Objective: To prospectively measure accuracy, reliability, and image quality of telemedical retinopathy of prematurity (ROP) diagnosis.

Methods: Two-hundred forty-eight eyes from 67 consecutive infants underwent wide-angle retinal imaging by a trained neonatal nurse at 31 to 33 weeks’ and/or 35 to 37 weeks’ postmenstrual age (PMA) using a standard protocol. Data were uploaded to a Web-based telemedicine system and interpreted by 3 expert retinal specialist graders who provided a diagnosis (no ROP, mild ROP, type 2 prethreshold ROP, treatment-requiring ROP) and an evaluation of image quality for each eye. Findings were compared with a reference standard of indirect ophthalmoscopy by an experienced pediatric ophthalmologist.

Results: At 35 to 37 weeks’ PMA, sensitivity and specificity for diagnosis of mild or worse ROP were 0.908 and 1.000 for grader A, 0.971 and 1.000 for grader B, and 0.908 and 0.977 for grader C. Sensitivity and specificity for diagnosis of type 2 prethreshold or worse ROP were 1.000 and 0.943 for grader A, 1.000 and 0.930 for grader B, and 1.000 and 0.851 for grader C. At 35 to 37 weeks’ PMA, weighted $\kappa$ for intergrader reliability was 0.791 to 0.889, and $\kappa$ for intragrad reliability for detection of type 2 prethreshold or worse ROP was 0.769 to 1.000. Image technical quality was rated as “adequate” or “possibly adequate” for diagnosis in 93.3% to 100% of eyes.

Conclusion: A telemedicine system using nurse-captured retinal images has the potential to improve existing shortcomings of ROP management, particularly at later PMAs.

Arch Ophthalmol. 2007;125(11):1531-1538

Retinopathy of prematurity (ROP) is a vasoproliferative disease that is diagnosed by serial dilated ophthalmoscopy. Progress has occurred in validation of treatment criteria through the Cryotherapy for Retinopathy of Prematurity (CRYO-ROP) and Early Treatment for Retinopathy of Prematurity (ETROP) trials and in development of an international classification system. However, ROP continues to be a leading cause of childhood blindness throughout the world.

See also pages 1523 and 1562

Retinopathy of prematurity management presents significant challenges: (1) Diagnosis at the neonatal intensive care unit (NICU) bedside requires extensive travel and coordination and is logistically difficult. (2) The number of infants requiring surveillance is increasing. In the United States, the rate of premature births has grown from 9.4% to 12.7% since 1981. Worldwide, ROP incidence is rising as neonatal survival improves. New guidelines have expanded the gestational-age cutoff for examination to decrease the likelihood of missing larger infants with disease. (3) Availability of ophthalmologists who perform ROP examination is limited. A 2006 American Academy of Ophthalmology survey found that only 54% of retinal specialists and pediatric ophthalmologists are managing ROP and that more than 20% plan to stop because of concerns including medicolegal liability and poor reimbursement.

One strategy for improving accessibility and delivery of ROP care is store-and-forward telemedicine, which is an emerging technology where medical data are captured for subsequent interpretation by a remote expert. Widespread adoption of telemedicine has been constrained by the lack of substantive evaluation data. Although studies have shown that interpretation of digital reti-
nal photographs may be accurate enough to identify clinically significant ROP, concerns have been raised about image quality. Furthermore, all published work to our knowledge has used images captured by ophthalmologists or ophthalmic photographers, and most studies have involved only a single image grader. Several studies have been confounded by designs in which the image grader was the same investigator who performed reference standard ophthalmoscopic examinations. The accuracy and reliability of telemedical ROP examination by multiple expert graders based on photographs obtained by non-ophthalmic personnel are not known. This is an important gap in knowledge because large-scale ROP telemedicine systems would likely require image capture by neonatal personnel available at the point of care.

This article describes a prospective study to determine the accuracy and reliability of telemedical ROP diagnosis among 3 expert graders and the quality of image capture by a trained neonatal nurse. Results are compared with a reference standard of ophthalmoscopy by 1 of 2 experienced examiners.

METHODS

NURSE TRAINING AND TELEMEDICINE SYSTEM DESIGN

This study was approved by the Columbia University institutional review board. A neonatal nurse was trained to perform wide-angle retinal imaging using a commercially available device (RetCam II; Clarity Medical Systems, Pleasanton, California). This included 2 day-long instructional sessions with the manufacturer, followed by 6 weekly sessions with 1 of us (M.F.C.) during regular ophthalmoscopic examinations. At each session, approximately 3 infants were photographed, and images were correlated with clinical findings.

A store-and-forward ROP telemedicine application was developed by 2 of us (L.W. and M.F.C.). This included a secure database system (SQL 2005; Microsoft, Redmond, Washington); a module allowing the photographer to upload data and images; and a Web-based interface for expert interpretation. The system was designed to represent real-world telemedicine examinations and scenarios. Images from both eyes were displayed side by side, along with the birth weight, gestational age, and postmenstrual age (PMA) at the time of examination (Figure 1).

OPHTHALMOSCOPIC EXAMINATION AND RETINAL IMAGING

Infants hospitalized in the Columbia University NICU from November 1, 2005, though October 31, 2006, were included if they met existing ROP examination criteria and if their parents provided informed consent for participation. Patients were excluded if they had structural ocular anomalies, had previously received laser or other ROP treatment, or were considered unstable for examination by their neonatologist.

Each subject underwent 2 procedures sequentially performed at the NICU bedside under topical anesthesia: (1) Dilated ophthalmoscopic examination by 1 of 2 pediatric ophthalmologists (S.A.K. and M.F.C.), according to standard protocols. Both ophthalmologists had served as certified investigators in the ETROP study. Findings were documented on clinical templates, based on the international classification standard. (2) Imaging by the study nurse (O.C.), according to a protocol by which an image set consisting of posterior, temporal, and nasal photographs was captured from each retina. Each image set included up to 2 additional photos from any
area of the eye, if felt by the nurse to contribute diagnostic information. Imaging was performed without input from the examining ophthalmologist. No subjects were excluded because of poor image quality or inability to capture photographs. No complications such as apnea or corneal injury occurred to prevent imaging.

Study infants were imaged during up to 2 sessions: (1) 31 to 33 weeks’ PMA, which was intended to represent the time at which initial examinations are performed,10,23 and (2) 35 to 37 weeks’ PMA, which was intended to optimize a time at which clinically significant disease occurs, while minimizing the number of study infants lost from hospital discharge or laser treatment. The best images were selected by the nurse, and data were uploaded to the telemedicine system.

TELEMEDICAL EXAMINATION

Three graders (T.C.L., D.J.W., and A.M.B.) independently performed telemedical examinations using the Web-based system. Each grader was a retina specialist with extensive experience reviewing RetCam images and was responsible for ROP examination and treatment in a tertiary care medical center. Two graders had authored peer-reviewed ROP manuscripts and the third had served as a principal investigator in the ETROP trial. No graders had previously examined any study infants.

Eyes were classified using an ordinal scale based on CRYO-ROP and ETROP criteria12: (1) no ROP; (2) mild ROP, defined as ROP less than type 2 disease; (3) type 2 prethreshold ROP (zone 1, stage 1 or 2, without plus disease, or zone 2, stage 3, without plus disease); (4) treatment-requiring ROP, defined as type 1 ROP (zone 1, any stage, with plus disease; zone 1, stage 3, without plus disease; or zone 2, stage 2 or 3, with plus disease) or threshold ROP (at least 5 contiguous or 8 noncontiguous clock hours of stage 3 in zone 1 or 2, with plus disease); or (5) unknown, meaning that the grader was uncomfortable making a diagnosis from the data provided. Graders also rated the “technical quality” and “retinal coverage” of each image set as adequate, possibly adequate, or inadequate for diagnosis. Finally, to measure intragrader reliability for interpreting the same images at different times, 20% of study examinations were randomly selected for repeated presentation by the system.

ANALYSIS

Eyes were analyzed by examination session (31-33 weeks, 35-37 weeks). Sensitivity, specificity, and area under the receiver operating characteristic curves (AUCs)36 were determined for presence of mild or worse, type 2 prethreshold, or worse, and treatment-requiring ROP. Ophthalmoscopic examination was used as the reference standard. “Unknown” responses were excluded from calculations, so as not to penalize graders for reporting that they were uncomfortable providing a diagnosis.

Intergrader and intragrader reliability of telemedical examination were determined from the κ and weighted κ statistics for chance-adjusted agreement in ordinal diagnosis, using a well-known scale: 0 to 0.20 = slight agreement, 0.21 to 0.40 = fair agreement, 0.41 to 0.60 = moderate agreement, 0.61 to 0.80 = substantial agreement, and 0.81 to 1.00 = near perfect agreement.37

Image quality was assessed from number of “unknown” diagnoses and from image acceptability ratings by graders. Logistic regression was used to determine whether there was a tendency toward improved diagnostic performance as additional photos were captured by the nurse and as additional telemedical examinations were performed by graders. This was performed with the outcomes being sensitivity and false-positive rate for diagnosis of mild or worse, type 2 prethreshold or worse, or treatment-requiring ROP by each grader and with the predictor being order of retinal imaging and grading.

Analysis was performed using statistical software (SPSS 15.0; SPSS Inc, Chicago, Illinois, and R 2.5.0; Free Software Foundation, Boston, Massachusetts). Standard errors were calculated by the jackknife method because both eyes were used. Statistical significance was considered to be a 2-sided P value <.05.

RESULTS

OVERVIEW OF INFANTS AND EXAMINATIONS

Sixty-seven infants participated in this study, of whom 21 (31.3%) received 1 set of examinations at 31 to 33 weeks’ PMA, 10 (14.9%) received 1 set at 35 to 37 weeks’ PMA, and 36 (53.7%) received examinations at each session. Both eyes were examined at all sessions, for a total of 206 unique eyes. Bilateral images from 21 examinations were repeated to test intragrader reliability, for an overall total of 248 study eyes.

Mean infant birth weight was 912.4 g (range, 398-1440 g), and mean gestational age was 26.7 weeks (range, 23-33 weeks). Examination results are summarized in Table 1. From ophthalmoscopic examinations, the incidence of mild or worse ROP was 36.8% (42 of 114 eyes) at 31 to 33 weeks’ PMA and 58.7% (54 of 92 eyes) at 35 to 37 weeks’ PMA. From telemedical examinations at 35 to 37 weeks’ PMA, grader B had a tendency to diagnose more severe disease than ophthalmoscopy (P=.005) and grader C had a tendency to diagnose more severe disease than ophthalmoscopy (P<.001), grader A (P<.001), and grader B (P=.001).

ACCURACY OF TELEMEDICINE EXAMINATION

Table 2 reports telemedicine accuracy when “unknown” responses were excluded. For infants 31 to 33 weeks’ PMA, all graders had sensitivity of 0.729 or greater, specificity of 0.893 or greater, and AUC of 0.840 or greater for diagnosis of mild or worse ROP. For infants 35 to 37 weeks’ PMA, all graders had sensitivity of 1.000, specificity of 0.851 or greater, and AUC of 0.955 or greater for diagnosis of type 2 prethreshold or worse ROP and sensitivity of 1.000, specificity of 0.806 or greater, and AUC of 0.903 or greater for diagnosis of treatment-requiring ROP.

RELIABILITY OF TELEMEDICAL EXAMINATION

Table 3 displays intergrader reliability of telemedical examination based on ordinal classification. The mean κ and weighted κ among all pairs of graders were 0.615 and 0.654 at 31 to 33 weeks’ PMA and 0.735 and 0.823 at 35 to 37 weeks’ PMA, indicating substantial to near-perfect agreement. Table 4 shows intragrader reliability results. Among infants 31 to 33 weeks’ PMA, κ was 0.462 to 0.769 (mod-
erate to substantial agreement) for detection of mild or worse ROP and was 1.000 (perfect agreement) for detection of treatment-requiring ROP by each grader. Among infants 35 to 37 weeks' PMA, intragrader $\kappa$ was 0.909 to 1.000 for detection of mild or worse ROP and 0.786 to 1.000 for detection of treatment-requiring ROP. Figure 2 displays ex-

### Table 1. Ordinal Diagnostic Classification of ROP in Study Eyes$^a$

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>31-33 wk</td>
<td>35-37 wk</td>
<td>31-33 wk</td>
<td>35-37 wk</td>
<td>31-33 wk</td>
<td>35-37 wk</td>
<td>31-33 wk</td>
<td>35-37 wk</td>
<td>31-33 wk</td>
<td>35-37 wk</td>
<td>31-33 wk</td>
<td>35-37 wk</td>
<td></td>
</tr>
<tr>
<td>No ROP$^d$</td>
<td>80 (62.5)</td>
<td>44 (36.7)</td>
<td>53 (51.0)</td>
<td>45 (39.5)</td>
<td>38 (50.0)</td>
<td>44 (39.3)</td>
<td>88 (68.8)</td>
<td>50 (41.7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mild ROP$^d$</td>
<td>41 (32.0)</td>
<td>50 (41.7)</td>
<td>43 (41.3)</td>
<td>38 (33.3)</td>
<td>27 (35.5)</td>
<td>36 (32.1)</td>
<td>30 (23.4)</td>
<td>30 (25.0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 2 prethreshold$^d$</td>
<td>7 (5.5)</td>
<td>14 (11.7)</td>
<td>6 (5.8)</td>
<td>13 (11.4)</td>
<td>11 (14.5)</td>
<td>13 (11.6)</td>
<td>2 (1.6)</td>
<td>7 (5.8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment-requiring ROP$^d$</td>
<td>0</td>
<td>12 (10.0)</td>
<td>2 (1.9)</td>
<td>18 (15.8)</td>
<td>0</td>
<td>19 (17.0)</td>
<td>8 (6.3)</td>
<td>33 (27.5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total assessed$^d$</td>
<td>128 (100)</td>
<td>120 (100)</td>
<td>128 (100)</td>
<td>114 (100)</td>
<td>76 (100)</td>
<td>112 (100)</td>
<td>128 (100)</td>
<td>120 (100)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown$^e$</td>
<td>0</td>
<td>0</td>
<td>24 (18.8)</td>
<td>6 (5.0)</td>
<td>52 (40.6)</td>
<td>8 (6.7)</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total sample size</td>
<td>128</td>
<td>120</td>
<td>128</td>
<td>120</td>
<td>128</td>
<td>120</td>
<td>128</td>
<td>120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Abbreviation: ROP, retinopathy of prematurity.*

$^a$Results are shown for reference standard ophthalmoscopic examination by an experienced pediatric ophthalmologist and for telemedical examination by 3 experienced retinal specialist graders reviewing images captured by a trained nurse. Results are shown for eyes at 31 to 33 weeks' postmenstrual age and eyes at 35 to 37 weeks' postmenstrual age.

$^b$Grader B was significantly more likely to diagnose more severe disease than the reference standard ($P=.005$) by Wilcoxon signed rank test.

$^c$Grader C was significantly more likely to diagnose more severe disease than the reference standard ($P<.001$), grader A ($P<.001$), or grader B ($P=.001$) by Wilcoxon signed rank test.

$^d$Percentages obtained by dividing by the total number of images assessed.

$^e$Percentages obtained by dividing by the total number of the sample size.

### Table 2. Accuracy of Telemedical ROP Diagnosis by 3 Experienced Retinal Specialist Graders Based on Sensitivity, Specificity, and AUCs$^a$

<table>
<thead>
<tr>
<th></th>
<th>Mild or Worse ROP</th>
<th>Type 2 Prethreshold or Worse ROP</th>
<th>Treatment-Requiring ROP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sensitivity</td>
<td>Specificity</td>
<td>AUC</td>
</tr>
<tr>
<td><strong>PMA 31-33 wk</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grader</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.938 (0.036)</td>
<td>0.893 (0.042)</td>
<td>0.924 (0.025)</td>
</tr>
<tr>
<td>B</td>
<td>0.860 (0.054)</td>
<td>0.970 (0.031)</td>
<td>0.919 (0.029)</td>
</tr>
<tr>
<td>C</td>
<td>0.729 (0.065)</td>
<td>0.938 (0.027)</td>
<td>0.840 (0.035)</td>
</tr>
<tr>
<td><strong>PMA 35-37 wk</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grader</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.908 (0.033)</td>
<td>1.000 (0.000)</td>
<td>0.954 (0.017)</td>
</tr>
<tr>
<td>B</td>
<td>0.971 (0.020)</td>
<td>1.000 (0.000)</td>
<td>0.986 (0.010)</td>
</tr>
<tr>
<td>C</td>
<td>0.908 (0.033)</td>
<td>0.977 (0.023)</td>
<td>0.949 (0.018)</td>
</tr>
</tbody>
</table>

*Abbreviations: AUC, area under the receiver operating characteristic curve; NA, not applicable; PMA, postmenstrual age; ROP, retinopathy of prematurity; ellipses, not estimable.*

$^a$Eyes graded as “unknown” were excluded from calculations. Data were calculated for ability to detect mild or worse ROP, type 2 prethreshold or worse ROP, and treatment-requiring ROP. Interpretations by masked graders were compared with reference standard of ophthalmoscopic examination by an experienced pediatric ophthalmologist. Results are displayed as proportion of eyes (SE).

$^b$Not applicable because no treatment-requiring ROP was diagnosed by reference standard examination at 31 to 33 weeks' PMA.

### Table 3. Intergrader Reliability of Telemedical ROP Diagnosis Based on $\kappa$ and Weighted $\kappa$ Statistics Between Each Pair of Graders$^a$

<table>
<thead>
<tr>
<th></th>
<th>PMA 31-33 wk</th>
<th>PMA 35-37 wk</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\kappa$ (SE)</td>
<td>Weighted $\kappa$ (SE)</td>
</tr>
<tr>
<td>Graders</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A vs B</td>
<td>0.676 (0.072)</td>
<td>0.735 (0.056)</td>
</tr>
<tr>
<td>B vs C</td>
<td>0.545 (0.069)</td>
<td>0.548 (0.068)</td>
</tr>
<tr>
<td>A vs C</td>
<td>0.624 (0.063)</td>
<td>0.679 (0.057)</td>
</tr>
</tbody>
</table>

*Abbreviations: PMA, postmenstrual age; ROP, retinopathy of prematurity.*

$^a$Telemedical examination findings were classified as no ROP, mild ROP, type 2 prethreshold ROP, or treatment-requiring ROP.
amples of variations in accuracy and reliability among graders in this study.

QUALITY OF IMAGE CAPTURE

BY TRAINED NURSE

At 31 to 33 weeks' PMA, grader A reported “unknown” diagnoses in 24 eyes (18.8%); grader B, in 52 eyes (40.6%); and grader C, in 0 eyes (0%). At 35 to 37 weeks' PMA, grader A reported “unknown” diagnoses in 6 eyes (5.0%); grader B, in 8 eyes (6.7%); and grader C, in 0 eyes (0%) (Table 1). Ratings of image technical quality and retinal coverage are displayed in Table 5. Based on logistic regression analysis, there were no statistically significant associations between order of retinal imaging and diagnostic performance by any grader.

This study prospectively evaluates performance of telemedical ROP diagnosis by 3 expert graders compared with a reference standard of dilated ophthalmoscopy. The key findings are (1) telemedicine is highly accurate and reproducible, using images captured by a trained nurse and (2) accuracy, reliability, and image quality are better at later PMAs.

Accuracy of telemedical diagnosis in this study was high, although there were variations among graders. To understand whether this performance is adequate, it is essential to consider the underlying goal of an ROP telemedicine system. It could be argued that this should be either to fully classify retinal findings in each eye (diagnosis) or simply to identify infants with disease requiring referral for complete examination (screening). To evaluate a diagnostic approach, the accuracy of multiple graders must be examined at all levels of severity (Table 2). For example, factors leading to decreased sensitivity for detection of mild or worse ROP (eg, failure to identify stage 1 disease) are different from those leading to decreased sensitivity for detection of treatment-requiring ROP (eg, failure to identify plus disease). In a

<table>
<thead>
<tr>
<th>Grader</th>
<th>Mild or Worse ROP</th>
<th>Type 2 or Worse ROP</th>
<th>Treatment-Requiring ROP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \kappa ) (SE)</td>
<td>( \kappa ) (SE)</td>
<td>( \kappa ) (SE)</td>
</tr>
<tr>
<td>A</td>
<td>0.769 (0.239)</td>
<td>1.000 (\ldots)</td>
<td>1.000 (\ldots)</td>
</tr>
<tr>
<td>B</td>
<td>0.462 (0.332)</td>
<td>. . .</td>
<td>. . .</td>
</tr>
<tr>
<td>C</td>
<td>0.689 (0.220)</td>
<td>1.000 (\ldots)</td>
<td>1.000 (\ldots)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grader</th>
<th>Mild or Worse ROP</th>
<th>Type 2 or Worse ROP</th>
<th>Treatment-Requiring ROP</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.909 (0.093)</td>
<td>1.000 (\ldots)</td>
<td>1.000 (\ldots)</td>
</tr>
<tr>
<td>B</td>
<td>1.000 (\ldots)</td>
<td>0.769 (0.129)</td>
<td>0.913 (0.089)</td>
</tr>
<tr>
<td>C</td>
<td>1.000 (\ldots)</td>
<td>0.857 (0.099)</td>
<td>0.786 (0.118)</td>
</tr>
</tbody>
</table>

Abbreviations: PMA, postmenstrual age; ROP, retinopathy of prematurity; \( \ldots \), not estimable.

a Data were calculated for ability to detect mild or worse ROP, type 2 prethreshold or worse ROP, and treatment-requiring ROP.
b Standard error was not estimable because all repeated images received the same grading.
c Grader B gave the same diagnosis for all images at one assessment, while different diagnoses were given at the other assessment. This resulted in inestimable \( \kappa \) (SE).

Table 4. Intrgrader Reliability of Telemmedal ROP Diagnosis Based on \( \kappa \) and Weighted \( \kappa \) Statistics for Each Grader

©2007 American Medical Association. All rights reserved.
1.00 at 38 to 40 weeks. This is not surprising given that the sensitivity and specificity for image-based detection of mild or worse ROP were 0.46 and 1.00 at 32 to 34 weeks' PMA and 0.76 and 1.00 at 35 to 37 weeks' PMA than at 31 to 33 weeks. This is consistent with published results determining that sensitivity and specificity of 0.84 and 0.79 among 3 graders,18 and Roth et al20 found sensitivity and specificity of 0.82 and 0.94. For detection of type 2 or worse ROP, Ells et al17 measured sensitivity and specificity of 0.84 and 0.79 among 3 graders,18 and Roth et al20 found sensitivity and specificity of 0.82 and 0.94. For detection of type 2 or worse ROP at 35 to 37 weeks' PMA would be 0.714 and 0.777 for grader A, 0.857 and 0.529 for grader B, and 0.729 and 0.400 for grader C. No eyes with type 2 or worse ROP according to ophthalmoscopy were classified as “unknown” by any grader, although the number of these particular eyes (n=7 at 31-33 weeks, n=26 at 35-37 weeks) was likely too small to draw firm conclusions. Accuracy and reliability were uniformly high at 35 to 37 weeks’ PMA, which is the most clinically relevant period.27,28 Because of this discrepancy in performance depending on infant age, a potential strategy for ROP management might combine ophthalmoscopy and telemedicine at different times.

This is the first data set to our knowledge that has examined ROP image capture by nonophthalmic personnel. It is useful to compare our accuracy results with prior studies where images were obtained by ophthalmology personnel. For detection of mild or worse ROP, we previously found mean sensitivity and specificity of 0.84 and 0.79 among 3 graders,18 and Roth et al20 found sensitivity and specificity of 0.82 and 0.94. For detection of type 2 or worse ROP, Ells et al17 measured sensitivity and specificity of 1.00 and 0.96. Although these prior studies did not systematically categorize findings based on PMA, telemedical accuracy in the present study appears comparable or better (Table 2). This suggests that it is feasible for a neonatal nurse to capture and select acceptable images. Furthermore, technical quality was rated as either “adequate” or “possibly adequate” in 81.2% to 98.4% of images at 31 to 33 weeks’ PMA and 93.3% to 100% of images at 35 to 37 weeks’ PMA. This is the first data set to our knowledge that has examined ROP image capture by nonophthalmic personnel. It is useful to compare our accuracy results with prior studies where images were obtained by ophthalmology personnel. For detection of mild or worse ROP, we previously found mean sensitivity and specificity of 0.84 and 0.79 among 3 graders,18 and Roth et al20 found sensitivity and specificity of 0.82 and 0.94. For detection of type 2 or worse ROP, Ells et al17 measured sensitivity and specificity of 1.00 and 0.96. Although these prior studies did not systematically categorize findings based on PMA, telemedical accuracy in the present study appears comparable or better (Table 2). This suggests that it is feasible for a neonatal nurse to capture and select acceptable images. Furthermore, technical quality was rated as either “adequate” or “possibly adequate” in 81.2% to 98.4% of images at 31 to 33 weeks’ PMA and 93.3% to 100% of images at 35 to 37 weeks’ PMA. This is the first data set to our knowledge that has examined ROP image capture by nonophthalmic personnel. It is useful to compare our accuracy results with prior studies where images were obtained by ophthalmology personnel. For detection of mild or worse ROP, we previously found mean sensitivity and specificity of 0.84 and 0.79 among 3 graders,18 and Roth et al20 found sensitivity and specificity of 0.82 and 0.94. For detection of type 2 or worse ROP, Ells et al17 measured sensitivity and specificity of 1.00 and 0.96. Although these prior studies did not systematically categorize findings based on PMA, telemedical accuracy in the present study appears comparable or better (Table 2). This suggests that it is feasible for a neonatal nurse to capture and select acceptable images. Furthermore, technical quality was rated as either “adequate” or “possibly adequate” in 81.2% to 98.4% of images at 31 to 33 weeks’ PMA and 93.3% to 100% of images at 35 to 37 weeks’ PMA.
images at 35 to 37 weeks’ PMA (Table 5). Trained technicians are responsible for performing sophisticated imaging studies in fields such as radiology and cardiology. Neonatal intensive care unit nurses would be a logical choice in an ROP telemedicine strategy, given their familiarity with neonatal physiology and their ability to perform complex procedures on infants.

Intergrader agreement in this study was near perfect, with a weighted \( \kappa \) of 0.791 to 0.889 at 35 to 37 weeks’ PMA (Table 4). This is higher than previously published findings, which showed a weighted \( \kappa \) of 0.671 to 0.834 among ophthalmologist graders. This difference may be because the current study involved a clearly defined imaging protocol and used graders with extensive ROP experience. By comparison, intergrader weighted \( \kappa \) for image-based diabetic retinopathy diagnosis using the Early Treatment Diabetic Retinopathy Study (ETDRS) 7-field criterion standard was 0.41 to 0.80, depending on the lesion type. Taken together, these results suggest that reliability of telemedical ROP diagnosis is comparable with that of well-accepted diagnostic tests, even when images are captured by a neonatal nurse.

There are several study limitations: (1) Three standard photographs were taken of each eye, with up to 2 additional images at the nurse’s discretion. This may have contributed to decreased sensitivity, lower retinal coverage ratings, and more “unknown” diagnoses if graders could not visualize sufficient peripheral findings. In designing the study, we established this protocol because we felt that the majority of clinically significant disease can be identified temporally and nasally and because from a practical perspective we hoped to obtain the most useful image data in the least time. Future research regarding diagnostic and logistical trade-offs of different protocols may be useful. (2) Telemedicine graders were not permitted to manipulate image parameters. These adjustments might either increase or decrease performance, particularly if some graders were less skilled at image manipulation than others. To avoid potential confounding effects, we did not incorporate this functionality into the telemedicine system. (3) Dilated ophthalmoscopy was considered the reference standard. Although this has been the design of previous telemedicine studies, image-based examinations may not be inherently less “correct” than ophthalmoscopy. In fact, we have demonstrated that there may be diagnostic disagreements between ophthalmoscopic and image-based examinations performed by the same physician and that there is photographic evidence that image-based diagnoses may often be more accurate. This has implications for the design of future telemedicine studies.

(4) Data were analyzed by eye, although ROP diagnoses in 2 eyes of the same patient are not independent. Because ophthalmoscopy is performed on both eyes together, telemedical images were also presented side by side to simulate a real-world scenario. This was done to minimize bias favoring either examination and to permit analysis of both eyes in each infant.

We believe that this is the most extensive study of telemedical ROP diagnosis performed to date. Our results show that accuracy, reliability, and image quality are very high at later PMAs, even when images are captured by a trained nurse. Telemedicine is a promising strategy for addressing limitations of the current paradigm for ROP care, such as quality and accessibility, and we have shown that it is more cost-effective than ophthalmoscopy. Unresolved issues include medicolegal liability, engineering of telemedicine strategies into existing neonatal workflows, standardization of imaging protocols, and uncertainty about image quality at earlier PMAs.

Submitted for Publication: May 14, 2007; final revision received June 19, 2007; accepted June 19, 2007.

Correspondence: Michael F. Chiang, MD, Columbia University College of Physicians and Surgeons, 635 W 165th St, Box 92, New York, NY 10032 (chiang@dbmi.columbia.edu).

Author Contributions: Dr Chiang had full access to all the data in the study and takes responsibility for the integrity of the data and accuracy of the data analysis.

Financial Disclosure: Dr Chiang is an unpaid member of the Scientific Advisory Board of Clarity Medical Systems, Pleasanton, California.

Funding/Support: This work was supported by a Career Development Award from Research to Prevent Blindness (Dr Chiang) and by grant EY13972 from the National Eye Institute (Dr Chiang).

Role of the Sponsors: The sponsors had no role in the design and conduct of the study; in the collection, analysis, and interpretation of data; or in the preparation, review, or approval of the manuscript.

REFERENCES


The President took as the subject of his presidential address some clinical experiences of primary chronic glaucoma and the value of iridectomy. . . . He thought that an operation should be done on every case in which the patient was strong enough to stand it, and the earlier it were done the better, but he would never hesitate to operate on an eye even if the field were contracted to fixation, provided there was any vision worth saving. Even in the premonitory stage he advocated operation.