Published in final edited form as:

Br J Ophthalmol. 2010 March ; 94(3): 351–356. doi:10.1136/bjo.2009.166348.

Telemedical Diagnosis of Retinopathy of Prematurity: Accuracy of Expert vs. Non-expert Graders

Steven L. Williams, MD¹, Lu Wang, MD¹, Steven A. Kane, MD, PhD¹, Thomas C. Lee, MD², David J. Weissgold, MD³, Audina M. Berrocal, MD⁴, Daniel Rabinowitz, PhD⁵, Justin Starren, MD, PhD^{6,7}, John T. Flynn, MD¹, and Michael F. Chiang, MD, MA^{1,7}

¹Department of Ophthalmology, Columbia University College of Physicians and Surgeons (New York, New York)

²The Vision Center, Childrens Hospital Los Angeles (Los Angeles, California)

³Retina Center of Vermont and University of Vermont (Burlington, Vermont)

⁴Department of Ophthalmology, Bascom Palmer Eye Institute (Miami, Florida)

⁵Department of Statistics, Columbia University (New York, New York)

⁶Biomedical Informatics Research Center, Marshfield Clinic (Marshfield, Wisconsin)

⁷Department of Biomedical Informatics, Columbia University College of Physicians and Surgeons (New York, New York)

Abstract

Background/Aims—To assess accuracy of telemedical retinopathy of prematurity (ROP) diagnosis by trained non-expert graders compared to expert graders.

Methods—248 eye examinations from 67 consecutive infants were captured using wide-angle retinal photography (RetCam-II, Clarity Medical Systems, Pleasanton, CA). Non-expert graders attended two hour-long training sessions on image-based ROP diagnosis. Using a web-based telemedicine system, 14 non-expert and 3 expert graders provided a diagnosis for each eye: no ROP, mild ROP, type-2 prethreshold ROP, or treatment-requiring ROP. All diagnoses were compared to a reference standard of dilated indirect ophthalmoscopy by an experienced pediatric ophthalmologist.

Results—For detection of type-2 or worse ROP, the mean (range) sensitivities and specificities were 0.95 (0.94–0.97) and 0.93 (0.91–0.96) for experts, 0.87 (0.71–0.97) and 0.73 (0.39–0.95) for resident non-experts, and 0.73 (0.41–0.88) and 0.91 (0.84–0.96) for student non-experts. For detection of treatment-requiring ROP, the mean (range) sensitivities and specificities were 1.00 (1.00–1.00) and 0.93 (0.88–0.96) for experts, 0.88 (0.50–1.00) and 0.84 (0.71–0.98) for resident non-experts, and 0.82 (0.42–1.00) and 0.92 (0.83–0.97) for student non-experts.

Conclusions—Mean sensitivity and specificity of trained non-experts are lower than that of experts, although several non-experts had high accuracy. Development of methods for training non-expert graders may help support telemedical ROP evaluation.

Corresponding author and reprints: Michael F. Chiang, MD, Columbia University College of Physicians and Surgeons, 635 West 165th Street, Box 92, New York, NY 10032, Tel: 212-342-3440 | Fax: 212-305-5962 | chiang@dbmi.columbia.edu.

COMPETING INTERESTS

MFC is an unpaid member of the Scientific Advisory Board for Clarity Medical Systems (Pleasanton, CA). The authors have no commercial, proprietary, or financial interest in any of the products or companies described in this article.

Keywords

retinopathy of prematurity; pediatric ophthalmology; telemedicine; medical informatics; retina

INTRODUCTION

Retinopathy of prematurity (ROP) is a vasoproliferative disease that affects low birth-weight infants. An international classification system has been developed,^{1,2} and multi-center studies such as the Cryotherapy for ROP (CRYO-ROP) and Early Treatment for ROP (ETROP) trials have identified criteria for disease treatment.^{3,4} Nevertheless, ROP remains a leading cause of childhood blindness throughout the world.⁵ There are major challenges in ROP care: (1) Diagnosis at the neonatal intensive care unit (NICU) bedside requires significant logistical coordination and travel. (2) There is growing need for ROP surveillance as the rates of premature birth and neonatal survival continue to rise in the United States and worldwide.⁶ (3) The availability of ophthalmologists who are willing to manage ROP is limited. A 2006 American Academy of Ophthalmology survey found that only 54% of retinal specialists and pediatric ophthalmologists are managing ROP, and that over 20% plan to stop for reasons including poor reimbursement and medico-legal liability.⁷

Store-and-forward telemedicine is a potential strategy for addressing many of these limitations. This would involve capture of clinical data and images from patients, with subsequent interpretation by a remote examiner.⁸ Telemedicine has been applied to a number of ophthalmic problems.⁹ In particular, it has been shown to be highly-accurate for ROP diagnosis, especially for detection of moderate to severe disease.^{10–13}

However, there may be significant practical difficulties associated with ROP management using telemedicine. For example, type-1 or worse ROP should be treated within 48–72 hours of diagnosis, and type-2 ROP must be monitored closely.^{3,4} Because of this rapid time course, some ophthalmologists may encounter challenges with integrating telemedical ROP management into their routine clinical workflow. Specialized ROP telemedicine reading centers staffed by trained non-physicians working under expert supervision, analogous to well-established models for diabetic retinopathy classification,^{14,15} might offer an alternative approach. The purpose of this pilot study was to measure the accuracy and reliability of telemedical diagnosis of ROP by trained non-expert graders.

METHODS

This study was approved by the Institutional Review Board at Columbia University Medical Center, and included waiver of consent for use of de-identified retinal images. Informed consent was obtained from all study participants. The study was performed in compliance with guidelines from the Declaration of Helsinki and the Health Insurance Portability and Accountability Act.

Ophthalmoscopic Examination and Retinal Imaging

During routine ROP examination at the Columbia University NICU, 1 of 2 experienced pediatric ophthalmologists (SAK, MFC) performed dilated ophthalmoscopy according to standard protocols on 67 consecutively hospitalized infants.¹⁶ Ophthalmoscopic findings were recorded for this study at up to 2 sessions for each of the 67 infants, resulting in a total of 206 unique eye examinations (103 unique bilateral examination sessions).

On the same day as ophthalmoscopic exam, all study infants also underwent wide-angle retinal imaging using a standard protocol (RetCam-II; Clarity Medical Systems, Pleasanton,

CA). Methods for image acquisition and interpretation have been described previously.¹² Data were uploaded to a secure web-based telemedicine system developed by the authors (LW, MFC). Approximately 20% of study image sets were randomly repeated by the telemedicine system for the purpose of measuring intra-grader reliability, for a total of 124 telemedical image sets (248 total eye examinations).

Graders and Training

A group of 14 non-expert study participants (1 retinal fellow, 4 ophthalmology residents, 8 medical students, 1 neonatal nurse) was invited to use the telemedicine system for image grading. All non-experts provided informed consent, and attended two hour-long training sessions on image-based ROP diagnosis conducted by the authors (JTF, MFC). Non-expert graders were given a handout, prepared by the authors, which described the definitions of each ordinal diagnostic level used in the study.² At the end of these training sessions, non-expert graders were required to review and correctly recognize a series of 10 ROP images.

A group of 3 retinal specialists (TCL, DJW, AMB) was used for comparison. Their telemedicine diagnoses have previously been shown to be highly accurate and reliable compared to a reference standard of dilated ophthalmoscopy.¹²

Telemedicine Examination

The web-based telemedicine system displayed 3–5 wide-angle retinal images from each eye, along with the birth weight, gestational age at birth, and post-menstrual age at time of examination. Data from both eyes were displayed together. Graders logged in to the system with a password, and were required to provide a single diagnosis for each eye on an ordinal scale: (1) No ROP; (2) Mild ROP; defined as ROP less than type-2 disease; (3) Type-2 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, or 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 non-contiguous clock hours of stage 3 in zone 1 or 2, with plus disease); or (5) Unknown, if the grader was uncomfortable making a diagnosis from the data provided. All results were stored in a secure database (SQL 2005; Microsoft, Redmond, WA).

Data Analysis

Sensitivity and specificity for each grader were determined for detection of ROP at 3 levels of disease severity: (1) mild or worse ROP, (2) type-2 or worse ROP, and (3) treatment-requiring ROP. Results of dilated ophthalmoscopy by the examining pediatric ophthalmologist were used as the reference standard. Unknown diagnoses were noted but excluded from initial calculations.

The 3 “best” and 3 “worst” non-expert graders were identified, based on the proportion of correctly-diagnosed eyes. To determine whether consensus of multiple graders could improve performance, accuracy was also measured using the “majority vote” diagnosis by the 3 experts, 3 best residents, 3 worst residents, 3 best medical students, and 3 worst medical students.

Finally, intra-grader reliability was determined using κ statistics for chance-adjusted agreement in ordinal diagnosis: slight agreement (0–0.20), fair agreement (0.21–0.40), moderate agreement (0.41–0.60), substantial agreement (0.61–0.80), and almost-perfect agreement (0.81–1.00).¹⁷ Analysis was performed using statistical software (Excel 2003; Microsoft, Redmond, WA; SPSS 15.0; SPSS Inc, Chicago, Illinois, and R 2.5.0; Free

Software Foundation, Boston, Massachusetts;). Statistical significance was considered to be 2-sided p -value <0.05 .

RESULTS

Distribution of Ophthalmoscopic and Telemedical Diagnoses

Reference standard ophthalmoscopic examination results for each eye have been described previously.¹² Of the 206 unique eyes, 75 (36%) were diagnosed as mild ROP, 15 (7%) were diagnosed as type-2 ROP, and 6 (3%) were diagnosed as treatment-requiring ROP by dilated ophthalmoscopy.

Figure 1 demonstrates the distribution of telemedical diagnoses by experts ($n=3$), ophthalmology residents ($n=4$), and medical students ($n=8$), when compared to reference standard ophthalmoscopy. For example, of the 12 eyes with a reference standard diagnosis of treatment-requiring ROP, 36/36 (100%) eyes received a diagnosis of treatment-requiring ROP during exams by 3 experts, 41/48 (85%) eyes received a diagnosis of treatment-requiring ROP during exams by 4 residents, and 67/96 (70%) eyes received a diagnosis of treatment-requiring ROP during exams by 8 medical students. When responses by each grader were considered individually, the 12 eyes with a reference standard diagnosis of treatment-requiring ROP were diagnosed as treatment-requiring ROP in 12/12 to 12/12 eyes by the 3 experts, 5/12 to 12/12 eyes by the 4 residents, and 6/12 to 12/12 eyes by the 8 medical students (data not displayed).

Accuracy of Telemedical Diagnoses

Figure 2 summarizes accuracy of telemedicine diagnosis by all graders, excluding “unknown” diagnoses. For detection of type-2 or worse ROP, the mean (range) sensitivities and specificities were 0.95 (0.94–0.97) and 0.93 (0.91–0.96) for experts, 0.87 (0.71–0.97) and 0.73 (0.39–0.95) for residents, and 0.73 (0.41–0.88), 0.91 (0.84–0.96) for students. For detection of treatment-requiring ROP, the mean (range) sensitivities and specificities were 1.00 (1.00–1.00) and 0.93 (0.88–0.96) for experts, 0.88 (0.50–1.00) and 0.84 (0.71–0.98) for residents, 0.82 (0.42–1.00) and 0.92 (0.83–0.97) for students.

Figure 3 summarizes accuracy of telemedicine diagnosis by consensus of the 3 experts, 3 best and 3 worst residents, and 3 best and 3 worst medical students, excluding “unknown” diagnoses. For detection of type-2 or worse ROP, the sensitivity and specificity were 1.00 and 0.94 by consensus of the 3 experts, 0.91 and 0.86 by consensus of the 3 best residents, and 0.82 and 0.96 by consensus of the 3 best students. For detection of treatment-requiring ROP, the sensitivity and specificity were 1.00 and 0.93 by consensus of the 3 experts, 1.00 and 0.87 by consensus of the 3 best residents, and 1.00 and 0.94 by consensus of the 3 best students.

Intra-Grader Reliability

Intra-grader agreement was measured for detection of treatment-requiring ROP, using the series of randomly-repeated image sets. Three of 3 (100%) experts, 2/4 (50%) residents, and 2/8 (25%) students had kappa 0.80–1.00 (near-perfect agreement). One of 4 (25%) residents and 2/8 (25%) students had kappa 0.60–0.80 (substantial agreement). One of 4 (25%) residents and 3/8 (38%) students had kappa 0.40–0.60 (moderate agreement), and 1/8 (13%) students had kappa 0.20–0.40 (fair agreement).

DISCUSSION

This study examined the accuracy of telemedical ROP diagnosis by a group of 14 trained non-expert graders compared to 3 expert graders. To our knowledge, it is the first study of this nature for ROP. Our main findings are: (1) Upon completion of the study training protocol, accuracy and intra-grader reliability of non-experts were lower than that of experts. (2) A small number of non-experts performed comparably to experts.

Telemedicine could serve either to fully classify ROP findings (“examination”) or to identify infants with disease that warrants referral for complete examination (“screening”).¹⁸ To evaluate both approaches, accuracy was measured at 3 levels of severity (Figures 2–3). The mean accuracy of non-experts was lower than that of experts at every diagnostic level. From a screening approach, a logical threshold for referral for full examination would be the presence of type-2 or worse ROP. In previous studies which measured the diagnostic performance of retinal specialists and a general ophthalmologist, compared to a reference standard of indirect ophthalmoscopy, sensitivity and specificity for detection of type-2 or worse ROP were 0.71–1.00 and 0.85–1.00, respectively.^{11,12} When compared to the lowest diagnostic accuracy among expert graders in our study, 2/14 (14%) non-experts achieved 94% sensitivity and 6/14 (43%) non-experts achieved 91% specificity for detection of type-2 or worse ROP, while 8/14 (57%) non-experts achieved 100% sensitivity and 10/14 (71%) non-experts achieved 88% specificity for detection of treatment-requiring ROP. These findings suggest that some trained non-experts have potential to perform comparably to experts for image-based ROP diagnosis.

Expert graders generally diagnosed eyes with moderate-to-severe ROP more accurately than non-experts. Among eyes with type-2 or worse ROP according to ophthalmoscopy, 5.1% (5/99) eyes were given a diagnosis of mild ROP, no ROP, or unknown by the expert graders (Figure 1). In comparison, 15.9% (21/132) of these eyes were given a diagnosis of mild ROP, no ROP, or unknown by the residents; and 32.1% (85/264) eyes were given a diagnosis of mild ROP, no ROP, or unknown by the medical students (Figure 1). This suggests that the lower sensitivity among non-experts for this clinically-significant level of disease may result from failure to identify the correct zone and stage of ROP. This has implications for development of training protocols for telemedicine as well as ophthalmoscopic examination.

This study examined the use of “consensus” diagnoses among graders as a strategy for improving accuracy (Figure 3). There was no tendency for the consensus diagnosis among non-experts to have greater accuracy than the expert graders. Alternate methods for determining the consensus diagnosis may be informative. For example, if the consensus diagnosis were defined as the most severe diagnosis among three graders, sensitivity and specificity for detection of type-2 or worse ROP would be 1.00 and 0.87 by consensus of the 3 experts, 0.91 and 0.74 by consensus of the 3 best residents, and 0.91 and 0.91 by consensus of the 3 best students; and sensitivity and specificity for detection of treatment-requiring ROP are 1.00 and 0.88 by consensus of the 3 experts, 1.00 and 0.80 by consensus of the 3 best residents, and 1.00 and 0.92 by consensus of the 3 best students. Based on these findings, the role of a consensus diagnosis for ROP remains unclear. The optimal way to improve diagnostic accuracy may inevitably require increasing the expertise of individual graders through enhanced training.

If telemedical ROP diagnosis is utilized on a large scale, eyes receiving “unknown” diagnoses must be dealt with appropriately. It may be argued that in a real-world situation, unknown diagnoses should either be promptly re-imaged or referred for ophthalmoscopy. Even if unknown diagnoses were treated as positive disease, there would still be cases of

moderate-to-severe ROP that were missed by non-expert graders in this study (Figure 1). To minimize the likelihood of missing clinically-significant disease, real-world telemedicine systems for ROP must implement protocols for timing of image acquisition and quality assurance of imaging grading, regardless of whether graders are “experts” or trained “non-experts.”

Retinal imaging has greatly impacted the management of other diseases such as diabetic retinopathy. It has been suggested that telemedical approaches to diabetic retinopathy could increase efficiency and improve patient and physician satisfaction, increase patient access and compliance, and reduce costs.^{19–21} Image-based diagnosis of diabetic retinopathy using standard photographic protocols has been shown to be more accurate than dilated examination by general ophthalmologists and retinal specialists.²² Based on this work, reading centers staffed by trained non-physicians working under expert supervision have become the “gold standard” for diabetic retinopathy diagnosis and classification in clinical trials.^{15,23} Development and validation of analogous protocols for training and certification of non-physician graders for telemedical ROP diagnosis could provide some benefits over the use of expert graders. This might improve standardization with screening and classification guidelines,^{2,4} reduce the clinical and medico-legal burden associated with ROP examinations, and provide better allocation of medical resources. However, progression in ROP is far more aggressive than in diabetic retinopathy, and there would need to be clear evidence of consistent accuracy by trained non-experts.

A key limitation of this study is that dilated ophthalmoscopy by one pediatric ophthalmologist was considered the reference standard, because this is what was done during routine clinical care. Though this has been used in many previous studies,^{10,11} image-based examination may not be inherently “less correct” than ophthalmoscopy.^{18,24} Furthermore, although ROP examiners adhere to the same guideline for diagnosis,² there is evidence that variation exists even among experts.²⁵ Therefore, it is not clear that the reference standard used for this study was necessarily always “correct.” A more rigorous approach might involve a large prospective study in which structural and functional outcomes of infants randomized to telemedicine vs. ophthalmoscopic management were compared. Because this latter approach is likely to be impractical, we feel that the design used for the current study is reasonable. We note that expert graders in this study tended to have higher sensitivities and specificities than non-experts, which supports the validity of the reference standard.

There are several other study limitations: (1) Telemedicine graders were not permitted to adjust image parameters such as brightness and contrast. This decision was made to eliminate any potential confounding effects created by differences in image manipulation among graders. (2) Non-expert graders received only limited training in telemedical ROP diagnosis, although all non-expert participants had some familiarity with ROP diagnosis and classification prior to this study. Therefore, the extent to which these study findings are generalizable, and the amount of training and certification that would be required for ROP telemedicine systems, are unclear.

Overall, this study shows that the diagnostic performance of telemedicine examination by non-expert graders is lower than that of expert retinal specialists. Although there may be logistical and workflow benefits to reading centers staffed by trained non-experts, this approach would require significant investigation and validation of training protocols.

Acknowledgments

FUNDING

This study was supported by the Starr Foundation, New York, NY (SLW), the Doris Duke Charitable Foundation, New York, NY (SLW), a Career Development Award from Research to Prevent Blindness, New York, NY (MFC), and grant EY13972 from the National Institutes of Health, Bethesda, MD (MFC).

References

1. The Committee for the Classification of Retinopathy of Prematurity. An international classification of retinopathy of prematurity. *Arch Ophthalmol*. 1984; 102:1130–4. [PubMed: 6547831]
2. An International Committee for the Classification of Retinopathy of Prematurity. The international classification of retinopathy of prematurity revisited. *Arch Ophthalmol*. 2005; 123:991–9. [PubMed: 16009843]
3. Cryotherapy for Retinopathy of Prematurity Cooperative Group. Multicenter trial of cryotherapy for retinopathy of prematurity. Preliminary results. *Arch Ophthalmol*. 1988; 106:471–9. [PubMed: 2895630]
4. Early Treatment for Retinopathy of Prematurity Cooperative Group. Revised indications for the treatment of retinopathy of prematurity: results of the early treatment for retinopathy of prematurity randomized trial. *Arch Ophthalmol*. 2003; 121:1684–94.
5. Gilbert C, Foster A. Childhood blindness in the context of VISION 2020- the right to sight. *Bull World Health Organ*. 2001; 79:227–32. [PubMed: 11285667]
6. Gilbert C, Fielder A, Gordillo L, et al. Characteristics of infants with severe retinopathy of prematurity in countries with low, moderate, and high levels of development: implications for screening programs. *Pediatrics*. 2005; 115:e518–25. [PubMed: 15805336]
7. American Academy of Ophthalmology. [Accessed June 28, 2008.] Ophthalmologists warn of shortage in specialists who treat premature babies with blinding eye condition. <http://www.aao.org/newsroom/release/20060713.cfm>
8. Perednia DA, Allen A. Telemedicine technology and clinical applications. *JAMA*. 1995; 273:483–8. [PubMed: 7837367]
9. Tang RA, Morales M, Ricur G, Schiffman JS. Telemedicine for eye care. *J Telemed and Telecare*. 2005; 11:391–6. [PubMed: 16356312]
10. Yen KG, Hess D, Burke B, et al. Telephotoscreening to detect retinopathy of prematurity: preliminary study of the optimum time to employ digital fundus camera imaging to detect ROP. *J AAPOS*. 2002; 6:64–70. [PubMed: 11997800]
11. Chiang MF, Keenan JD, Starren J, et al. Accuracy and reliability of remote retinopathy of prematurity diagnosis. *Arch Ophthalmol*. 2006; 124:322–7. [PubMed: 16534051]
12. Chiang MF, Wang L, Busuioc M, et al. Telemedical retinopathy of prematurity diagnosis: accuracy, reliability, and image quality. *Arch Ophthalmol*. 2007; 125:1531–8. [PubMed: 17998515]
13. Chiang MF, Starren J, Du YE, et al. Remote image based retinopathy of prematurity diagnosis: a receiver operating characteristic analysis of accuracy. *Br J Ophthalmol*. 2006; 90:1292–6. [PubMed: 16613919]
14. Early Treatment Diabetic Retinopathy Study Research Group. Grading diabetic retinopathy from stereoscopic color fundus photographs--an extension of the modified Airlie House classification. ETDRS report number 10. *Ophthalmology*. 1991; 98:786–806. [PubMed: 2062513]
15. Fransen SR, Leonard-Martin TC, Feuer WJ, Hildebrand PL. Clinical evaluation of patients with diabetic retinopathy: accuracy of the Inoveon diabetic retinopathy-3DT system. *Ophthalmology*. 2002; 109:595–601. [PubMed: 11874767]
16. Section on Ophthalmology American Academy of Pediatrics, American Academy of Ophthalmology, American Association for Pediatric Ophthalmology and Strabismus. Screening examination of premature infants for retinopathy of prematurity. *Pediatrics*. 2006; 117:572–6. [PubMed: 16452383]
17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977; 33:159–174. [PubMed: 843571]
18. Trese MT. What Is the Real Gold Standard for ROP Screening? *Retina*. 2008; 28:S1–2. [PubMed: 18317337]

19. Rotvold GH, Knarvik U, Johansen MA, Fossen K. Telemedicine screening for diabetic retinopathy: staff and patient satisfaction. *J Telemed and Telecare*. 2003; 9:109–13. [PubMed: 12699582]
20. Cummings DM, Morrissey S, Barondes MJ, et al. Screening for diabetic retinopathy in rural areas: the potential of telemedicine. *J Rural Health*. 2001; 17:25–31. [PubMed: 11354719]
21. Sculpher MJ, Buxton MJ, Ferguson BA, et al. A relative cost-effectiveness analysis of different methods of screening for diabetic retinopathy. *Diabet Med*. 1991; 8:644–50. [PubMed: 1833116]
22. Lee VS, Kingsley RM, Lee ET, et al. The diagnosis of diabetic retinopathy. Ophthalmoscopy versus fundus photography. *Ophthalmology*. 1993; 100:1504–12. [PubMed: 8414411]
23. Taylor CR, Merin LM, Salunga AM, et al. Improving diabetic retinopathy screening ratios using telemedicine-based digital retinal imaging technology: the Vine Hill study. *Diabet Care*. 2007; 30:574–8.
24. Scott KE, Kim DY, Wang L, et al. Telemedical Diagnosis of Retinopathy of Prematurity Intrapphysician Agreement between Ophthalmoscopic Examination and Image-Based Interpretation. *Ophthalmology*. 2008; 115:1222–8. [PubMed: 18456337]
25. Chiang MF, Jiang L, Gelman R, et al. Interexpert agreement of plus disease diagnosis in retinopathy of prematurity. *Arch Ophthalmol*. 2007; 125:875–80. [PubMed: 17620564]

Figure 1. Mean distribution of telemedical retinopathy of prematurity (ROP) diagnoses by expert retinal specialists (n=3), ophthalmology residents (n=4), and medical students (n=8) Diagnoses are divided based on reference standard of dilated ophthalmoscopy by a pediatric ophthalmologist (x-axis). Diagnoses are reported for each eye as no ROP, mild ROP, type 2 ROP, treatment-requiring ROP, or unknown.*

* In 248 study eye examinations, the mean number (range) of unknown diagnoses was 30 (0–60) among experts, 31 (11–69) among residents, and 14 (2–64) among medical students.

Figure 2. Accuracy of telemedical retinopathy of prematurity (ROP) diagnosis by 3 retinal specialists (experts), 1 retina fellow (fellow), 4 ophthalmology residents (residents), 8 medical students (students), and 1 neonatal nurse (nurse)

Sensitivity and specificity are measured for detection of (A) Mild or worse ROP, (B) Type-2 or worse ROP, and (C) Treatment-requiring ROP. All unknown diagnoses, although record, are excluded from the analysis.

Figure 3. Accuracy of telemedical retinopathy of prematurity (ROP) diagnosis by 3 individual retinal specialists (experts), consensus of 3 retinal specialists (expert consensus), consensus of 3 residents (3 best residents and 3 worst residents) and the consensus of 3 medical students (3 best students and 3 worst students)

* Sensitivity and specificity are measured for detection of (A) Mild or worse ROP, (B) Type-2 or worse ROP, and (C) Treatment-requiring ROP. The designations of “best” and “worst” are based on the proportion of images sets that were correctly diagnosed by each individual image grader. All unknown diagnoses, although recorded, are excluded from the analysis.

* Consensus diagnoses were determined by: (1) the majority vote diagnosis among three graders, or (2) the most severe diagnosis when there was no majority vote diagnosis among the three graders.