng, PhD; Songfu Feng, PhD; Honghua Yu, PhD

Abstract

**IMPORTANCE** Retinopathy of prematurity (ROP) is the leading cause of childhood blindness worldwide. Prediction of ROP before onset holds great promise for reducing the risk of blindness.

**OBJECTIVE** To develop and validate a deep learning (DL) system to predict the occurrence and severity of ROP before 45 weeks' postmenstrual age.

**DESIGN, SETTING, AND PARTICIPANTS** This retrospective prognostic study included 7033 retinal photographs of 725 infants in the training set and 763 retinal photographs of 90 infants in the external validation set, along with 46 characteristics for each infant. All images of both eyes from the same infant taken at the first screening were labeled according to the final diagnosis made between the first screening and 45 weeks' postmenstrual age. The DL system was developed using retinal photographs from the first ROP screening and clinical characteristics before or at the first screening in infants born between June 3, 2017, and August 28, 2019.

**EXPOSURES** Two models were specifically designed for predictions of the occurrence (occurrence network [OC-Net]) and severity (severity network [SE-Net]) of ROP. Five-fold cross-validation was applied for internal validation.

**MAIN OUTCOMES AND MEASURES** Area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, and specificity to evaluate the performance in ROP prediction.

**RESULTS** This study included 815 infants (450 [55.2%] boys) with mean birth weight of 1.91 kg (95% CI, 1.87-1.95 kg) and mean gestational age of 33.1 weeks (95% CI, 32.9-33.3 weeks). In internal validation, mean AUC, accuracy, sensitivity, and specificity were 0.90 (95% CI, 0.88-0.92), 52.8% (95% CI, 49.2%-56.4%), 100% (95% CI, 97.4%-100%), and 37.8% (95% CI, 33.7%-42.1%), respectively, for OC-Net to predict ROP occurrence and 0.87 (95% CI, 0.82-0.91), 68.0% (95% CI, 61.2%-74.8%), 100% (95% CI, 93.2%-100%), and 46.6% (95% CI, 37.3%-56.0%), respectively, for SE-Net to predict severe ROP. In external validation, the AUC, accuracy, sensitivity, and specificity were 0.94, 33.3%, 100%, and 7.5%, respectively, for OC-Net, and 0.88, 56.0%, 100%, and 35.3%, respectively, for SE-Net.

**CONCLUSIONS AND RELEVANCE** In this study, the DL system achieved promising accuracy in ROP prediction. This DL system is potentially useful in identifying infants with high risk of developing ROP.
Introduction

The reduction of blindness in children is one of the major goals of the World Health Organization’s Vision 2020 program. As the leading cause of childhood blindness, retinopathy of prematurity (ROP) is estimated to affect more than 184,700 children worldwide. The incidence and severity of ROP are associated with various factors, such as birth weight (BW), gestational age (GA), oxygen exposure, blood transfusions, and several systemic risk factors. Early screening can effectively reduce childhood blindness and health care costs associated with ROP. Because of the lack of widespread implementation of a cost-effective screening strategy, the cost of ROP screening to the global health system remains high. While the window for treating ROP is brief, treatment has multigenerationally beneficial consequences for individuals, families, societies, and national economies. Thus, early screening, regular monitoring, and timely treatment are critical for premature infants.

Multiple longitudinal examinations using fundus photography (eg, by RetCam3) have been widely adopted for ROP screening, which are stressful to infants, laborious to ophthalmologists, and often unavailable in remote areas. There is also a growing shortage of pediatric ophthalmologists or retinal specialists capable of performing ROP screening. Moreover, less than 10% of mild ROP cases progress to sight-threatening ROP that requires further treatment. Given the extensive efforts required for ROP screening, low rates of treatment-requiring ROP, and misdiagnosis of severe ROP because of interrupted follow-up, cost-effective programs that can identify infants with high risk of severe ROP are needed. Previous studies have used traditional regression analysis to predict ROP based on BW, GA, gender, and/or postnatal weight gain, among other factors. However, some of these prediction models are flawed because of their neglect retinal status, which is also important for ROP prediction.

Deep learning (DL) has been recently used for the automated diagnosis of many ocular diseases. Several DL systems have been developed for automated diagnosis of the plus disease of ROP with high accuracy (>89.6%) based on RetCam images. However, to our knowledge, there is no DL system to date capable of predicting the occurrence and severity of ROP. In the present study, we developed a DL system for the prediction of occurrence and severity of ROP before 45 weeks’ postmenstrual age (PMA) based on retinal photographs from the first ROP screening and clinical characteristics before or at the first ROP screening.

Methods

Data Set Preparation

The study flowchart is presented in Figure 1. We collected data retrospectively on 988 infants (delivered between June 3, 2017, and August 28, 2019) who underwent ROP examinations at the Zhujiang Hospital of Southern Medical University (ZHSMU) and the Second Nanning People’s Hospital (SNPH). The ROP screening criteria were developed according to an international guideline, including a BW of less than 1501 g, a GA of 30 weeks or younger, or risk for ROP as determined by the pediatrician or neonatologist. The study was conducted according to the Declaration of Helsinki and was approved by the Research Ethics Committee of ZHSMU and SNPH. Informed consent was obtained from parents of all infants enrolled. This study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guideline.

In total, data of 988 infants who underwent ROP screening were retrieved from the screening centers and electronic medical records (EMR) in ZHSMU and SNPH. Of those, 71 infants (7.2%) were excluded because retinal photographs were labeled as unacceptable for evaluation; 37 infants (3.7%) were excluded for invalid, repetitive, and missing records; 42 infants (4.3%) were excluded for failing to complete follow-up; and 23 infants (2.3%) were excluded for having clinical features of ROP at the first screening. There were 815 infants (82.5%) with complete data and eligible for the
development and validation of the DL system who had a known ROP outcome before 45 weeks’ PMA but no clinical features of ROP on the retinal photographs at the first screening. There was no significant difference in the birth characteristics between the included and excluded infants (eTable 1 in the Supplement). A total of 7033 images from the first screening of 725 infants were collected from ZHSMU by certified ophthalmologists and used for training and cross-validating the DL system. An additional 763 images from the first screening of 90 infants were obtained from SNPH following the same criteria of the training set and were used for external validation. A total of 46 clinical characteristics for each infant were also collected from EMR in ZHSMU and SNPH before or at the first screening. Screening schedule details performed at each follow-up are reported in eAppendix 1 in the Supplement.

During each screening, multiple retinal photographs are captured using a commercially available camera (RetCam; Natus Medical Inc). A total of 46 clinical characteristics of each infant from the EMR were extracted, including 7 maternal factors, 18 neonatal factors, 7 treatment factors, and 14 laboratory factors. These clinical characteristics were selected according to the reported risk factors associated with the development of ROP. Detailed definitions, the time of collection, and references to related studies of the characteristics are provided in eAppendix 2 in the Supplement.

Image Labeling
All retinal photographs used for the development and validation of the DL system were collected only from the first screening. All cases were annotated independently by 2 ophthalmologists (S.F. and Q.W.) with more than 3 years’ experience of ROP care according to the retinal photographs. Since the

Figure 1. Workflow Diagram

988 Infants in data set

Data set preparation:
First ROP screening
Electronic medical records
Clinical characteristics

173 Infants excluded
71 With low-quality images
42 Did not complete follow-up
23 Clinical features of ROP at first screening

815 Infants in included data set

725 Infants from ZHSMU in training set
33350 Clinical characteristics
7033 Retinal photographs

183 Infants with ROP (25.2%)
542 Infants without ROP (74.8%)

Training and internal validation
(5-fold cross-validation)

116 Infants with mild ROP (61.4%)
67 Infants with severe ROP (36.6%)

Training and internal validation
(5-fold cross-validation)

25 Infants with ROP (27.8%)
65 Infants without ROP (72.2%)

Training and internal validation
(5-fold cross-validation)

17 Infants with mild ROP (68.0%)
8 Infants with severe ROP (32.0%)

External validation

ROP indicates retinopathy of prematurity; SNPH, Second Nanning People’s Hospital; and ZHSMU, Zhujiang Hospital of Southern Medical University.
ROP prediction was performed on each case rather than each retinal photograph independently, all retinal photographs of both eyes from the same infant were defined and labeled as 1 case. A case would be excluded if 1 of 2 ophthalmologists labeled the quality of the retinal photographs unacceptable for evaluation. All ophthalmoscopic and image-based examination findings for each case were documented according to the international classification of ROP, with zone, stage, and plus disease. A case was annotated ROP if the clinical features of ROP were found in any retinal photographs of this case; otherwise, it was annotated normal. To predict treatment-requiring ROP, cases with type II ROP; zone II stage 1 or stage 2 ROP without plus disease; and zone III stage 1, 2, or 3 ROP were graded as mild ROP, which only requires regular follow-up; cases with threshold disease, stage 4 or 5 ROP, and type I or aggressive posterior-ROP were graded as severe ROP, which needs prompt treatment. If a case had different severities in each eye, the case would be graded according to the higher ROP grade of either eye.

The κ was 0.81 for annotation of ROP occurrence and 0.72 for annotation of severe ROP, suggesting good agreement of the 2 clinical ophthalmologists in image labeling. Moreover, the labels were double confirmed by a senior retinal specialist (H.Y.) with more than 15 years of clinical retina experience to generate a final annotation. These annotations were used as the ground-truth labels in the training and validation of the DL system.

**DL System Development and Validation**

An illustration of our DL system is presented in Figure 2. The ROP prediction consists of 2 tasks. One is to predict ROP occurrence before 45 weeks’ PMA; the other is to predict severity of ROP. An occurrence network (OC-Net) and a severity network (SE-Net) with similar network architectures were developed for these 2 tasks. The prediction labels used in OC-Net were normal and ROP; and the prediction labels used in SE-Net were mild ROP and severe ROP.

Multiple preprocessing steps were performed to standardize input images before deep feature learning. After preprocessing and characterization by the residual network–50 (Microsoft Corp), 512 features of each retinal photograph, extracted from the global average pool layer, were concatenated with 46 clinical characteristics of the same case into a final vector of 558 dimensions. A deep neural network was trained on the representative feature vector of 558 values to generate the predictive probability of each retinal photograph.

We adopted 3 different training schemes (ie, majority voting, 1-vote veto, and image-level methods) to output the predictive labels based on the probability threshold. Details about the

---

Figure 2. Illustration of the Proposed Algorithmic Pipeline

A deep learning system was developed to predict occurrence and severity of retinopathy of prematurity (ROP) before 45 weeks’ postmenstrual age (PMA) using a major voting scheme. OC-Net indicates occurrence network; ResNet, residual network; and SE-Net, severity network.
development and validation of the DL system and the different training schemes are provided in eAppendix 3 in the Supplement.

For internal validation of the 3 training schemes, the 5-fold cross-validation manner was used on the training set, including the occurrence data set and the severity data set. The mean and 95% CI of the performing metrics of the 5 runs were recorded to measure the overall performance of the 3 training schemes. Then, the best of the 3 training schemes was deployed in the DL system for further validation, with a requirement of 100% sensitivity. OC-Net and SE-Net with 100% sensitivity were used to perform the external validation. Furthermore, we obtained all weights of each characteristic from the last fully connected layers and calculated weight ratios of all characteristics by normalization to reflect the importance of characteristics to the DL system. The higher the weight ratios are, the more important the corresponding characteristics are for the prediction task.

ROP Prediction Using ROPScore
To better demonstrate the prediction accuracy of our DL system, performance of a commonly used algorithm, ROPScore, was also evaluated. ROPScore is accessible online and also used to predict occurrence and severity of ROP for each infant before 45 weeks' PMA. Details about the validation using ROPScore are provided in eAppendix 4 in the Supplement. The predictive performance of ROPScore was compared with the DL system.

Visualization Method
To visualize the critical regions in retinal photographs highly associated with the ROP prediction, Grad-CAM was used to generate saliency maps and enhance the interpretability of the DL system. Grad-CAM calculates feature weight to generate heat maps that highlight the regions that play an important role in the final prediction. By visualizing the saliency maps, we can observe and determine the contribution of potential fundus features to the final ROP prediction.

Statistical Analysis
Breakdowns of the predictive labels with reference to the ground-truth labels were depicted as confusion matrices, which were used to calculate the accuracy, sensitivity, and specificity of OC-Net and SE-Net with the 3 training schemes (Python version 3.7.0 [Python Software Foundation]). We also used the area under the receiver operating characteristic curve (AUC) to determine the predictive performance of the DL system with reference to the ground-truth labels and conducted calibration plots to examine overestimation or underestimation of risks in the DL system.

Results
Description of the Study Cohort
Among 815 included infants, 450 (55.2%) were boys, the mean GA was 33.1 weeks (95% CI, 32.9-33.3 weeks; median, 33 weeks; range, 25-40 weeks), and the mean BW was 1.91 kg (95% CI, 1.87-1.95 kg; median, 1.92 kg; range, 0.75-3.80 kg) (eFigure 1 in the Supplement). Demographic characteristics of the training and the external validation sets are summarized in Table 1.

Internal Validation Under Requirement of 100% Sensitivity
Of the 725 cases included in the training set, 542 (74.8%) were annotated normal, and 183 (25.2%) were annotated ROP. Of the 183 ROP cases, 116 (63.4%) were graded mild ROP, and 67 (36.6%) were graded severe ROP. The results of the 3 training schemes of OC-Net and SE-Net are summarized in eTable 2 in the Supplement. Among the 3 training schemes, the major voting scheme achieved the best overall performance for both OC-Net and SE-Net. Thus, the major voting scheme was deployed for further validation under the requirement of 100% sensitivity. After adjusting thresholds of OC-Net to 0.10 and SE-Net to 0.26, both OC-Net and SE-Net achieved 100% sensitivity to capture all ROP and severe ROP cases, respectively, with the same model structure.
## Demographic Characteristics

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>All (N = 815)</th>
<th>Training set (n = 725)</th>
<th>External validation set (n = 90)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Maternal factors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maternal age, mean (SD), y</td>
<td>30.9 (5.6)</td>
<td>30.9 (5.6)</td>
<td>31.2 (5.7)</td>
</tr>
<tr>
<td>Cesarean delivery</td>
<td>445 (54.6)</td>
<td>392 (54.1)</td>
<td>53 (58.9)</td>
</tr>
<tr>
<td>IVF-ET</td>
<td>98 (12.0)</td>
<td>79 (10.9)</td>
<td>19 (21.1)</td>
</tr>
<tr>
<td>Gestational hypertension</td>
<td>76 (9.3)</td>
<td>63 (8.7)</td>
<td>13 (14.4)</td>
</tr>
<tr>
<td>Gestational diabetes</td>
<td>154 (18.9)</td>
<td>131 (18.1)</td>
<td>23 (25.6)</td>
</tr>
<tr>
<td>Intrauterine infection</td>
<td>26 (3.2)</td>
<td>24 (3.3)</td>
<td>2 (2.2)</td>
</tr>
<tr>
<td>Use of dexamethasone</td>
<td>160 (19.6)</td>
<td>142 (19.6)</td>
<td>18 (20.0)</td>
</tr>
<tr>
<td><strong>Neonatal factors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gestational age, mean (SD), wk</td>
<td>33.1 (3.2)</td>
<td>33.4 (3.2)</td>
<td>30.7 (2.2)</td>
</tr>
<tr>
<td>Birth weight, mean (SD), kg</td>
<td>1.91 (0.6)</td>
<td>2.0 (0.6)</td>
<td>1.5 (0.3)</td>
</tr>
<tr>
<td>Male sex</td>
<td>450 (55.2)</td>
<td>401 (55.3)</td>
<td>50 (55.6)</td>
</tr>
<tr>
<td>Small for gestational age</td>
<td>61 (7.5)</td>
<td>54 (7.5)</td>
<td>10 (11.1)</td>
</tr>
<tr>
<td>Multiple gestations</td>
<td>202 (24.8)</td>
<td>175 (24.1)</td>
<td>27 (30.0)</td>
</tr>
<tr>
<td>Apgar Score, mean (SD)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-min</td>
<td>8.5 (1.9)</td>
<td>8.5 (1.9)</td>
<td>8.6 (1.5)</td>
</tr>
<tr>
<td>5-min</td>
<td>9.4 (1.3)</td>
<td>9.3 (1.3)</td>
<td>9.4 (0.9)</td>
</tr>
<tr>
<td>Intrauterine distress</td>
<td>81 (9.9)</td>
<td>75 (10.3)</td>
<td>6 (6.7)</td>
</tr>
<tr>
<td>Asphyxia</td>
<td>140 (17.2)</td>
<td>126 (17.4)</td>
<td>14 (15.6)</td>
</tr>
<tr>
<td>Bronchopulmonary dysplasia</td>
<td>148 (18.2)</td>
<td>121 (16.7)</td>
<td>27 (30.0)</td>
</tr>
<tr>
<td>Intracranial hemorrhage</td>
<td>166 (20.4)</td>
<td>139 (19.2)</td>
<td>27 (30.0)</td>
</tr>
<tr>
<td>Sepsis</td>
<td>103 (12.6)</td>
<td>89 (12.8)</td>
<td>14 (15.6)</td>
</tr>
<tr>
<td>Hypoxic ischemic encephalopathy</td>
<td>77 (9.5)</td>
<td>76 (10.5)</td>
<td>8 (8.9)</td>
</tr>
<tr>
<td>Respiratory distress syndrome</td>
<td>334 (41.0)</td>
<td>266 (36.7)</td>
<td>36 (40.0)</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>430 (52.8)</td>
<td>359 (49.5)</td>
<td>47 (52.2)</td>
</tr>
<tr>
<td>Necrotizing enterocolitis</td>
<td>47 (5.8)</td>
<td>41 (5.7)</td>
<td>6 (6.7)</td>
</tr>
<tr>
<td>Neonatal jaundice</td>
<td>578 (70.9)</td>
<td>504 (69.5)</td>
<td>73 (81.1)</td>
</tr>
<tr>
<td>Patent ductus arteriosus</td>
<td>229 (28.1)</td>
<td>197 (27.2)</td>
<td>33 (36.7)</td>
</tr>
<tr>
<td><strong>Treatment factors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>History of oxygen exposure</td>
<td>560 (68.7)</td>
<td>497 (68.6)</td>
<td>63 (70.0)</td>
</tr>
<tr>
<td>Use of oxygen in MV</td>
<td>422 (51.8)</td>
<td>366 (50.5)</td>
<td>56 (62.2)</td>
</tr>
<tr>
<td>Duration of oxygen in MV, mean (SD), d</td>
<td>15.5 (20.0)</td>
<td>14.3 (19.4)</td>
<td>25.3 (22.1)</td>
</tr>
<tr>
<td>Blood transfusion</td>
<td>482 (59.1)</td>
<td>423 (58.3)</td>
<td>59 (65.6)</td>
</tr>
<tr>
<td>RBC transfusion</td>
<td>423 (51.9)</td>
<td>367 (50.6)</td>
<td>56 (62.2)</td>
</tr>
<tr>
<td>No. of RBC transfusions, mean (SD)</td>
<td>1.1 (1.6)</td>
<td>1.1 (1.6)</td>
<td>1.6 (1.9)</td>
</tr>
<tr>
<td>Volume of RBC transfusion, mean (SD), U</td>
<td>0.4 (0.6)</td>
<td>0.4 (0.6)</td>
<td>0.5 (0.6)</td>
</tr>
<tr>
<td><strong>Laboratory factors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-sensitivity CRP, mean (SD), mg/dL</td>
<td>0.7 (2.3)</td>
<td>0.8 (2.4)</td>
<td>0.3 (1.0)</td>
</tr>
<tr>
<td>WBC count, mean (SD), /μL</td>
<td>11 900 (6800)</td>
<td>12 200 (7000)</td>
<td>9900 (4900)</td>
</tr>
<tr>
<td>RBC count, mean (SD), 10⁷/μL</td>
<td>4.3 (1.1)</td>
<td>4.3 (1.0)</td>
<td>4.6 (1.4)</td>
</tr>
<tr>
<td>Hemoglobin concentration, mean (SD), g/dL</td>
<td>14.97 (3.68)</td>
<td>14.82 (3.54)</td>
<td>16.17 (4.53)</td>
</tr>
<tr>
<td>Hematocrit, mean (SD), mg/L</td>
<td>0.4 (0.1)</td>
<td>0.4 (0.1)</td>
<td>0.5 (1.0)</td>
</tr>
<tr>
<td>Corpuscular volume, mean (SD), fl</td>
<td>101.4 (11.5)</td>
<td>100.9 (11.6)</td>
<td>104.9 (10.5)</td>
</tr>
<tr>
<td>RBC distribution width, mean (SD), %</td>
<td>16.7 (2.3)</td>
<td>16.7 (2.4)</td>
<td>16.7 (1.5)</td>
</tr>
<tr>
<td>Absolute neutrophil count, mean (SD), /μL</td>
<td>7000 (5600)</td>
<td>7200 (5700)</td>
<td>5400 (3700)</td>
</tr>
<tr>
<td>Absolute lymphocyte count, mean (SD), /μL</td>
<td>3600 (3100)</td>
<td>3600 (3200)</td>
<td>3200 (1500)</td>
</tr>
<tr>
<td>Absolute monocyte count, mean (SD), /μL</td>
<td>1300 (1200)</td>
<td>1400 (1300)</td>
<td>1200 (1200)</td>
</tr>
<tr>
<td>Platelet count, mean (SD), 10¹²/μL</td>
<td>261.6 (122.9)</td>
<td>262.9 (127.2)</td>
<td>251.2 (80.0)</td>
</tr>
<tr>
<td>Platelet volume, mean (SD), fl</td>
<td>10.3 (1.1)</td>
<td>10.3 (1.1)</td>
<td>10.2 (0.8)</td>
</tr>
<tr>
<td>Platelet distribution width, mean (SD), %</td>
<td>12.3 (2.5)</td>
<td>12.4 (2.6)</td>
<td>11.4 (1.7)</td>
</tr>
<tr>
<td>Serum total bilirubin, mean (SD), mg/dL</td>
<td>5.99 (4.13)</td>
<td>5.97 (4.27)</td>
<td>6.10 (2.70)</td>
</tr>
</tbody>
</table>

Abbreviations: CRP, C-reactive protein; IVF-ET, in vitro fertilization and embryo transfer; MV, mechanical ventilation; RBC, red blood cell; WBC, white blood cell. SI conversion factors: To convert CRP to milligrams per liter, multiply by 10; hemoglobin to grams per liter, multiply by 10.0; lymphocyte count, monocyte count, neutrophil count, and WBC to ×10⁹ cells per liter, multiply by 0.001; platelet count to ×10¹² cells per liter, multiply by 1; RBC to ×10¹² cells per liter, multiply by 1; and total bilirubin to micromoles per liter, multiply by 17.104.
Applying the threshold of 0.10, the mean AUC, accuracy, sensitivity, and specificity of OC-Net to predict ROP occurrence were 0.90 (95% CI, 0.88-0.92), 52.8% (95% CI, 49.2%-56.4%), 100% (95% CI, 97.4%-100%), and 37.8% (95% CI, 33.7%-42.1%), respectively (Table 2 and Figure 3A). Applying the threshold of 0.26, the mean AUC, accuracy, sensitivity, and specificity of SE-Net to predict severe ROP were 0.87 (95% CI, 0.82-0.91), 68.0% (95% CI, 61.2%-74.8%), 100% (95% CI, 93.2%-100%), and 46.6% (95% CI, 37.3%-56.0%), respectively (Table 2 and Figure 3B). Moreover, calibration plots

Table 2. Overall Performance of the Deep Learning System Under the Requirement of 100% Sensitivity and ROP Score in Prediction of ROP

<table>
<thead>
<tr>
<th>System</th>
<th>AUC (95% CI)</th>
<th>% (95% CI)</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal validation for prediction of ROP occurrence, with threshold of 0.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OC-Net, major voting scheme</td>
<td>0.90 (0.88-0.92)</td>
<td>52.8 (49.2-56.4)</td>
<td>100 (97.4-100)</td>
<td>37.8 (33.7-42.1)</td>
<td></td>
</tr>
<tr>
<td>ROP Score</td>
<td>0.76 (0.72-0.80)</td>
<td>69.0 (65.6-72.3)</td>
<td>61.7 (54.3-68.7)</td>
<td>71.4 (67.4-75.1)</td>
<td></td>
</tr>
<tr>
<td>Internal validation for prediction of ROP severity, with threshold of 0.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SE-Net, major voting scheme</td>
<td>0.87 (0.82-0.91)</td>
<td>68.0 (61.2-74.8)</td>
<td>100 (93.2-100)</td>
<td>46.6 (37.3-56.0)</td>
<td></td>
</tr>
<tr>
<td>ROP Score</td>
<td>0.67 (0.59-0.75)</td>
<td>67.8 (64.4-71.2)</td>
<td>53.7 (41.2-65.8)</td>
<td>75.9 (66.9-83.1)</td>
<td></td>
</tr>
<tr>
<td>External validation for prediction of ROP occurrence, with threshold of 0.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OC-Net, major voting scheme</td>
<td>0.94</td>
<td>33.3</td>
<td>100</td>
<td>7.5</td>
<td></td>
</tr>
<tr>
<td>ROP Score</td>
<td>0.76 (0.72-0.80)</td>
<td>58.9 (55.3-62.5)</td>
<td>100 (83.4-100)</td>
<td>43.1 (31.1-55.9)</td>
<td></td>
</tr>
<tr>
<td>External validation for prediction of ROP severity, with threshold of 0.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SE-Net, major voting scheme</td>
<td>0.88</td>
<td>56.0</td>
<td>100</td>
<td>35.3</td>
<td></td>
</tr>
<tr>
<td>ROP Score</td>
<td>0.80 (0.62-0.98)</td>
<td>60.0 (56.4-63.6)</td>
<td>100 (59.8-100)</td>
<td>41.2 (19.4-66.5)</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: AUC, area under the receiver operating characteristic curve; OC-Net, occurrence network; ROP, retinopathy of prematurity; SE-Net, severity network.
of the major voting scheme showed that the model was well adapted (eFigure 3 in the Supplement), and improvement in the DL system performance by clinical characteristics is described in eAppendix 5 in the Supplement.

**External Validation**
Of the 90 cases included in the external validation set, 65 (72.2%) were annotated normal, and 25 (27.8%) were annotated ROP. Of the 25 ROP cases, 17 (68.0%) were graded mild ROP, and 8 (32.0%) were graded severe ROP. Using the same threshold as internal validation, the AUC, accuracy, sensitivity, and specificity were 0.94, 33.3%, 100%, and 7.5%, respectively, for OC-Net and 0.88, 56.0%, 100%, and 35.3%, respectively, for SE-Net (Table 2 and Figure 3).

**Characteristics of Infants With False-Negative and True-Positive Results for the Prediction of ROP**
To explore the characteristics of cases with incorrect predictions, we compared the birth characteristics between the false-negative and true-positive results of occurrence and severity predictions for ROP. For OC-Net, the major voting scheme predicted that 21 infants (10.1%) with ROP would not have ROP, while 187 infants (89.9%) infants with ROP were correctly predicted. Compared with infants with true-positive results, those with false-negative results were more likely to be born via cesarean delivery with a later GA, greater BW, better systemic condition, and less need for blood transfusions and oxygen therapy (eTable 3 in the Supplement). For SE-Net, the major voting scheme predicted that 10 infants (13.3%) with severe ROP would have mild ROP, while 65 infants (86.7%) with severe ROP were correctly predicted. Compared with those with true-positive results, the infants with false-positive results had greater BW, lower prevalence of intracranial hemorrhage, and less need for red blood cell (RBC) transfusions (eTable 4 in the Supplement).

**ROP Prediction Using ROPScore**
Using the training data set, the mean AUC, accuracy, sensitivity, and specificity were 0.76 (95% CI, 0.72-0.80), 69.0% (95% CI, 65.6%-72.3%), 61.7% (95% CI, 54.3%-68.7%), and 71.4% (95% CI, 67.4%-75.1%), respectively, for ROPscore to predict ROP occurrence and 0.67 (95% CI, 0.59-0.75), 67.8% (95% CI, 64.4%-71.2%), 53.7% (95% CI, 41.2%-65.8%), and 75.9% (95% CI, 66.9%-83.1%), respectively, for ROPscore to predict severe ROP (Table 2). Using the external validation data set, the mean AUC, accuracy, sensitivity, and specificity were 0.76 (95% CI, 0.72-0.80), 58.9% (95% CI, 55.3%-62.5%), 100% (95% CI, 83.4%-100%), and 43.1% (95% CI, 31.1%-55.9%), respectively, for ROPscore to predict ROP occurrence and 0.80 (95% CI, 0.62-0.98), 60.0% (95% CI, 56.4%-63.6%), 100% (95% CI, 59.8%-100%), and 41.2% (95% CI, 19.4%-66.5%), respectively, for ROPscore to predict severe ROP (Table 2).

**Visualization Saliency Map**
The saliency map helps highlight areas used by the DL system for ROP prediction. The heat in saliency maps represents the importance of the regions on retinal photographs for the final prediction. A redder color indicates a greater contribution to the DL system; a bluer color indicates a smaller contribution to the DL system. For OC-Net, the saliency map mainly highlighted the optic disc and the vessels and areas around the optic disc. For SE-Net, the saliency map mainly highlighted the nasal and inferior areas around the optic disc (eFigure 4 in the Supplement). These results suggested that the fundus features of the optic disc and surrounding areas might have certain predictive value in ROP development.

**Weight Ratios of Different Characteristics**
To identify the most important clinical characteristics in the DL system, we calculated the weight ratios of all characteristics in the last fully connected layer of OC-Net and SE-Net. The results identified GA, BW, use of RBC transfusion, use of oxygen in mechanical ventilation, and occurrence
of respiratory distress syndrome (total weight ratios ≥36%) as the 5 most critical characteristics for OC-Net to predict ROP occurrence. BW, GA, cesarean delivery, use of RBC transfusion, and occurrence of intraventricular hemorrhage (total weight ratios ≥34%) were the 5 most critical characteristics for SE-Net to predict severe ROP (eFigure 2 in the Supplement).

Discussion

In the present study, a DL system was developed to predict the occurrence and severity of ROP before 45 weeks’ PMA based on retinal photographs from the first screening and clinical characteristics collected before or at the first screening. Our results indicated that ROP prediction using the DL system with the major voting scheme was promising. With the help of our DL system, ophthalmologists and parents could be alerted to ensure that regular screening is performed before the onset of ROP. This DL system may also be applied to reduce the workload for pediatric ophthalmologists and to increase the treatment rate of severe ROP.

Some researchers have used regression analysis to predict ROP based on prenatal and postnatal factors.16,20-23 These prediction models that can identify infants with high risk of developing sight-threatening ROP early may improve the efficacy of ROP screening. However, retinal status is also important for ROP prediction. Previous studies have shown that minor vascular dilation, vascular tortuosity insufficiency in plus disease, the extent of temporal vascularization, and retinal immaturity at the first screening have prognostic significance in the early course of ROP.25-27 Thus, our DL system was developed to predict occurrence and severity of ROP using retinal photographs from the first screening along with clinical characteristics. Our results show that the integration of retinal and other clinical information using DL technology can achieve reliable accuracy in ROP prediction. In addition, we compared the birth characteristics between the false-negative and true-positive results for predicting the occurrence and severity of ROP (eTable 3 and eTable 4 in the Supplement). This investigation will help us better understand the characteristics of infants whose ROP diagnosis is missed or whose severity is incorrectly predicted and develop strategies to minimize the risk of misdiagnosis in the future.

To select the best output pattern of the DL system, we adopted 3 different training schemes. The majority voting scheme may be more similar to the clinical reality than the others, given that ophthalmologists make a diagnosis of ROP according to all retinal photographs taken in screenings. However, the 1-vote veto method may be used when high sensitivity is required, such as infants with multiple major risk factors for ROP. We also validated and compared the predictive performance of ROPSscore with the DL system under the requirement of 100% sensitivity. The specificity and accuracy of SE-Net and ROPSscore were comparable in predicting severe ROP.

Our DL system has 2 promising outcomes. First, early identification of infants with high risk of ROP, especially severe ROP, may prompt ophthalmologists and parents to perform fundus screening, which can reduce the risk of infants with severe ROP being misdiagnosed and untreated because of interrupted follow-up, thereby further reducing childhood blindness and societal burden.13-15 That was also why we classified ROP cases into mild and severe categories, so more alerts would be raised if the prediction outcome was severe ROP. Second, the global cost of ROP screening remains high, especially in remote areas.12,14,18 Regional nurseries without pediatric ophthalmologists typically transfer high-risk infants to tertiary hospitals for screening, which will incur substantial costs and delay treatment delivery.11,12,35 However, screening programs using digital photography and telemedicine can improve the cost-effectiveness of the health system and reduce the risk of infant transport.11,12,35

To minimize the black-box effect and increase model transparency,30,31 we used Grad-CAM to generate saliency maps to highlight the potential regions in retinal photographs from the first screening that contributed most toward ROP prediction. Notably, the saliency maps showed that OC-Net predicted ROP occurrence mainly according to the optic disc and the vessels and areas around the optic disc, while SE-Net predicted severe ROP mainly according to the nasal and inferior
areas around the optic disc (Figure 4 in the Supplement). These exploratory findings might help advance our understanding of the pathophysiological mechanisms and evolution of vascular abnormalities in the course of ROP. Previous studies have demonstrated decreased global average, nasal, and superior disc retinal nerve fiber layer (RNFL) thicknesses in preterm children and thinner average RNFL thickness in children with severe ROP. Therefore, the fundus features of the optic disc and surrounding areas may have a certain predictive value in ROP prediction. Additionally, the DL algorithms could identify and interpret pathological features that physicians could not perceive, regardless of their clinical experience.

In the future, our DL system may be integrated into cloud computing platforms for telemedicine ROP screening. After uploading retinal photographs from the first screening and clinical characteristics of an infant through a website, an automated ROP prediction could be generated. After an initial prediction report is completed, human experts would evaluate it online and generate a final report with a personalized follow-up scheme and a clinical decision.

Limitations
Our study has some limitations. First, the generalizability of the DL system may be limited by the differences in the prevalence of clinical characteristics between different studies. Our DL system needs to be further validated by other prospective multiple-center data sets. Second, the DL system can only predict the occurrence of ROP and whether it is mild or severe ROP. More advanced DL systems that can provide more detailed predictions (e.g., plus disease, stage, and zone) need to be developed in the future.

Conclusions
In this study, our DL system provided accurate predictions of occurrence and severity of ROP before 45 weeks' PMA based on retinal photographs from the first ROP screening and clinical characteristics measured before or at the first screening. The DL system is potentially helpful to identify infants at high risk of developing ROP and reduce blindness resulting from ROP.

ARTICLE INFORMATION
Accepted for Publication: April 29, 2022.
Published: June 16, 2022. doi: 10.1001/jamanetworkopen.2022.17447

Open Access: This is an open access article distributed under the terms of the CC-BY License. © 2022 Wu Q et al. JAMA Network Open.

Corresponding Authors: Honghua Yu, PhD, Guangdong Eye Institute, Department of Ophthalmology, Guangdong Provincial People’s Hospital, Guangdong Academy of Medical Sciences, Second Zhongshan Rd 106#, Guangzhou 510080, China (yuhonghua@gdph.org.cn); and Songfu Feng, PhD, Department of Ophthalmology, Zhujiang Hospital of Southern Medical University, Industrial Ave Middle 253#, Guangzhou 510280, China (fsf516@163.com).

Author Affiliations: Guangdong Eye Institute, Department of Ophthalmology, Guangdong Provincial People’s Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China (Q. Wu, Hu, Zhang, Xiao, Zeng, Lin, Fang, Yu); Department of Ophthalmology, General Hospital of Central Theater Command, Wuhan, China (Q. Wu, Song); Guangdong Provincial Key Laboratory of Computational Intelligence and Cyberspace Information, School of Computer Science and Engineering, South China University of Technology, Guangzhou, China (Mo, Ng); Department of Ophthalmology, Zhujiang Hospital of Southern Medical University, Guangzhou, China (R. Wu, Lu, Feng); State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China (Yang, Liu); Department of Neonatology, Second Nanning People’s Hospital, Nanning, China (Wang).

Author Contributions: Dr Yu had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Drs Q. Wu, Hu, and Mo contributed equally and are considered co-first authors.
Concept and design: Q. Wu, Hu, Mo, R. Wu, Ng, Feng, Yu.

Acquisition, analysis, or interpretation of data: Q. Wu, Hu, Mo, R. Wu, Zhang, Yang, Liu, Xiao, Zeng, Lin, Fang, Wang, Lu, Song, Feng, Yu.

Drafting of the manuscript: Q. Wu, Hu, R. Wu, Liu, Wang, Ng.

Critical revision of the manuscript for important intellectual content: Q. Wu, Hu, Mo, Zang, Yang, Liu, Xiao, Zeng, Lin, Fang, Lu, Song, Feng, Yu.

Statistical analysis: Q. Wu, Mo, R. Wu, Zhang, Liu, Xiao, Zeng, Fang, Yu.

Obtained funding: Song, Yu.

Administrative, technical, or material support: Feng, Yu.

Supervision: Hu, Zhang, Yang, Lu, Song, Ng, Feng, Yu.

Conflict of Interest Disclosures: None reported.

Funding/Support: Dr Yu was supported by grants 81870663 and 82171075 from the National Natural Science Foundation of China; 202206010092 from the Guangzhou Municipal Science and Technology Bureau; KJ012019087 from the Outstanding Young Talent Trainee Program of Guangdong Provincial People's Hospital; and YO12018145 from the Talent Introduction Fund of Guangdong Provincial People's Hospital.

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.

Additional Information: The data sets can be obtained from the corresponding author upon request.

REFERENCES


56. Park KA, Oh SY. Retinal nerve fiber layer thickness in prematurity is correlated with stage of retinopathy of prematurity. Eye (Lond). 2015;29(12):1594-1602. doi:10.1038/eye.2015.166

**SUPPLEMENT.**

- eAppendix 1. Retinopathy of Prematurity Screening and Follow-up Schedule
- eAppendix 2. References, Detailed Definitions, and Time Point of Collection of the Clinical Characteristics
- eAppendix 3. Deep Learning System Development and Validation
- eAppendix 4. Prediction of Retinopathy of Prematurity Using ROPScore
- eAppendix 5. Improvement in Deep Learning System Performance by Clinical Characteristics
- eFigure 1. Distributions of Birth Weight and Gestational Age of All Included Infants
- eFigure 2. Weight Ratios of Different Characteristics
- eFigure 3. Calibration Plots for Evaluating the Calibration Ability of the Deep Learning System Using Major Voting Scheme
- eFigure 4. Original Retinal Photographs and Saliency Maps of Normal and Retinopathy of Prematurity Cases from Grad-CAM
- eTable 1. Comparison of Birth Characteristics Between the Included and Excluded Infants
- eTable 2. Performance of the Deep Learning System Using Different Training Schemes for Predicting Occurrence and Severity of Retinopathy of Prematurity in Internal Validation
- eTable 3. Comparison of Birth Characteristics of Infants With False-Negative and True-Positive Results for Prediction of Occurrence of Retinopathy of Prematurity
- eTable 4. Comparison of Birth Characteristics of Infants With False-Negative and True-Positive Results for Prediction of Severity of Retinopathy of Prematurity
- eReferences.