MAJOR REVIEW

Teledicine for Retinopathy of Prematurity Diagnosis: Evaluation and Challenges
Grace M. Richter, MD,1,2 Steven L. Williams, MD,1 Justin Starren, MD, PhD,3,4 John T. Flynn, MD,1 and Michael F. Chiang, MD1,4

Abstract. Retinopathy of prematurity (ROP) is a vasoproliferative disorder affecting low birth weight infants. Although timely diagnosis and treatment can significantly reduce the risk of severe complications, ROP remains a leading cause of childhood blindness worldwide. Limitations of current disease management strategies include extensive travel and logistical coordination requirements for ophthalmologists and neonatologists, decreasing availability of adequately trained ophthalmologists at the point of care, variability in how retinal findings are diagnosed and documented, and a growing need for ROP care worldwide. Store-and-forward telemedicine is an emerging technology by which medical data are captured for subsequent interpretation by a remote expert. This has potential to improve accessibility, quality, and cost of ROP management. In this article, we summarize the current evaluation data on applications of telemedicine for ROP, particularly involving the diagnostic accuracy and reliability of remote image interpretation by experts. We also address challenges such as the cost-effectiveness of telemedicine, and highlight potential barriers to implementation of these systems. Understanding these principles is essential to determine future directions in research and development of telemedicine systems for ROP, as well as for other ophthalmic diseases. (Surv Ophthalmol 54:671–685, 2009. © 2009 Elsevier Inc. All rights reserved.)

Key words. retinopathy of prematurity • telemedicine • diagnostic imaging • medical informatics • retina • diagnostic techniques and procedures • blindness

I. Retinopathy of Prematurity
A. INTRODUCTION AND EPIDEMIOLOGY

Retinopathy of prematurity (ROP) is a vasoproliferative retinal disorder affecting low birth weight infants. Treatment criteria for severe disease, using laser photocoagulation and cryotherapy, have been established through the Cryotherapy for Retinopathy of Prematurity (CRYO-ROP) and Early Treatment for Retinopathy of Prematurity (ETROP) trials.18,25,30,58 Nonetheless, of the approximately 4 million infants born each year in the United States, roughly 2,100 premature babies are affected by long-term sequelae of ROP such as strabismus, myopia, and retinal detachment.52,59 Among these infants, 400–900 develop ROP-related blindness annually.59 Management of ROP has been supported by universal, evidence-based standards. In particular, research on the natural progression of disease has led to strategies for diagnosis and classification.57,58
Development of an international classification system has standardized reporting of the severity, extent, and location of retinal abnormalities in ROP.\textsuperscript{17,37} Policy statements for examination of at-risk infants have been published jointly by the American Academy of Ophthalmology, American Academy of Pediatrics, and American Association for Pediatric Ophthalmology and Strabismus.\textsuperscript{67,68} These statements recommend that at-risk infants with birth weight (BW) < 1,500 grams or gestational age < 30 weeks, and infants with BW 1,500–2,000 grams or gestational age > 30 weeks with an unstable clinical course, should receive dilated ophthalmoscopic examinations.

ROP remains a leading cause of treatable childhood blindness in the United States and throughout the world. In developed countries, the number of infants at risk for ROP has been increasing. This is in part because of higher premature birth rates due to reasons such as assisted conception, increasing maternal age, possible genetic etiologies, and other socioeconomic issues.\textsuperscript{16,70} In the United States, the rate of prematurity has risen from 9.4\% of all births in 1981 to 12.7\% in 2005.\textsuperscript{34,52} In regions such as Latin America, Eastern Europe, and Asia, the number of ROP cases has grown dramatically because of higher overall birth rates, as well as improved neonatal survival from greater availability of neonatal care.\textsuperscript{79,80} Concerns have been raised about an emerging international ROP “epidemic” due to a shortage of adequately trained ophthalmologists and persistent variability in quality of neonatal care.\textsuperscript{8,27,29,75} There is also evidence that larger infants may develop ROP in some regions, highlighting the need for region-specific ROP management guidelines.\textsuperscript{9,27,28,75}

II. Telemedicine

A. GENERAL PRINCIPLES

Telemedicine is defined as the use of information technologies to support health care between participants who are separated from each other.\textsuperscript{25} It has potential to improve the accessibility, quality, and cost of healthcare, and may also contribute to medical education and research. These consultations can be divided into two categories: \textit{synchronous} telemedicine utilizes telecommunications for real-time interactions between participants (e.g., video conferencing), and \textit{store-and-forward} telemedicine involves the capture of patient data for subsequent interpretation by a remote expert (e.g., digital radiology). Telemedicine programs emerged in the United States over 40 years ago with mixed success. Many of these projects were discontinued after their initial period of funding due to technological limitations and lack of financial sustainability.\textsuperscript{5,25} Advances in information technology, along with growing concerns about rising costs and inequities in healthcare accessibility, have revived interest in telemedicine. In particular, large-scale store-and-forward applications have been implemented in specialties such as radiology and pathology, which rely heavily on the interpretation of data from image-based modalities.\textsuperscript{33,45}

B. APPLICATIONS TO ROP

Telemedicine has the potential to address some of the challenges associated with current ROP management and quality of care. In store-and-forward telemedicine programs for ROP, images would likely be captured by trained neonatal personnel for subsequent grading by a remote ophthalmologist. This could improve travel time for ophthalmologists, logistical coordination with neonatal staff, and accessibility to expert care for patients. Serial retinal imaging may offer a more objective method for documentation of disease findings and progression. In addition, widespread retinal imaging for ROP would provide opportunities to create digital
libraries for educational and research purposes. This could potentially improve the uniformity of ROP diagnosis. Retinal photography might also cause less physiologic stress to infants than ophthalmoscopy with scleral depression. Of note, digital retinal imaging has been successfully utilized in other pediatric retinal conditions such as retinoblastoma and shaken baby syndrome.

III. Feasibility and Evaluation Studies

A. STUDY DESIGN CONSIDERATIONS

Several studies have examined the application of telemedicine for ROP diagnosis (Tables 1, 2, and 3). In general, these studies have compared performance of image-based diagnosis by remote experts to a reference standard of dilated examination by an ophthalmologist. Recognizing key differences in experimental design is necessary to compare these studies.

First, imaging methods have differed based on:

1) Type of device used. This has included wide-angle contact cameras (Fig. 1) (RetCam; Clarity Medical Systems, Pleasanton, CA) as well as narrow-angle non-contact cameras (Nidek NM200-D; Nidek, Aichi, Japan).

2) Number of images captured from each eye.

3) Expertise of the camera operator. This has included trained neonatal nurses, ophthalmic photographers, and ophthalmologists.

Second, conditions for telemedical image grading have varied based on:

1) Expertise of graders. These have included general ophthalmologists, as well as retinal specialists and pediatric ophthalmologists.

2) Clinical information provided to telemedical graders. Some studies have apparently displayed only images, whereas others have provided additional data such as demographics and clinical history.

3) Presentation of images. Some studies have displayed them one eye at a time, whereas others have shown them both eyes at a time to simulate real-world clinical situations.

Third, study populations have varied from a selection of ROP cases to consecutively recruited infants in real-world NICU populations. Fourth, outcome measures have differed based on the diagnostic cut-off. This has included detection of any ROP, moderate or worse ROP, or severe ROP. Finally, some studies have examined telemedicine performance at individual examinations, whereas others have analyzed performance over multiple longitudinal examinations.

B. ACCURACY

1. Accuracy of Diagnosing Any ROP

Several studies have evaluated telemedicine for the detection of mild or worse ROP using multiple wide-angle retinal images (RetCam-120 or RetCam-II; Clarity Medical Systems), compared to a reference standard of dilated ophthalmoscopy by an expert (Table 1). Roth et al (2001; 100 exams from 32 infants), found sensitivity of 0.82 and a specificity of 0.94. Shah et al (2006; 87 exams from 27 infants) reported similar findings, with sensitivity of 0.86 and a specificity of 0.92. Chiang et al (2006; 163 exams from 64 infants) evaluated telemedicine for detection of any ROP by three image graders (one general ophthalmologist and two retinal specialists), and demonstrated a range in sensitivity of 0.82–0.86 and specificity of 0.49–0.96. When data from the general ophthalmologist were excluded, the range of sensitivity and specificity was 0.82–0.85 and 0.91–0.96, respectively. These studies attributed lower sensitivity to missed cases of mild peripheral ROP that were not captured during retinal imaging.

Yen et al (2002; 82 exams from 25 infants) demonstrated a higher accuracy for detection of any ROP in infants at later post-menstrual ages (PMA). For infants at 32–34 weeks PMA, sensitivity was 0.47 and specificity was 1.00. For infants at 38–40 weeks PMA, sensitivity was 0.76 and specificity was 1.00. This tendency toward higher accuracy at later PMA was also demonstrated prospectively by Chiang et al (2007; 248 exams from 68 infants), who evaluated accuracy of three pediatric retinal specialists. Sensitivity improved from 0.73–0.94 (at 31–33 weeks PMA) to 0.91–0.97 (at 35–37 weeks PMA), and specificity improved from 0.89–0.97 (at 31–33 weeks PMA) to 0.98–1.00 (at 35–37 weeks PMA). Finally, Dhaliwal et al (2009; 245 exams from 81 infants) conducted a masked, double-observer prospective study in which both wide-angle retinal imaging and ophthalmoscopy were performed by two pediatric ophthalmologists who were randomized to the examination technique. Almost all examinations were performed between 32-36 weeks PMA. Compared to a reference standard of ophthalmoscopy, the sensitivity of telemedicine for diagnosis of any ROP was 0.60 and specificity was 0.91. Higher accuracy at later PMAs
<table>
<thead>
<tr>
<th>Study</th>
<th># infants, # exams (proportion of exams with ROP)</th>
<th>a) camera and # images per eye</th>
<th>b) camera operator</th>
<th>c) image grader</th>
<th>d) clinical data provided to graders; uni- or bilateral presentation of eyes</th>
<th>e) study design and population</th>
<th>Outcome</th>
<th>Measure</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roth et al&lt;sup&gt;59&lt;/sup&gt; (2001)</td>
<td>32 infants, 100 exams (68%)</td>
<td>a) RetCam-120; 1-6 images per eye</td>
<td>b) nurse or ophthalmic photographer</td>
<td>c) consensus of 1 pediatric ophthalmologist and 1 ophthalmology resident</td>
<td>d) none; unilateral</td>
<td>e) consecutive enrollment of screening population (Miami, FL)</td>
<td>any ROP</td>
<td></td>
<td></td>
<td>0.82</td>
</tr>
<tr>
<td>Yen et al&lt;sup&gt;77&lt;/sup&gt; (2002)</td>
<td>25 infants, 96 exams (55%)</td>
<td>a) RetCam-120; 4-8 images per eye</td>
<td>b) nurse</td>
<td>c) consensus of 1 pediatric ophthalmologist and 1 pediatric ophthalmologist fellow</td>
<td>d) none; unilateral (few bilateral)</td>
<td>e) consecutive enrollment of screening population (Miami, FL)</td>
<td>any ROP at 32--34 wks PMA</td>
<td></td>
<td></td>
<td>0.46</td>
</tr>
<tr>
<td>Shah et al&lt;sup&gt;65&lt;/sup&gt; (2006)</td>
<td>27 infants, 87 exams (72%)</td>
<td>a) RetCam-120; 4-5 images per eye</td>
<td>b) vitreoretinal specialist</td>
<td>c) 1 vitreoretinal specialist</td>
<td>d) none; unilateral</td>
<td>e) consecutive enrollment of screening population (India)</td>
<td>any ROP</td>
<td></td>
<td></td>
<td>0.86</td>
</tr>
<tr>
<td>Chiang et al&lt;sup&gt;15&lt;/sup&gt; (2006)</td>
<td>64 infants, 163 exams (61%)</td>
<td>a) RetCam-120; 1-7 images/eye</td>
<td>b) ophthalmic photographer</td>
<td>c) 2 retinal specialists and 1 general ophthalmologist with limited ROP experience</td>
<td>d) none; unilateral</td>
<td>e) consecutive enrollment of screening population (Miami, FL)</td>
<td>any ROP</td>
<td>0.82-0.86</td>
<td></td>
<td>0.49-0.96</td>
</tr>
<tr>
<td>Chiang et al&lt;sup&gt;15&lt;/sup&gt; (2007)</td>
<td>68 infants, 248 exams (50%)</td>
<td>a) RetCam-II; 3-5 images/eye</td>
<td>b) nurse</td>
<td>c) 3 retinal specialists experienced in ROP</td>
<td>d) BW, GA, PMA; bilateral</td>
<td>e) consecutive enrollment of screening population (New York, NY)</td>
<td>any ROP at 31-33 wks PMA</td>
<td>0.73-0.94</td>
<td></td>
<td>0.89-0.97</td>
</tr>
<tr>
<td>Dhaliwal et al&lt;sup&gt;21&lt;/sup&gt; (2009)</td>
<td>81 infants, 245 exams (18%)</td>
<td>a) RetCam-II; 5-15 images/eye</td>
<td>b) pediatric ophthalmologist</td>
<td>c) pediatric ophthalmologist</td>
<td>d) BW, GA, PMA, all other data available during routine care; bilateral</td>
<td>e) consecutive enrollment of screening population (Edinburgh, United Kingdom)</td>
<td>any ROP</td>
<td></td>
<td></td>
<td>0.60</td>
</tr>
</tbody>
</table>

BW = birth weight; GA = gestational age; PMA = post-menstrual age; ROP = Retinopathy of prematurity.
has been attributed to improved image quality and capture because of larger palpebral fissures, larger eyes, and less corneal and vitreous haze.15,79,80

2. Accuracy of Diagnosing Moderate to Severe ROP

Other studies have examined the accuracy of telemedicine for diagnosing moderate to severe ROP using multiple wide-angle images (Table 2). In a group of pre-selected cases with relatively severe ROP, Schwartz et al (2000; 19 exams from 10 infants) reported that 18/19 (95%) eyes showed agreement between telemedicine and ophthalmoscopy for plus disease diagnosis and 17/19 (89%) eyes showed agreement between telemedicine and ophthalmoscopy for presence of prethreshold or worse ROP.65

Several studies have examined cohorts of consecutively enrolled infants in real-world NICU situations, and compared telemedicine interpretation to a reference standard of dilated ophthalmoscopy at the time of each individual examination. Wu et al (2006; 86 exams from 43 infants) examined accuracy of telemedicine for detection of prethreshold or worse ROP, and reported sensitivity of 1.00 and specificity of 0.98.70 Similarly, Chiang et al (2006; 163 exams from 64 infants) examined accuracy for detection of type 2 or worse ROP by three image graders, and found sensitivity ranging from 0.72–0.83 and specificity ranging from 0.90–0.99.13 For detection of treatment-requiring ROP (defined as type 1 or worse disease), the same study found sensitivity of 0.85–0.90 and specificity of 0.95–0.97.14 In a follow-up study, Chiang et al (2007; 248 exams from 68 infants), prospectively examined performance by three pediatric retinal specialist graders for detection of type 2 or worse ROP. At 31–33 weeks PMA, sensitivity was 0.71–0.86 and specificity was 0.93–0.97.15 At 35–37 weeks PMA, sensitivity was 1.00–1.00 and specificity was 0.85–0.94.15

Dhaliwal et al (2009; 245 exams from 81 infants) conducted a masked, double-observer prospective study in which both wide-angle retinal imaging and ophthalmoscopy were performed by two pediatric ophthalmologists who were randomized to the examination technique. Almost all examinations were performed between 32–36 weeks PMA. Sensitivity of telemedicine for diagnosis of stage 3 ROP was 0.57, and specificity was 0.98. Sensitivity for diagnosis of plus disease was 0.80, and specificity was 0.98. Absolute agreement between ophthalmoscopy and telemedicine was 0.96 for detection of stage 3 ROP, and 0.97 for detection of plus disease.21

Finally, two studies have assessed diagnostic performance by analyzing the development of moderate to severe ROP over multiple longitudinal examinations (Table 3). These studies consisted of paired serial exams by ophthalmoscopy and telemedicine, and determined whether telemedicine could detect the presence of moderate to severe ROP at any point during the cumulative clinical course. Ells et al (2003; 371 exams from 36 infants) examined detection of “referral-warranted ROP” (defined similarly to type 2 or worse disease: any ROP in zone 1, presence of plus disease, or presence of any stage 3 ROP at any time during infant’s hospital course) at any time during longitudinal inpatient examinations, and found sensitivity of 1.00 and specificity of 0.96.24 The prospective, multi-center Photographic Screening for ROP (Photo-ROP) study (2008; 300 exams from 51 infants) examined detection of “clinically significant ROP” [defined similarly to type 2 or worse disease, or inability to rule-out disease: a) zone 1, any ROP; without vascular dilation or tortuosity; b) zone II, stage 2, with up to one quadrant of vascular dilation and tortuosity; c) zone II, stage 3, with up to one quadrant of vascular dilation and tortuosity; d) any vascular dilation and tortuosity noted in eyes for which ridge characteristics were not interpretable (not imaged or poor image quality); or e) any ROP noted in eyes for which disc features (plus disease) were not interpretable (not imaged or poor image quality)] at any time during multiple longitudinal examinations.3,4 This study showed that “clinically significant ROP” was detected with sensitivity of 0.92 and specificity of 0.37.4

Overall, these findings suggest that the factors limiting accuracy for diagnosis of moderate to severe ROP (e.g., failure to detect plus disease) are likely to be different from those that limit the accuracy for diagnosis of any ROP (e.g., failure to detect peripheral disease).

Lorenz et al (2009; 6460 exams from 1222 infants) conducted a six-year prospective study at five NICUs in Germany, in which local general (4 NICUs) or pediatric (1 NICU) ophthalmologists were asked to continue standard ophthalmoscopic exams while also capturing wide-angle retinal images that were interpreted at a central reading center. All infants found to have “suspected treatment-requiring ROP” (defined as type 1 ROP or worse, or as anything else felt to represent possible treatment-requiring ROP that could not be reliably classified from retinal images) by telemedicine were referred for complete ophthalmoscopic examination. Sensitivity of telemedicine for detecting suspected treatment-requiring ROP was 1.00, positive predictive value for detecting...
<table>
<thead>
<tr>
<th>Study</th>
<th># infants, # eye exams (portion of exams with ROP)</th>
<th>a) camera and # images per eye</th>
<th>b) camera operator</th>
<th>c) image grader</th>
<th>d) clinical data provided to graders; unilateral or bilateral presentation of eyes</th>
<th>e) study design and population</th>
<th>Outcome Measure</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schwartz et al (2000)</td>
<td>10 infants, 19 exams (100%) a) RetCam-120; ~9 images per eye b) ophthalmologist c) 2 specialists d) BW and PMA; bilateral e) selection of 19 eyes with severe disease (Los Angeles, CA)</td>
<td>plus disease</td>
<td>1.00</td>
<td>0 (n=1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chiang et al (2006)</td>
<td>64 infants, 163 exams (61%) a) RetCam-120; 1–7 images per eye b) ophthalmic photographer c) 2 retinal specialists and 1 general ophthalmologist with limited ROP experience d) none; unilateral e) consecutive enrollment of screening population (Miami, Florida)</td>
<td>type 2 or worse ROP</td>
<td>0.72–0.83</td>
<td>0.90–0.99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wu et al (2006)</td>
<td>43 infants, 86 exams (42%) a) RetCam-II; 1–5 images per eye b) pediatric ophthalmologist or ophthalmic photographer c) 1 pediatric ophthalmologist d) BW, GA, PMA, race, sex, and birth multiplicity; unknown e) consecutive enrollment of screening population (Boston, MA)</td>
<td>treatment-requiring ROP</td>
<td>0.85–0.90</td>
<td>0.95–0.97</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chiang et al (2007)</td>
<td>68 infants, 248 exams (50%) a) RetCam-II; 3–5 images per eye b) nurse c) 3 retinal specialists experienced in ROP d) BW, GA, PMA; bilateral e) consecutive enrollment of screening population (New York, NY)</td>
<td>prethreshold or worse ROP</td>
<td>1.00</td>
<td>0.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dhaliwal et al (2009)</td>
<td>81 infants, 245 exams (18%) a) RetCam-II; 5-15 images/eye b) pediatric ophthalmologist c) pediatric ophthalmologist d) BW, GA, PMA, all other data available during routine care; bilateral e) consecutive enrollment of screening population (Edinburgh, United Kingdom)</td>
<td>stage 3 or worse ROP</td>
<td>0.57</td>
<td>0.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BW = birth weight; GA = gestational age; NA = not applicable; PMA = post-menstrual age; ROP = Retinopathy of prematurity.

*Study population consisted of eyes that were pre-selected for severe ROP. Telemedicine agreed with ophthalmoscopy in 18/19 (95%) eyes for diagnosis of plus disease, and in 17/19 (89%) eyes for diagnosis of prethreshold or worse disease.
treatment-requiring ROP at the initial referral was 0.82, and all treatment-requiring ROP was felt to be detected in time.50 Lastly, a telemedicine program involving four NICUs, in which nurses were trained to capture serial images from all at-risk infants, has been ongoing at Stanford University since 2005. That study did not compare results of each telemedicine interpretation to opthalmoscopic findings. However, among 97 infants (194 eyes) examined in that program, no known cases of referral-warranted or treatment-requiring ROP were missed, the positive predictive value of telemedicine for identifying treatment-requiring ROP was 85.7%, and there were no adverse outcomes such as retinal detachment.55,71

C. RELIABILITY

1. Inter-grader Reliability

Inter-grader reliability of telemedicine has been evaluated by comparing the diagnostic responses of multiple expert image graders who have evaluated the same images. One study demonstrated weighted kappa of 0.67–0.83 for pairs of graders reviewing the same images,15 whereas another found weighted kappa of 0.55–0.74 among images at 31–33 weeks PMA and 0.79–0.89 at 35–37 weeks PMA.15 Dhaliwal et al found 95% absolute agreement between two telemedicine graders on presence or absence of plus disease, 94% agreement on stage of ROP, and 97% agreement on zone of ROP.21 For comparison, the Early Treatment Diabetic Retinopathy Study (ETDRS) seven-field criterion standard demonstrated weighted kappa of 0.41–0.80 for inter-grader reliability of image-based diabetic retinopathy diagnosis, depending on the lesion type.22 In a study involving fluorescein angiogram interpretation for photodynamic therapy eligibility in age-related macular degeneration (AMD), the kappa was 0.37–0.40 for inter-grader reliability.35 Although these comparisons should be taken with caution because they are made in different retinal diseases, these findings suggest that inter-grader reliability for telemedical ROP detection is comparable to or better than that of other well-accepted diagnostic tests in ophthalmology.

2. Intra-grader Reliability

Intra-grader reliability of telemedicine has been evaluated by analyzing the diagnosis of randomly-repeated images presented to the same grader. Chiang et al demonstrated near-perfect to perfect agreement in a study involving three retinal specialist graders, with intra-grader kappa of 0.91–1.00 for detection of mild or worse ROP at 35–37 weeks PMA, and kappa of 0.79–1.00 for detection of treatment-requiring ROP at 35–37 weeks PMA.15 For comparison, the kappa for intra-grader reliability of fluorescein angiogram interpretation for determining photodynamic therapy eligibility in AMD has been found to be 0.44–0.89,35 and the intra-grader concordance of fluorescein angiography for detection of classic choroidal neovascularization has been shown to be 0.66–1.00.42

D. IMAGE QUALITY

Assessment of image quality is an important consideration for telemedical diagnosis of ROP. Ells et al found that wide-angle images were captured successfully in 96% of examinations, and that 94% of retinal images sets could be graded remotely to identify plus disease, zone, or stage of ROP. In the remaining 6%, repeatable readable photographs were obtained within 1–4 weeks.24 Wu et al reported that 21% of initial retinal images and 22% of repeated images were considered unacceptable for diagnosis by graders.79 The Photo-ROP Cooperative Group found that 8% of image sets were uninterpretable.4 Chiang et al showed that 3 telemedicine graders reported an “unknown” diagnosis (due to inadequate image quality or incomplete retinal coverage) in 0–41% of eye exams among infants at 31–33 weeks PMA, and in 0–7% of eye exams among infants at 35–37 weeks PMA.15 Lorenz et al found that, among 6460 telemedicine imaging sessions conducted at five NICUs over a six-year period, nearly 98% were of high enough quality to evaluate for degree of retinal maturity, and for risk of progression to “suspected treatment-requiring ROP” within 1–2 weeks.50 Studies have observed that heavy fundus pigmentation and factors associated with younger infants, such as increased corneal and vitreous haze, and smaller palpebral fissures and pupils, could lead to decreased image quality.15,79,80

E. EXAMINATION VS SCREENING

In principle, telemedicine could either be used to perform ROP “examination” by reporting comprehensive retinal findings, or to perform “screening” by identifying infants with sufficiently severe disease to require referral for full examination.74 Some of the discussed telemedicine studies have evaluated the efficacy of using multiple wide-angle images from an “examination” perspective (Table 1), whereas others have evaluated performance of “screening”-type functions such as detection of moderate or worse ROP requiring referral (Tables 2 and 3). It is not clear whether the ultimate role of real-world telemedicine systems should be to substitute for full examinations, or whether it
### TABLE 3

**Summary of Reported Sensitivities and Specificities of Telemedicine for Detection of Clinically Relevant ROP, as Compared to Reference Standard of Indirect Ophthalmoscopy, Over the Infant’s Cumulative Clinical Course.**

<table>
<thead>
<tr>
<th>Study</th>
<th># infants, # eye exams (proportion of exams with ROP)</th>
<th>a) camera and # images per eye</th>
<th>b) camera operator</th>
<th>c) image grader</th>
<th>d) clinical data provided to graders; unilateral or bilateral presentation of eyes</th>
<th>e) study design and population</th>
<th>Outcome Measure</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ells et al&lt;sup&gt;2&lt;/sup&gt; (2003)</td>
<td>36 infants, 371 exams (32% eyes with “referral-warranted ROP”)</td>
<td>a) RetCam-120; 5 images per eye</td>
<td>b) pediatric ophthalmologist</td>
<td>c) pediatric ophthalmologist</td>
<td>d) none; unilateral</td>
<td>e) consecutive enrollment of eligible population (Alberta, Canada)</td>
<td>referral-warranted ROP&lt;sup&gt;*&lt;/sup&gt;</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td>Photo-ROP Cooperative Group&lt;sup&gt;3&lt;/sup&gt; (2008)</td>
<td>51 infants, 300 exams (58% eyes with “clinically significant ROP”)</td>
<td>a) RetCam-120; 5 images per eye</td>
<td>b) ophthalmologist</td>
<td>c) 2 ROP specialists, by consensus</td>
<td>d) none; unilateral</td>
<td>e) consecutive enrollment of eligible population at 6 U.S. study centers</td>
<td>clinically-significant ROP&lt;sup&gt;y&lt;/sup&gt; or worse</td>
<td>0.92</td>
<td>0.37</td>
</tr>
<tr>
<td>Lorenz et al&lt;sup&gt;50&lt;/sup&gt; (2009)</td>
<td>1222 infants, 6460 exams (28% infants with ROP)</td>
<td>a) RetCam-120; 1-60 images/eye</td>
<td>b) general pediatric ophthalmologist</td>
<td>c) pediatric ophthalmologist</td>
<td>d) BW, GA, PMA, all other data available during routine care; bilateral</td>
<td>e) consecutive enrollment of eligible population at 5 German study centers</td>
<td>suspected treatment-requiring ROP&lt;sup&gt;*&lt;/sup&gt;</td>
<td>1.00</td>
<td>–</td>
</tr>
</tbody>
</table>

<sup>BW = birth weight; GA = gestational age; PMA = post-menstrual age; ROP = Rectionopathy of prematurity.  
*Referral-warranted ROP is defined as any ROP in zone 1, presence of plus disease, or presence of any stage 3 ROP at any time during infant’s hospital course.  
Clinically significant ROP defined as: a) zone 1, any ROP, without vascular dilation or tortuosity; b) zone II, stage 2, with up to one quadrant of vascular dilation and tortuosity; c) zone II, stage 3, with up to one quadrant of vascular dilation and tortuosity; d) any vascular dilation and tortuosity noted in eyes for which disc features (plus disease) were not interpretable (not imaged or poor image quality); or e) any ROP noted in eyes for which disc features (plus disease) were not interpretable (not imaged or poor image quality).  
**Suspected treatment-requiring ROP** was defined as type 1 ROP or worse, or as anything else felt to represent possible treatment-requiring ROP that could not be reliably classified from retinal images. Positive predictive value for detecting treatment-requiring ROP at the initial referral was 0.82, and all treatment-requiring ROP was felt to be detected in time.</sup>
should be to screen for the subset of infants with disease requiring evaluation by experts.

Two telemedicine studies have explored the efficacy of ROP “screening” using posterior pole images. Lajoie et al (2008; 248 exams from 67 infants) examined the detection of moderate to severe ROP from a single wide-angle, posterior pole image by three experienced graders. For detection of type 2 or worse ROP from single images, sensitivity was 0.69–1.00, and specificity was 0.87–1.00. For detection of plus disease from single images, sensitivity was 1.00–1.00, and specificity was 0.79–0.95. Skalet et al (2008; 110 exams from 28 infants) conducted a feasibility study involving capture of three images from each eye (1 posterior and 2 peripheral) by a narrow-angle retinal camera (Nidek NM200-D) for detection of “referral-warranted ROP” (any ROP in zone 1, presence of plus disease, or presence of any stage 3 ROP at any time during infant’s hospital course) at any time during longitudinal examinations. Images were captured at the NICU bedside by nurses in Peru, and interpreted by five experienced graders. Images were successfully captured in 56/58 (97%) sessions, and preliminary studies showed that sensitivity was 0.46–0.95 and specificity was 0.62–0.96 for detection of “referral-warranted ROP.” Taken together, these pilot studies raise the possibility that retinal imaging might eventually be used to support single-image telemedical screening for moderate to severe ROP. This may be applicable in areas with limited access to resources for ROP care.

### F. COST EFFECTIVENESS

The long-term viability of a telemedicine program for ROP will also depend on economic factors. Two studies, one from the US and another from the UK, have directly compared the cost-effectiveness of telemedicine versus ophthalmoscopy for ROP management. Both studies assumed the diagnostic accuracy of telemedicine and ophthalmoscopy to be identical, based on the notion that previous research does not suggest one modality to systematically outperform the other. These studies were similar in that they created decision tree models based on published research involving ROP epidemiology and treatment efficacy. Both analyses found that several telemedicine strategies were more cost-effective than current ophthalmoscopic management strategies.

The US study modeled two scenarios: 1) standard examinations by an experienced ophthalmologist, and 2) telemedicine examinations by non-ophthalmic personnel using wide-angle imaging devices. It was conducted from a third-party payer perspective using 2006 Medicare costs for ophthalmoscopic and photographic examinations, and did not consider non-reimbursed costs such as equipment and training. Accuracy of telemedical diagnosis, as well
as outcome data from clinical management, was derived from published studies.\textsuperscript{13,14,29,30,54–56} This showed that telemedicine for ROP management costs US$3,193 per quality-adjusted life year (QALY) gained, whereas traditional management involving serial ophthalmoscopy costs $5,617 per QALY gained.\textsuperscript{38} Sensitivity analysis also showed that telemedicine had greater cost-effectiveness than ophthalmoscopy over a range of parameter values including accuracies of telemedicine and ophthalmoscopy, incidence of treatment-requiring ROP, and percentage of telemedicine images that were readable.

The UK study modeled five potential strategies for ROP surveillance: 1) telemedicine with both image capture and grading by visiting nurses, 2) telemedicine with image capture by visiting nurses and image grading by remote ophthalmologists, 3) telemedicine with both image capture and grading by neonatal nurses, 4) telemedicine with image capture by neonatal nurses and image grading by remote ophthalmologists, and 5) traditional serial bedside ophthalmoscopy by an ophthalmologist. This study was conducted from a single-party payer perspective, and clinical outcome data were derived from published studies.\textsuperscript{30,31,41,56–58} Compared to traditional bedside ophthalmoscopy (\£321 per infant examined), this study found that telemedicine with both image capture and grading by visiting nurses (\£172 per infant examined) and telemedicine with image capture by visiting nurses and image grading by remote ophthalmologists (\£201 per infant examined) would be the most cost-effective strategies.\textsuperscript{7} Although using trained nurses to capture adequate-quality retinal images has been shown to be effective,\textsuperscript{15,55,66,71,80} the accuracy of trained nurses for image grading must be examined with additional research.

To anticipate the impact of a telemedicine program for ROP, it may also be necessary to consider societal costs and benefits that are more difficult to quantify. For example, telemedicine may offer benefits for ophthalmologists such as time savings, subjectively greater provider satisfaction, and availability of retinal images for secondary purposes such as education, research, and quality improvement. Richter et al found that telemedicine examinations required significantly less physician time than standard indirect ophthalmoscopy (1.02–1.75 minutes per telemedicine exam of 125 infants vs. 4.17–6.63 minutes per ophthalmoscopic exam of 72–105 infants by three ophthalmologists).\textsuperscript{61} These differences might contribute to economic benefits. Telemedicine might lead to either decreased risk for medicolegal liability because of improved documentation of examination findings or to increased risk due to wider availability of objective imaging data for scrutiny. Finally, telemedicine might cause dissatisfaction for some families who prefer the possibility of face-to-face discussion with an ophthalmologist. These factors may require further study.

### IV. Challenges

**A. WHAT IS THE TRUE GOLD STANDARD?**

In all of the studies discussed herein, accuracy of telemedical ROP diagnosis was compared to a reference standard of dilated indirect ophthalmoscopy by an experienced examiner. Yet it is not clear that this represents a true gold standard. For instance, ophthalmoscopy might inadvertently lead to inadequate assessments due to infant movements and small pupils, and paper-based documentation of retinal findings from indirect ophthalmoscopy is also less objective and subject to observer variation.\textsuperscript{74} An epidemiological study in Australia and New Zealand suggested that the varying rates of ROP incidence across geographical regions may result from observer bias.\textsuperscript{19,20} A clinical study to directly examine the accuracy of serial indirect ophthalmoscopy by multiple examiners on the same infants has never been conducted, and might be impractical because of concerns about infant safety.\textsuperscript{51}

Scott et al performed a study that controlled for inter-physician variability by examining ophthalmoscopic and telemedical examinations of 67 infants by the same graders. There was substantial to near-perfect agreement in these diagnostic modalities, with absolute agreement of 86% (178/206 eyes) and kappa of 0.66–0.85 between ophthalmoscopy and telemedicine.\textsuperscript{66} Among the 14% (28/206 eyes) discrepancies in this study, some cases provided photographic documentation suggesting that ophthalmoscopic examination may have missed signs of mild ROP. In addition, there were several discrepancies between presence of zone-1 ROP and presence of plus disease, in which telemedicine may have provided the theoretical advantages of allowing examiners to review their diagnoses, make more exact measurements of anatomical landmarks defining zone 1 of the retina, and directly compare images to the standard photograph for plus disease.\textsuperscript{66}

Ells et al suggested that telemedicine may detect moderate to severe ROP before ophthalmoscopy. This was based on their finding that “referral-warranted ROP” was diagnosed earlier by telemedicine than ophthalmoscopy in 43% (10/23) eyes, whereas it was diagnosed earlier by
ophthalmoscopy than telemedicine in only 13% (2/23) eyes.

Overall, research has suggested that there may be non-trivial variability in ROP diagnosis, even among experts.\(^{1,11,60,77}\) During ophthalmoscopic examinations in the CRYO-ROP trial, 12% of eyes diagnosed with threshold disease by one study-certified expert were diagnosed with less than threshold disease by a second certified expert who was asked to perform a confirmatory examination.\(^{60}\) The second examiner was unmasked to the fact that the first examiner had diagnosed threshold ROP, and this level of disagreement may have been even higher in the setting of a fully masked study design. A study of image-based plus disease detection found that 22 experts agreed on the same diagnosis (plus vs not plus) in only 21% (7/34) of images, and that the mean kappa for each expert compared to all others was 0.19–0.66.\(^{11}\) Compared to a reference standard defined as the diagnosis selected by a majority of experts, the sensitivity of those 22 experts for plus disease diagnosis ranged from 0.31–1.00 and the specificity ranged from 0.57–1.00.\(^{26}\) Limitations in reliability may be caused by differing interpretations regarding the level of dilation and tortuosity sufficient for plus disease. Widespread retinal imaging, with subsequent development of image libraries and computer-based tools for diagnosis of plus disease based on quantitative vessel properties, may result in improved consistency.\(^{10,26,39,40,43,76,78}\)

**B. SAFETY AND TECHNICAL CONSIDERATIONS IN RETINAL IMAGING**

The physiologic stress caused to premature infants by indirect ophthalmoscopy are well known.\(^{6,32,46,48,63,64,73}\) Several studies have examined the relative safety of wide-angle retinal imaging using contact retinal cameras. In a pilot study comparing the systemic effects of three methods (imaging with the RetCam-120, indirect ophthalmoscopy with scleral depression, and indirect ophthalmoscopy without scleral depression), Mehta et al reported a greater stress response when retinal imaging or scleral depression were employed.\(^{53}\) A larger study by Mukherjee et al compared infant stress response during RetCam-120 imaging to the response during indirect ophthalmoscopy with scleral depression. This showed that both modalities were stressful to infants, but that imaging was significantly less stressful even though it required approximately twice as much time.\(^{54}\)

With regard to ocular complications, there has been one reported case of retinal hemorrhages following RetCam imaging and ophthalmoscopy with scleral depression.\(^{1}\) A follow-up investigation which involved performing indirect ophthalmoscopy 1 hour after ROP RetCam imaging in 50 eyes of 25 premature infants reported no retinal hemorrhages.\(^{2}\) Some papers involving telemedical ROP diagnosis have specifically reported no cases of ocular or systemic complications following contact retinal imaging.\(^{15,80}\) Although no ocular complications have been reported in other telemedicine studies, safety was not an explicit outcome measure in those other studies.

In addition, there are technical considerations in retinal imaging. Some investigators have raised concerns that imaging the retinal periphery using the RetCam-120 device (Clarity Medical Systems) was difficult, particularly in small infants with narrow palpebral fissures.\(^{13,62,80}\) Newer camera lenses may have improved this problem.\(^{24,55}\) and others have argued that imaging of mild peripheral ROP disease is less clinically important.\(^{12,55,71,74}\) The RetCam contact camera offers several lenses with different fields of view, and other non-contact devices are available with narrower fields of view (e.g., Nidek NM200-D). The trade-offs associated with these devices may warrant further research, particularly because image magnification may affect plus disease diagnosis.\(^{11}\)

A final technical concern relates to the training and expertise of both camera operators and image graders. Telemedicine studies with trained non-ophthalmologist camera operators have had high accuracy,\(^{15,53,56,71}\) comparable to accuracy in studies using ophthalmologist camera operators. However, specific training protocols emphasizing important technical considerations during image capture may be important. For example, contact cameras have potential to alter the appearance of retinal vessels. One case report cited variation in the appearance of peripheral ROP that was attributed to excessive force from the camera lens.\(^{44}\) Image graders may also benefit from specific training protocols. Compared to indirect ophthalmoscopy, RetCam images are wider-angle and lacking in stereo view. To adapt for such differences, some have suggested the possibility of training graders using real-time imaging videos.\(^{51}\)

**C. UNRESOLVED ISSUES**

Several key issues involving telemedical ROP diagnosis remain unresolved. It is not clear what constitutes “adequate” accuracy for implementation of real-world systems, particularly given that the accuracy of ophthalmoscopic examinations is not known. A reasonable diagnostic cutoff might be to refer all infants who are felt by telemedicine graders to have type 2 or worse ROP, although cost-benefit
trade-offs must be examined. Low sensitivity may result in missed cases of disease, whereas low specificity may result in unnecessary over-referrals. From this perspective, the societal impacts of missed ROP cases contributing to avoidable childhood blindness versus the inefficient use of limited ophthalmology resources must be examined. Also, a high proportion of unreadable images in a telemedicine program could pose significant challenges, because these infants would require either repeat imaging or referral for ophthalmoscopic exam. “Aggressive-posterior ROP,” which requires timely detection and treatment, may present at earlier PMA when image quality has been shown to be lower. The role of telemedicine as a tool for “examination” vs. “screening” must also be clearly defined, with protocols specifying the appropriate number of images, criteria for referral, and type of camera in various settings.

Ultimately, the feasibility of telemedicine for ROP will depend largely on resolution of logistical challenges such as integration into existing neonatology workflow, reimbursement, licensure, development of standard protocols for retinal image capture, creation of training protocols for photographers and graders, and malpractice liability. For example, retinal imaging could protect physicians through improved documentation, but could also increase medico-legal risk by subjecting images to heavy scrutiny. These issues must be addressed before creation of a large-scale telemedicine infrastructure for ROP.

V. Conclusions and Future Directions

Validation studies of telemedicine for ROP diagnosis have demonstrated that its accuracy, inter-grader reliability, and intra-grader reliability are high and are comparable to or better than that of other widely accepted diagnostic tests. This is particularly true for detection of moderate-to-severe levels of ROP and for examination of infants at later post-menstrual ages. Accuracy and reliability of telemedical imaging at earlier post-menstrual ages are less well-established in the published literature.

Based on these findings, several pilot telemedicine programs have been implemented in the United States and internationally, often using trained non-expert personnel to capture images and transfer data to remote ophthalmologists. For example, a network involving five German NICUs has managed over 1200 infants using telemedicine since 2001. Similarly, a network involving four NICUs, in which nurses are trained to capture serial wide-angle retinal images, has been providing remote care for at-risk infants at Stanford University since 2005. These programs have demonstrated that it is possible to successfully incorporate image capture, remote telemedicine interpretation, and timely referral of high-risk infants into neonatal workflow. In the longer term, expansion of these programs will require addressing many of the remaining unresolved issues. This may allow the ophthalmology expertise necessary for ROP management to be utilized more efficiently, delivered more accessibly to the growing population of premature infants, and explored further for application to other ophthalmic and medical disorders.

VI. Method of Literature Search

The Medline database was queried from January 1980 to May 2009 without date limitations. The following search terms were used: retinopathy of prematurity AND telemedicine, retinopathy of prematurity AND diagnosis, retinopathy of prematurity AND digital imaging, retinopathy of prematurity AND retinal imaging, retinopathy of prematurity AND digital imaging, retinopathy of prematurity AND incidence, retinopathy of prematurity AND cost-effectiveness AND screening. Criteria for inclusion included the relevance, clinical importance, and scientific importance of articles to the subject of this paper. Articles cited in the reference lists of other articles were reviewed and included when considered appropriate. All articles with English abstracts were reviewed, but only English-language articles were used for this paper. Additional sources included “popular”, non-peer-reviewed media regarding current telemedicine applications and ophthalmologists’ attitudes, and these are cited in the text rather than included in the references section.

References


70. Shennan AH, Bewley S. Why should preterm births be rising? BMJ. 2006;332(7547):924–5
74. Trese MT. What is the real gold standard for ROP screening? Retina. 2008;28(3):S1–2

The authors reported no proprietary or commercial interest in any product mentioned or concept discussed in the article.

M.F.C. is supported by a Career Development Award from Research to Prevent Blindness (New York, NY) and by grant EY13972 from the National Institutes of Health (Bethesda, MD). M.F.C. is an unpaid member of the Scientific Advisory Board for Clarity Medical Systems (Pleasanton, CA).

Reprint address: Michael F. Chiang, MD, Herbert Irving Assistant Professor of Ophthalmology and Biomedical Informatics, Columbia University College of Physicians and Surgeons, 635 West 165th Street, Box 92, New York, NY 10032. e-mail: chiang@dbmi.columbia.edu.
Outline

I. Retinopathy of prematurity
   A. Introduction and epidemiology
   B. Limitations of current care strategies

II. Telemedicine
   A. General principles
   B. Applications to ROP

III. Feasibility and evaluation studies
   A. Study design considerations
   B. Accuracy
      1. Accuracy of diagnosing any ROP
      2. Accuracy of diagnosing moderate to severe ROP
   C. Reliability
      1. Inter-grader reliability
      2. Intra-grader reliability
   D. Image quality
   E. Examination vs screening
   F. Cost effectiveness

IV. Challenges
   A. What is the true gold standard?
   B. Safety and technical considerations in retinal imaging
   C. Unresolved issues

V. Conclusions and future directions

VI. Method of literature search