Validated System for Centralized Grading of Retinopathy of Prematurity
Telemedicine Approaches to Evaluating Acute-Phase Retinopathy of Prematurity (e-ROP) Study

Ebenezer Daniel, MBBS, MS, MPH, PhD; Graham E. Quinn, MD, MSCE; P. Lloyd Hildebrand, MD; Anna Ells, MD, FRCSC; G. Baker Hubbard III, MD; Antonio Capone Jr, MD; E. Revell Martin, BA; Candace P. Ostroff, BA; Eli Smith, BA; Maxwell Pistilli, MS; Gui-Shuang Ying, PhD; for the e-ROP Cooperative Group

**IMPORTANCE** Measurable competence derived from comprehensive and advanced training in grading digital images is critical in studies using a reading center to evaluate retinal fundus images from infants at risk for retinopathy of prematurity (ROP). Details of certification for nonphysician trained readers (TRs) have not yet been described.

**OBJECTIVE** To describe a centralized system for grading ROP digital images by TRs in the Telemedicine Approaches to Evaluating Acute-Phase Retinopathy of Prematurity (e-ROP) Study.

**DESIGN, SETTING, AND PARTICIPANTS** Multicenter observational cohort study conducted from July 1, 2010, to June 30, 2014. The TRs were trained by experienced ROP specialists and certified to detect ROP morphology in digital retinal images under supervision of an ophthalmologist reading center director. An ROP reading center was developed with standard hardware, secure Internet access, and customized image viewing software with an electronic grading form. A detailed protocol for grading was developed. Based on results of TR gradings, a computerized algorithm determined whether referral-warranted ROP (RW-ROP; defined as presence of plus disease, zone I ROP, and stage 3 or worse ROP) was present in digital images from infants with birth weight less than 1251 g enrolled from May 25, 2011, through October 31, 2013. Independent double grading was done by the TRs with adjudication of discrepant fields performed by the reading center director.

**EXPOSURE** Digital retinal images.

**MAIN OUTCOMES AND MEASURES** Intrgrader and intergrader variability and monitoring for temporal drift.

**RESULTS** Four TRs underwent rigorous training and certification. A total of 5520 image sets were double graded, with 24.5% requiring adjudication for at least 1 component of RW-ROP. For individual RW-ROP components, the adjudication rate was 3.9% for plus disease, 12.4% for zone I ROP, and 16.9% for stage 3 or worse ROP. The weighted κ for intergrader agreement (n = 80 image sets) was 0.72 (95% CI, 0.52-0.93) for RW-ROP, 0.57 (95% CI, 0.37-0.77) for plus disease, 0.43 (95% CI, 0.24-0.63) for zone I ROP, and 0.67 (95% CI, 0.47-0.88) for stage 3 or worse ROP. The weighted κ for grade-regrade agreement was 0.77 (95% CI, 0.57-0.97) for RW-ROP, 0.87 (95% CI, 0.67-1.00) for plus disease, 0.70 (95% CI, 0.51-0.90) for zone I ROP, and 0.77 (95% CI, 0.57-0.97) for stage 3 or worse ROP.

**CONCLUSIONS AND RELEVANCE** These data suggest that the e-ROP system for training and certifying nonphysicians to grade ROP images under the supervision of a reading center director reliably detects potentially serious ROP with good intrgrader and intergrader consistency and minimal temporal drift.

Published online March 26, 2015.

Copyright 2015 American Medical Association. All rights reserved.
Worldwide, there is limited availability of ophthalmologists experienced in detection of severe retinopathy of prematurity (ROP).1,2 A telemedicine system that can accurately identify infants with potentially severe ROP can maximize the likelihood of detecting eyes with referral-warranted (RW) ROP, ie, morphological features associated with severe ROP, such as plus disease, zone I ROP, or stage 3 or worse ROP, that may indicate a need for intervention.3 The Telemedicine Approaches to Evaluating Acute-Phase Retinopathy of Prematurity (e-ROP) Study was a large multicenter, National Eye Institute–funded, clinical study undertaken to evaluate the validity of an ROP telemedicine system to detect eyes that have RW-ROP.4 It compared results of digital image grading vs findings of binocular indirect ophthalmoscopic examinations performed by study-certified and ROP-experienced ophthalmologists.5 Similar to telemedicine approaches in diabetic retinopathy, the e-ROP Study established a reading center (RC) where trained, certified nonphysician readers supervised by an ophthalmologist RC director graded standardized image sets of eyes at risk for ROP from infants with birth weight less than 1251 g enrolled from May 25, 2011, through October 31, 2013. We describe the training, certification, operational workflow, and quality assurance in an ROP RC that supported the e-ROP Study.

Methods

RC Infrastructure

Standardized independent workstations with secure Internet access were provided to all trained readers (TRs) in the e-ROP Study; they included similarly configured computers with monitors that were calibrated every 2 weeks to maintain consistency in brightness and hue. Software was developed for displaying and manipulating contrast, brightness, and magnification in the ROP images and data from grading were captured using web-based forms (Figure).

Parental informed consent was provided for all infants prior to study enrollment. The institutional review boards allowed for verbal consent followed by written consent by the infant’s parent or legal guardian at 4 clinical centers; at all other centers, written consent had to be obtained prior to enrollment. After obtaining approval from the institutional review boards at each of the participating clinical centers, standard 6-image sets were acquired for each eye, sorted by field of view with respect to the ideal field, focus, and clarity, and uploaded to the Inoveon Data Center by nonphysician imagers. The image sets were assigned to reading queues that could be randomly assigned to a general reader queue or assigned to a specific reader. The TRs used a structured grading protocol to...
document morphological ROP features on electronic grading forms. All image sets were graded independently by 2 TRs, with discrepancies adjudicated by the RC director. All images and grading data were stored in the e-ROP Inoveon Data Center and then exported to the data coordination center for review and statistical analysis. This study was conducted from July 1, 2010, to June 30, 2014.

Training and Certification of TRs

The Department of Ophthalmology, University of Pennsylvania, Philadelphia, had an existing RC with 3 TRs with extensive experience in reading color and fluorescein digital retinal images in studies related to age-related macular degeneration and diabetic retinopathy but with no previous experience in grading ROP images. Requirements for TRs included demonstrating ability to grade digital images systematically and to adhere strictly to the protocol. The TRs had diverse undergraduate backgrounds. They underwent 3-phase training, a precertification process, and a final certification process for the e-ROP Study.

Phase 1 training included didactic lectures, interactive sessions, and assigned readings that covered classification of ROP. The e-ROP Study protocol, telemedicine principles, the grading protocol, and current ROP treatments. A broad spectrum of ROP clinical images was shown and discussed during interactive training sessions through face-to-face meetings and webinars with participation by expert graders, the RC director, and the study chair. The TRs visited the neonatal intensive care unit at The Children’s Hospital of Philadelphia to observe the imaging of premature babies. To complete phase 1, the trainees were required to successfully complete a knowledge assessment test.

In phase 2 training, the TRs independently viewed and graded training image sets with known ROP grading from a previous ROP study database. Prior to the images’ use in training, a group of experts generated consensus final grading results for each image set that were used as the answer keys to assess TR grading performance. Using a paper grading form, the TRs graded each training image set for the presence or absence of plus disease, zone I ROP, and stage 3 or worse ROP. The percentage of agreement of each grading variable from each reader was determined by comparing TR grading results with expert consensus results. Each training session included independent grading of an average of 15 image sets by each TR, followed by the review of training image sets and discussion of their grading results. The study biostatistician (G.-S.Y.) reviewed the analysis of the grader training results with the RC director and study chair to identify areas that warranted additional training.

In phase 3 training, the TRs graded additional ROP RetCam image sets using the electronic form and grading protocol. Deidentified training image sets that included classic ROP morphology, various artifacts, and different aspects of quality related to focus, clarity, and field were provided. The TRs graded and reviewed 100 ROP training image sets. They graded image sets individually and met once a week with the study chair, RC director, and a clinical expert (A.E.) through teleconference to compare findings and discuss discrepancies with the image sets displayed on shared monitors.

During the precertification process, the TRs were required to demonstrate good agreement with the consensus grading of training image sets and then 10 image sets from the e-ROP pilot study were provided for independent grading. Grading results for RW-ROP and its components were compared with the consensus grading for that image set. An agreement of 85% or higher was judged to be satisfactory, and retraining was required if the TR did not achieve a satisfactory score.

Final certification was conducted once 85% agreement in precertification images was reached. An additional 15 image sets derived from e-ROP pilot submissions were queued for the final certification process. If less than 80% agreement was achieved on these image sets, retraining for a week was performed; then, another 15 image sets were queued and the process was repeated until there was at least 80% agreement with the consensus grading, resulting in TR certification.

Grading Workflow

The RC workflow (eFigure in the Supplement) was developed using the defined roles of a data manager, TRs, RC director, and study chair. The TRs graded all images from infants who developed RW-ROP based on the diagnostic examination. Approximately 80% of infants were not expected to develop RW-ROP; therefore, we had decided a priori to select for the primary outcome article a random sample of approximately 60% of infants who never developed RW-ROP. All image sets from this selected subsample of infants were graded by TRs. The data manager at the data coordination center selected and assigned image sets for grading. Two TRs independently graded each image set. The RC director oversaw the operations of the RC and provided adjudication for discrepancies arising from the TR double grading that were above a predetermined threshold (eTable 1 in the Supplement). On rare occasions, the study chair provided adjudication for grading disparities when referred by the RC director. More details of the grading process are given in eAppendix 1 in the Supplement.

Grading Protocol

The e-ROP Study grading protocol required evaluation of both image quality and the key morphological features of ROP. In developing the final e-ROP Study grading protocol, there were 2 iterations of protocol and a final third version was used to grade all of the study image sets. Details of grading of the image quality and morphological features (posterior pole vessels, zone of ROP, and stage of ROP) are described in eAppendix 2 in the Supplement.

Quality Assurance

To ensure the integrity and completeness of the image evaluation, the TRs were completely masked to all infant demographic information including birth weight and gestational age, clinical data on ROP findings from the diagnostic eye examination, and the grading results from image sets of previous visits and image sets from the fellow eye. In addition, real-time consistency checks were performed and automatic edit queries were generated once the TRs finalized the evaluation of
Results

Four TRs were trained and certified for the study in 2 groups. After 1 member of the initial group of 3 TRs left for another job, a replacement TR was recruited and trained. Certification of the TRs required at least 85% agreement in interpreting RW-ROP with consensus grading. As shown in eTable 2 in the Supplement, the agreements are very high in precertification (80%-100%) and final certification (93%-100%) for each TR.

For the primary outcome article, a total of 5520 image sets were double graded by TRs with adjudication of discrepancies. In the image set grading, 56.4% had discrepancies in at least 1 grading field (either image-quality fields or ROP-morphology fields) that required adjudication by the RC director. A small number of discrepancies (246 image sets [4.5%]) were also reviewed by the study chair at the beginning of the study to assure that the nonphysician readers and the RC director were following the grading protocol. Overall, 24.5% of the image sets required adjudication for a feature that determined whether RW-ROP was present. For individual RW-ROP components, the adjudication rate was 3.9% for plus disease, 12.4% for zone I ROP, and 16.9% for stage 3 or worse ROP (Table 1).

The temporal drift for each of the TRs is shown in Table 2. The temporal drift sample of 25 image sets graded during November 2012 was regraded 3 times by each of the TRs. The grade-regrade agreement for RW-ROP was high with weighted κ ranging from 0.57 to 0.94 across the TRs. A total of 80 image sets from 4 samples of the contemporaneous variability sample were regraded by each TR during the grading period. The weighted κ for intergrader agreement from these 80 image sets is shown in Table 3. The weighted κ was 0.72 (95% CI, 0.52-0.93) for RW-ROP; 0.57 (95% CI, 0.37-0.77) for plus disease; 0.43 (95% CI, 0.24-0.63) for zone I ROP; and 0.67 (95% CI, 0.47-0.88) for stage 3 or worse ROP. Any RO had a weighted κ score of 0.89 (95% CI, 0.68-1.00). The weighted κ for grade-regrade agreement of final consensus grading from these 80 image sets is given in Table 4. The weighted κ was 0.77 (95% CI, 0.57-0.97) for RW-ROP; 0.87 (95% CI, 0.67-1.00) for plus disease; 0.70 (95% CI, 0.51-0.89) for zone I ROP; and 0.77 (95% CI, 0.57-0.97) for stage 3 or worse ROP. Any RO had a perfect agreement.

Discussion

 measurable competence developed during comprehensive and advanced training in grading digital images is critical in studies that use a centralized RC system to evaluate retinal fundus images from infants at risk for ROP. Previously published reports studying telemedicine approaches to ROP have not detailed the processes involved in certifying nonphysician TRs, except one that described a brief training session for nonexpert graders consisting mostly of medical students and ophthalmology residents. Previous studies on ROP telemedicine have used different types of readers: single ROP-experienced ophthalmologist as an unmasked reader; single masked ROP-experienced ophthalmologists; 2 retinal spe-
chologists and 1 general ophthalmologist briefly trained by an ROP-experienced pediatric ophthalmologist; more than 2 masked ROP-experienced ophthalmologists; ROP-experienced ophthalmologists who performed clinical examinations and evaluated retinal images from the same infants after a few months; and nonphysician imagers who also read the images they had taken. Previous clinical studies have not studied a system to evaluate the competency of remote nonphysician graders.

The e-ROP Study RC developed an ROP curriculum, training, and certification for nonphysician TRs and developed and implemented a standardized grading protocol using electronic data capture. The e-ROP Study protocol required 2 TRs to grade image sets independently and significant discrepancies were adjudicated by the RC director and, if needed, by the study chair. By implementing an extensive quality management system that included quality assessment of images, intragrader and intragrader variability, and temporal and contemporaneous drift, the e-ROP Study maintained a consistent and repeatable grading system throughout the study period.

It was important to have a robust certification system that keeps TR agreement levels high, and this was particularly important in accurately identifying morphological features of plus disease, zone I ROP, and stage 3 or worse ROP. An important feature in the development of the TR certification system was establishing a standard criterion of reference not only for the different morphology of eyes with RW-ROP but also for any ROP and preplus disease. This was done through a process of integrating the grading of 3 expert readers, the RC director, and the study chair and using this consensus grading as the standard criterion against which the grading of the TRs was compared. The excellent agreement between TRs reflects the extensive and rigorous training and certification process.

Table 2. Temporal Drift Among Trained Readers in the Telemedicine Approaches to Evaluating Acute-Phase Retinopathy of Prematurity (e-ROP) Study

<table>
<thead>
<tr>
<th>Variable</th>
<th>Grader 1</th>
<th></th>
<th>Grader 2</th>
<th></th>
<th>Grader 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Agreement (%</td>
<td>Weighted κ (95% CI)</td>
<td>Agreement (%</td>
<td>Weighted κ (95% CI)</td>
<td>Agreement (%)</td>
</tr>
<tr>
<td>Any ROP</td>
<td>100</td>
<td>1.00 (1.00 to 1.00)</td>
<td>96</td>
<td>0.94 (0.59 to 1.00)</td>
<td>96</td>
</tr>
<tr>
<td>Ridge</td>
<td>92</td>
<td>0.84 (0.48 to 1.00)</td>
<td>92</td>
<td>0.87 (0.51 to 1.00)</td>
<td>92</td>
</tr>
<tr>
<td>Flat preretinal neovascular proliferation</td>
<td>100</td>
<td>1.00 (1.00 to 1.00)</td>
<td>96</td>
<td>0.65 (0.28 to 1.00)</td>
<td>96</td>
</tr>
<tr>
<td>Zone I ROP</td>
<td>100</td>
<td>1.00 (1.00 to 1.00)</td>
<td>96</td>
<td>0.65 (0.28 to 1.00)</td>
<td>96</td>
</tr>
<tr>
<td>Stage 3 or worse ROP</td>
<td>88</td>
<td>0.77 (0.42 to 1.00)</td>
<td>84</td>
<td>0.71 (0.36 to 1.00)</td>
<td>84</td>
</tr>
<tr>
<td>RW-ROP</td>
<td>88</td>
<td>0.77 (0.42 to 1.00)</td>
<td>84</td>
<td>0.71 (0.36 to 1.00)</td>
<td>84</td>
</tr>
</tbody>
</table>

Abbreviations: NA, not applicable; ROP, retinopathy of prematurity; RW-ROP, referral-warranted ROP.

* The original sample was graded November 30, 2012. Regrade period 1 took place February 28, 2013, through March 1, 2013; regrade period 2, June 17 and 18, 2013; and regrade period 3, October 24 and 25, 2013.
Table 3. Intergrader Variability for the Samples of Contemporaneous Variability in the Telemedicine Approaches to Evaluating Acute-Phase Retinopathy of Prematurity (e-ROP) Study*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Agreement, %</th>
<th>Weighted κ (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pupil quality</td>
<td>94</td>
<td>0.73 (0.56-0.91)</td>
</tr>
<tr>
<td>Disc center quality</td>
<td>66</td>
<td>0.47 (0.30-0.65)</td>
</tr>
<tr>
<td>Disc up quality</td>
<td>74</td>
<td>0.56 (0.43-0.69)</td>
</tr>
<tr>
<td>Disc down quality</td>
<td>78</td>
<td>0.40 (0.25-0.56)</td>
</tr>
<tr>
<td>Disc temporal quality</td>
<td>81</td>
<td>0.66 (0.51-0.81)</td>
</tr>
<tr>
<td>Disc nasal quality</td>
<td>74</td>
<td>0.49 (0.34-0.64)</td>
</tr>
<tr>
<td>Posterior pole vessels</td>
<td>80</td>
<td>0.60 (0.39-0.82)</td>
</tr>
<tr>
<td>Preplus or plus disease in each quadrant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Superonasal quadrant</td>
<td>78</td>
<td>0.59 (0.42-0.76)</td>
</tr>
<tr>
<td>Inferonasal quadrant</td>
<td>71</td>
<td>0.49 (0.32-0.67)</td>
</tr>
<tr>
<td>Inferotemporal quadrant</td>
<td>71</td>
<td>0.56 (0.42-0.76)</td>
</tr>
<tr>
<td>Total quadrants of plus or preplus disease</td>
<td>65</td>
<td>0.68 (0.51-0.84)</td>
</tr>
<tr>
<td>Total quadrants of plus disease</td>
<td>88</td>
<td>0.50 (0.32-0.68)</td>
</tr>
<tr>
<td>Dominant feature of vessels</td>
<td>71</td>
<td>0.58 (0.42-0.75)</td>
</tr>
<tr>
<td>Any ROP</td>
<td>95</td>
<td>0.89 (0.68-1.00)</td>
</tr>
<tr>
<td>Demarcation line</td>
<td>99</td>
<td>0.75 (0.56-0.92)</td>
</tr>
<tr>
<td>Ridge</td>
<td>83</td>
<td>0.65 (0.45-0.86)</td>
</tr>
<tr>
<td>Extraretinal fibrovascular proliferation</td>
<td>83</td>
<td>0.67 (0.47-0.88)</td>
</tr>
<tr>
<td>Flat preretinal neovascular proliferation</td>
<td>100</td>
<td>1.00 (1.00-1.00)</td>
</tr>
<tr>
<td>Retinal detachment</td>
<td>100</td>
<td>1.00 (1.00-1.00)</td>
</tr>
<tr>
<td>Highest stage</td>
<td>81</td>
<td>0.85 (0.67-1.00)</td>
</tr>
<tr>
<td>Lowest zone</td>
<td>80</td>
<td>0.81 (0.63-0.99)</td>
</tr>
<tr>
<td>Plus disease</td>
<td>90</td>
<td>0.57 (0.37-0.77)</td>
</tr>
<tr>
<td>Zone I ROP</td>
<td>83</td>
<td>0.43 (0.24-0.63)</td>
</tr>
<tr>
<td>Stage 3 or worse ROP</td>
<td>83</td>
<td>0.67 (0.47-0.88)</td>
</tr>
<tr>
<td>RW-ROP</td>
<td>85</td>
<td>0.72 (0.52-0.93)</td>
</tr>
</tbody>
</table>

Abbreviations: ROP, retinopathy of prematurity; RW-ROP, referral-warranted ROP.
* All samples combined (N = 80).

Table 4. Intragrader Variability Among Trained Readers in the Telemedicine Approaches to Evaluating Acute-Phase Retinopathy of Prematurity (e-ROP) Study

<table>
<thead>
<tr>
<th>Variable</th>
<th>Agreement, %</th>
<th>Weighted κ (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pupil quality</td>
<td>95</td>
<td>0.76 (0.59-0.93)</td>
</tr>
<tr>
<td>Disc center quality</td>
<td>85</td>
<td>0.67 (0.46-0.88)</td>
</tr>
<tr>
<td>Disc up quality</td>
<td>91</td>
<td>0.78 (0.63-0.93)</td>
</tr>
<tr>
<td>Disc down quality</td>
<td>96</td>
<td>0.81 (0.63-0.99)</td>
</tr>
<tr>
<td>Disc temporal quality</td>
<td>90</td>
<td>0.76 (0.61-0.91)</td>
</tr>
<tr>
<td>Disc nasal quality</td>
<td>89</td>
<td>0.64 (0.50-0.78)</td>
</tr>
<tr>
<td>Posterior pole vessels</td>
<td>86</td>
<td>0.73 (0.51-0.94)</td>
</tr>
<tr>
<td>Preplus or plus disease in each quadrant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Superonasal quadrant</td>
<td>88</td>
<td>0.81 (0.64-0.98)</td>
</tr>
<tr>
<td>Inferonasal quadrant</td>
<td>81</td>
<td>0.73 (0.56-0.90)</td>
</tr>
<tr>
<td>Inferotemporal quadrant</td>
<td>89</td>
<td>0.64 (0.49-0.84)</td>
</tr>
<tr>
<td>Total quadrants of plus or preplus disease</td>
<td>74</td>
<td>0.75 (0.57-0.94)</td>
</tr>
<tr>
<td>Total quadrants of plus disease</td>
<td>98</td>
<td>0.87 (0.69-1.00)</td>
</tr>
<tr>
<td>Dominant feature of vessels</td>
<td>69</td>
<td>0.65 (0.48-0.82)</td>
</tr>
<tr>
<td>Any ROP</td>
<td>100</td>
<td>1.00 (1.00-1.00)</td>
</tr>
<tr>
<td>Demarcation line</td>
<td>100</td>
<td>1.00 (1.00-1.00)</td>
</tr>
<tr>
<td>Ridge</td>
<td>98</td>
<td>0.95 (0.74-1.00)</td>
</tr>
<tr>
<td>Extraretinal fibrovascular proliferation</td>
<td>88</td>
<td>0.77 (0.57-0.97)</td>
</tr>
<tr>
<td>Flat preretinal neovascular proliferation</td>
<td>100</td>
<td>1.00 (1.00-1.00)</td>
</tr>
<tr>
<td>Retinal detachment</td>
<td>100</td>
<td>1.00 (1.00-1.00)</td>
</tr>
<tr>
<td>Highest stage</td>
<td>89</td>
<td>0.95 (0.77-1.00)</td>
</tr>
<tr>
<td>Lowest zone</td>
<td>93</td>
<td>0.95 (0.77-1.00)</td>
</tr>
<tr>
<td>Plus disease</td>
<td>98</td>
<td>0.87 (0.67-1.00)</td>
</tr>
<tr>
<td>Zone I ROP</td>
<td>90</td>
<td>0.70 (0.51-0.90)</td>
</tr>
<tr>
<td>Stage 3 or worse ROP</td>
<td>88</td>
<td>0.77 (0.57-0.97)</td>
</tr>
<tr>
<td>RW-ROP</td>
<td>88</td>
<td>0.77 (0.57-0.97)</td>
</tr>
</tbody>
</table>

Abbreviations: ROP, retinopathy of prematurity; RW-ROP, referral-warranted ROP.

Intergrader variability also measured differences in judging the quality of the retinal images. The weighted κ of agreement ranged from 0.40 to 0.66 for each of the 5 retinal images. The nasal and superior fields had the lowest agreements in terms of image quality. The quality of images becomes extremely important when the TR is unable to determine RW-ROP while perceiving morphological changes in poor-quality digital images. However, this study showed uncertainty in grading RW-ROP pathology in fewer than 2% of all image sets graded, and fewer than 1% required adjudication by the RC director for such uncertainty (Table 3). Enhancing the appearance of the ROP morphology and attenuating background noise in poor-quality images by manipulating the contrast, brightness, magnification, and gray tone brings more certainty to detecting ROP pathology in digital images. While some studies have refrained from using image enhancement to avoid the confounding effects of difference in the readers’ skills in enhancing images, we allowed the TRs in our study to use the functional modalities of the e-ROP grading software. Because all TRs’ graded images used these functionalities, we are unable to assess what effect these enhancements had on reducing undecided readings or the extent to which they increased or decreased agreement between readers; we propose investigating this in future studies.

Identifying plus and preplus disease showed an intergrader-variability weighted κ of 0.57. Using International Classification of Retinopathy of Prematurity images as standards for the tortuosity and dilation in the 4 quadrants to identify plus and preplus disease did not appear to adequately minimize intergrader variability and dilation in the 4 quadrants to identify plus and preplus disease.
variability. This was not surprising as identification of plus disease in ROP among experts also appears to be highly variable over several studies.\(^{16-18}\) Four experienced ROP ophthalmologists grading high-quality RetCam images disagreed 10% of the time on the presence of plus disease, and the disagreement increased almost 3-fold when grading was confined to images having only plus and preplus disease.\(^{16}\) Among 22 experienced ROP experts who interpreted wide-angle images for the presence of plus disease, only 27% had mean k scores over 0.80 (substantial agreement), while 18% had scores below 0.41 (slight or fair agreement).\(^{18}\) Unlike our study in which TRs were allowed to look at the 4 peripheral retinal images before identifying plus disease in the disc center image, the studies had these expert readers looking only at the disc center image. In the e-ROP Study, TR disagreement on presence of plus disease was 55%. Thus, these disagreements in identifying plus disease persist in telemedicine ROP studies and need more rigorous refinements on the definition and quantitative methods of detecting plus disease in digital images.

Study strengths included having the TRs and the RC director completely masked to all infant details and having them grade the image sets of each eye independent of the morphological changes that could be present in the other eye. Para-
doxically, this could also have been a limitation of the study as allowing access to information on the gestational age and birth weight of the infant together with the findings in the other eye could have improved the sensitivity and specificity in the study. This hypothesis is being tested in a future study. Another limitation of this study and other similar past studies is that there was no gold standard to assess the competency of the TR in identifying morphological features in the retinal images; the consensus opinion of a few experts in ROP, which is subject to error, was used as the standard for comparison for training and certification of TRs.

**Conclusions**

The results of this study suggest that reliable, comprehensive, systematic training and certification of nonphysician readers of digital image sets of premature infants at risk for ROP are feasible. To our knowledge, this is the first study that has demonstrated consistent and good agreement between and among nonphysician TRs grading ROP from digital images using a centralized reading facility.

---

**ARTICLE INFORMATION**

Submitted for Publication: October 15, 2014; final revision received January 16, 2015; accepted January 29, 2015.

Published Online: March 26, 2015. doi:10.1001/jamaophthalmol.2015.0460.

**Author Affiliations:** Department of Ophthalmology, University of Pennsylvania, Philadelphia (Daniel, Quinn, Martin, Ostroff, Smith, Pistilli, Ying); Division of Pediatric Ophthalmology, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania (Quinn); Dean McGee Eye Institute, University of Oklahoma, Oklahoma City (Hildebrand); Department of Ophthalmology, University of Calgary, Calgary, Alberta, Canada (Elli); Department of Ophthalmology, Emory University, Atlanta, Georgia (Hubbard); Associated Retinal Consultants, Royal Oak, Michigan (Capone); William Beaumont Hospital, Oakland University School of Medicine, Auburn Hills, Michigan (Capone).

**Author Contributions:** Drs Quinn and Ying 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.

Study concept and design: Daniel, Quinn, Ying. Acquisition, analysis, or interpretation of data: All authors.

Drafting of the manuscript: Daniel, Quinn, Hubbard. Critical revision of the manuscript for important intellectual content: Daniel, Quinn, Hildebrand, Elli, Capone, Martin, Ostroff, Smith, Pistilli, Ying.

Statistical analysis: Daniel, Quinn, Pistilli, Ying. Obtained funding: Quinn.

Administrative, technical, or material support: Daniel, Hildebrand, Elli, Martin, Ostroff, Smith. Study supervision: Daniel, Quinn, Hildebrand, Ying.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Dr Hildebrand reported receiving consulting fees from Inoveon Corp, serving as chairman of the board of directors for Inoveon Corp, and receiving royalties for US patents 5940802 and 6470320, “Digital Disease Management System” (assignee: Board of Regents, University of Oklahoma). Dr Ellis reported receiving compensation from VisionQuest Biomedical, LLC for grading photographs outside this work. Dr Capone reported being a founding partner of FocusROP, LLC. Dr Ying reported serving as a statistical consultant for Janssen Research and Development, LLC. No other disclosures were reported.

Funding/Support: This work was supported by cooperative agreement grant U10 EY017014 from the National Eye Institute.

Role of the Funder/Sponsor: The funder 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.

**Group Information:** Members of the e-ROP Cooperative Group include the following: Office of Study Chair, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania: Graham E. Quinn, MD, MSCE (principal investigator [PI]), Kelly C. Wade, MD, PhD, MSCE, Agnieszka Baumritter, MS, Trang B. Duros, and Lisa Erbring, Johns Hopkins University, Baltimore, Maryland: Michael X. Repka, MD (PI), Jennifer A. Shepard, CRNP, David Emmert, BA, and C. Mark Herring; Boston Children's Hospital, Boston, Massachusetts: Deborah VanderVeen, MD (PI), Suzanne Johnston, MD, Carolyn Wu, MD, Jason Montagis, MD, Danielle Ledoux, MD, Tamir Winter, RN, BSN, IBCLC, Frank Weng, and Theresa Mansfield, RN; Nationwide Children's Hospital and Ohio State University Hospital, Columbus: Don L. Brenner, MD (PI), Mary Lou McGregor, MD, Catherine Olson Jordan, MD, David L. Rogers, MD, Rae R. Fellows, MD, CCRC, Suzanne Brandt, RNC, BSN, and Brenda Mann, RNC, BSN, Duke University, Durham, North Carolina: David Wallace, MD (PI), Sharon Freedman, MD, Sarah K. Jones, Du Tran-Viet, and Rhonda “Michelle” Young; University of Louisville, Louisville, Kentucky: Charles C. Barr, MD (PI), Rahul Bhola, MD, Craig Douglas, MD, Peggy Fishman, MD, Michelle Bottorff, Brandi Hubbuch, RN, MSN, NNP-BC, and Rachel Keith, PhD; University of Minnesota, Minneapolis: Erick D. Bothun, MD (PI), Inge DeBecker, MD, Jill Anderson, MD, Ann Marie Holleschau, BA, CCRP, Nichole E. Miller, MA, RN, NNP, and Darla N. Nyquist, MA, RN, NNP, University of Oklahoma, Oklahoma City: R. Michael Siatkowska, MD (PI), Lucas Trigler, MD, Marilyn Escobedo, MD, Karen Cortt, MS, ARNP, NNP-BC, Michelle Huyhn, MS, ARNP, and Kelli Sabatine, MS, ARNP, NNP-BC; The Children's Hospital of Philadelphia: Monte D. Mills, MD (PI), Will Anninger, MD, Gil Binenbaum, MD, MSCE, Graham E. Quinn, MD, MSCE, Karen A. Karp, BSN, and Denise Pearson, COMT, University of Texas Health Science Center, San Antonio: Alice Gong, MD (PI), John Stokes, MD, Clio Armitage Harper, MD, Laurie Weaver, Carmen McHenry, BSN, Kathryn Conner, Rosalind Heener, and Elonna Coley, RN; University of Utah, Salt Lake City: Robert Hoffman, MD (PI), David Dries, MD, Katie Jo Farnsworth, Deborah Harrison, MS, Bonnie Carlstrom, and Cyrie Ann Frye, CRA, OCT-C, Vanderbilt University, Nashville, Tennessee: David Morrison, MD (PI), Sean Donahue, MD, Nancy Benegas, MD, Sandy Owings, COA, CCRP, Sandra Phillips, COT, CRI, and Scott Ruark; Hospital of the Foothills Medical Center, Calgary, Alberta, Canada: Anna Eells, MD, FRCSC (C) (PI), Patrick Mitchell, MD, April Ingram, and Rosie Sorbie, RN, Dto Coordinating Center, University of Pennsylvania School of Medicine, Philadelphia: Gui-Shuang Ying, PhD (PI), Maureen Maguire, PhD, Mary Brightwell-Arnold, BA, SCP, Maxwell Pistilli, MS, Kathleen McWilliams, CCRP, Sandra Harris, and Claressa Whearly; Image Reading Center, University of Pennsylvania School of
REFERENCES


