Published in final edited form as:

Ophthalmology. 2016 August ; 123(8): 1795–1801. doi:10.1016/j.ophtha.2016.04.035.

Diagnostic discrepancies in retinopathy of prematurity classification

J. Peter Campbell, MD, MPH[*,1], **Michael C. Ryan, MD**[*,1], **Emily Lore**[2], **Peng Tian, BE**[3], **Susan Ostmo, MS**[1], **Karyn Jonas, RN**[4], **R.V. Paul Chan, MD**[4], and **Michael F. Chiang, MD**[1,2] **for the Imaging & Informatics in Retinopathy of Prematurity (i-ROP) Research Consortium**

[1]Department of Ophthalmology, Casey Eye Institute, Oregon Health & Science University, Portland, OR, USA

[2]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA

[3]Cognitive Systems Laboratory, Northeastern University, Boston, MA

[4]Department of Ophthalmology and Visual Sciences, Illinois Eye and Ear Infirmary, University of Illinois at Chicago, Chicago, IL

Abstract

Objective—To identify the most common areas for discrepancy in retinopathy of prematurity (ROP) classification between experts.

Design—Prospective cohort study.

Subjects, Participants, and/or Controls—281 infants were identified as part of a multi-center, prospective, ROP cohort study from 7 participating centers. Each site had participating ophthalmologists who provided the clinical classification after routine examination using binocular indirect ophthalmoscopy (BIO), and obtained wide-angle retinal images, which were independently classified by two study experts.

Methods—Wide-angle retinal images (RetCam; Clarity Medical Systems, Pleasanton, CA) were obtained from study subjects, and two experts evaluated each image using a secure web-based module. Image-based classifications for zone, stage, plus disease, overall disease category

Address for reprints: Michael F. Chiang, MD, Departments of Ophthalmology & Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, 3375 SW Terwilliger Boulevard, Portland, OR 97239, Tel: 503-418-3087 | Fax: 503-494-5347 | chiangm@ohsu.edu.

[*]J. Peter Campbell and Michael C. Ryan contributed equally to the development of this manuscript.

Meeting Presentation:
An earlier version of this paper was presented at the American Academy of Ophthalmology Meeting in November 2014. (**AAO MEETING PAPER**)

Conflict of interest: No conflicting relationship exists for any author. MFC is an unpaid member of the Scientific Advisory Board for Clarity Medical Systems (Pleasanton, CA), and a consultant for Novartis (Basel, Switzerland).

No funding organizations had any role in the design or conduct of this research.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

(no ROP, mild ROP, Type II or pre-plus, and Type I) were compared between the two experts, and to the clinical classification obtained by BIO.

Main Outcome Measures—Inter-expert image-based agreement and image-based vs. ophthalmoscopic diagnostic agreement using absolute agreement and weighted kappa statistic.

Results—1553 study eye examinations from 281 infants were included in the study. Experts disagreed on the stage classification in 620/1553 (40%) of comparisons, plus disease classification (including pre-plus) in 287/1553 (18%), zone in 117/1553 (8%), and overall ROP category in 618/1553 (40%). However, agreement for presence vs. absence of type 1 disease was >95%. There were no differences between image-based and clinical classification except for zone III disease.

Conclusions—The most common area of discrepancy in ROP classification is stage, although inter-expert agreement for clinically-significant disease such as presence vs. absence of type 1 and type 2 disease is high. There were no differences between image-based grading and the clinical exam in the ability to detect clinically-significant disease. This study provides additional evidence that image-based classification of ROP reliably detects clinically significant levels of ROP with high accuracy compared to the clinical exam.

Retinopathy of prematurity (ROP) is a leading cause of childhood blindness worldwide.[1,2] Over the last 30 years, the multicenter Cryotherapy for Retinopathy of Prematurity (CRYO-ROP) and the Early Treatment for Retinopathy of Prematurity (ETROP) clinical trials established treatment criteria and algorithms that have been shown to improve structural and functional outcomes.[3,4] Subsequently, the American Academy of Pediatrics, American Academy of Ophthalmology, and American Association for Pediatric Ophthalmology and Strabismus have developed consensus guidelines for identification of at-risk infants who require ROP screening and surveillance.[5] Despite the development of a standardized nomenclature known as the International Classification of ROP (ICROP),[6] numerous studies have demonstrated imperfect inter-expert reliability in the diagnosis of ROP.[7–20] In the real world, this suggests that children with the same level of disease may be treated differently by different clinicians, and conversely children who are treated may not all meet clinical criteria to justify treatment. Even more troubling, ROP is one of the most costly medicolegal subjects within ophthalmology which has made it more challenging to find qualified clinicians willing to take on this area of practice.[21]

The vast majority of the previous work on inter-expert variability has focused on plus disease,[7–18,22] with relatively little work focused on zone and stage differences.[7,19,20] In addition, whether there are any differences between image-based and indirect ophthalmoscopic diagnosis (long considered the gold standard) has not been studied in detail.[23] Little research has examined the factors that contribute to which ICROP factors are most likely to be discordant between graders using different modalities, especially for zone and stage.[23] This is an important gap in knowledge that should be understood to improve real-world clinical diagnosis, as well as the diagnostic accuracy of large-scale telemedicine systems for ROP screening.

As part of a large prospective observational cohort study, the Imaging & Informatics in ROP (i-ROP) study group has developed a large repository of images obtained during routine clinical care, which are classified in terms of zone, stage, plus disease, and category both by the examining expert clinician using binocular indirect ophthalmoscopy (BIO), and separately by masked expert image graders. The purpose of this study is to evaluate the inter-expert image-based agreement on all categories of ICROP classification (zone, stage, plus, and category) as well as compare image-based ROP classification to that obtained by BIO. This study thus fills a gap in knowledge in terms of overall ICROP classification differences between experts, and may illuminate any differences between image-based classifications (as used in telemedicine ROP programs) and BIO.

METHODS

Study population

Infants were included in the study if they were admitted to a participating neonatal intensive care unit (NICU) or were transferred to a participating center for specialized ophthalmic care between July 2011 and November 2014, met published criteria for ROP screening examination, and their parents provided informed consent for data collection. Clinical data and images for this study were obtained from 7 participating institutions: (1) Oregon Health & Science University (OHSU), (2) Weill Cornell Medical College, (3) University of Miami, (4) Columbia University Medical Center, (5) Children's Hospital Los Angeles, (6) Cedars-Sinai Medical Center, and (7) Asociación para Evitar la Ceguera en México (APEC). This study was conducted in accordance with Health Insurance Portability and Accountability Act (HIPAA) guidelines, obtained approval from each institution from the Institutional Review Board (IRB), and adhered to the tenets of the Declaration of Helsinki.

Clinical Grading and Image Acquisition

Infants underwent serial dilated ophthalmoscopic examinations that were performed by a study ophthalmologist and were in accordance with current, evidence-based guidelines. Study clinicians were practicing, board certified, ophthalmologists who had undergone specialty training in either pediatric ophthalmology or vitreoretinal surgery. The clinical diagnosis was determined by BIO performed by the experienced ROP study clinician. All study clinicians were either principal investigators or certified investigators in the ETROP study, and/or had published >2 peer-reviewed articles on ROP. Exam findings were documented using ICROP criteria. A trained photographer took retinal images after the clinical exam using a commercially-available wide-angle camera (RetCam; Clarity Medical Systems, Pleasanton, CA). A typical image set for each retina included five images: posterior pole, temporal retina, nasal retina, superior retina, and inferior retina. The photographer could obtain up to 5 supplemental images if it was felt that they provided additional diagnostic information. De-identified clinical data and images were uploaded to a secure web-based database system developed by the authors. Exclusion criteria included a clinical diagnosis of stage 4 or 5, in order to focus on identification of the onset of clinically-significant disease.

Telemedical Image Reading

Two study experts (MFC, RVPC), each with more than 10 years of clinical ROP experience and over 40 ROP-related publications independently conducted remote, image-based interpretation of all of the images via an SSL-encrypted web-based module (Figure 1). In some cases, the physician readers were the same ophthalmologists who had performed the clinical examination. To minimize recall bias, images were generally reviewed several months after acquisition and no clinical data other than the retinal images and basic demographic information (birth weight, gestational age, post-menstrual age at time of examination) was available.

Images were interpreted using an ordinal scale representing overall disease category, which was based on CRYO-ROP and ETROP criteria: (1) no ROP; (2) mild ROP, defined as ROP less than type-2 disease; (3) type-2 or pre-plus ROP (zone I, stage 1 or 2, without plus disease; or zone II, stage 3, without plus disease; or any ROP less than type-1 but with pre-plus disease); (4) treatment-requiring ROP, defined as type-1 ROP (zone I, any stage, with plus disease; zone I, stage 3, without plus disease; or zone II, stage 2 or 3, with plus disease) or worse.

Data Analysis

All data were analyzed using Stata v. 11.0 (College Station, TX. Inter-expert agreement was calculated for each ordinal sub-category of zone (I–III), stage (0–5), plus (none, pre-plus, plus), and overall disease category. Agreement was similarly calculated between each expert image-based classification and the ophthalmoscopic classification. Inter-expert agreement was also calculated for clinically-significant binary classifications for zone (zone I vs. not), stage (≥stage 3 vs. not), and vascular morphology (plus disease vs. not). Chi-square tests were used to compare distributions of ordinal data, and jackknife sampling was used to generate probability sampling P values for comparing frequencies of discrepancies. $P < 0.05$ was considered statistically significant. Inter-expert agreement was reported as absolute agreement and as κ statistic for chance-adjusted agreement. Interpretation of the κ statistic utilized a commonly accepted 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.[24,25]

RESULTS

Study Population

A total of 281 infants were enrolled in this study. Infants underwent serial examinations in accordance with current ROP management guidelines (mean 3.7 examinations/infant, range 1–14), and each session included an evaluation of each eye, for a total of 1576 study eye examinations. 23 eye examinations were excluded due to a clinical diagnosis of stage 4 or 5, yielding 1553 total study eye examinations for comparison.

Table 1 summarizes the distribution of ICROP classifications for the two experts using image-based examination, and the clinical examinations from the database. The distributions between experts and between each expert and the BIO exam were significantly different for each zone, stage, plus, and category distribution (P<0.05 for all comparisons). The most notable difference was for the classification of zone, as zone III was almost never classified by image-based interpretation (<1% by the two graders vs. 14% clinically, P<0.01 for all comparisons). For each subgroup of stage, the differences between the distributions of the expert classifications was greater than the difference between either grader and the BIO exam, except for stage 3 which was classified more frequently on the BIO exam (P<0.05). For the classification of plus disease, both experts classified pre-plus disease more frequently than the clinical classification (15 and 23% vs. 7%, respectively, P<0.05 for all comparisons), however there were no differences between imaging and BIO examination for plus disease. Both experts classified eyes as pre-plus more frequently than the clinical exam (19 and 23% vs. 12%, respectively, P<0.01 for all comparisons), however there were no differences in the detection of type 1 or 2 disease.

Image-based and ophthalmoscopic discrepancy in ROP Classification

Table 2 summarizes inter-expert image-based and ophthalmoscopic discrepancies for 1553 study eye examinations for each level of zone, stage, plus disease, and overall disease category. Discrepancies were most common in the diagnosis of stage. For example, out of 1553 comparisons between expert image-based grading, there were 620 (40%) stage discrepancies, though the majority of these discrepancies (356/620, 57%) were between stage 0 and 1, with similar findings for the image-based versus ophthalmoscopic agreement. For zone, when comparing image-based vs. clinical diagnosis, there were frequent discrepancies between zone II and III classification (14% for each expert versus the ophthalmoscopic classification), in every case because the clinical diagnosis identified zone III but the image-based grade was zone II. Experts disagreed on zone I versus II in 114/1553 (7%) comparisons. For plus disease, the majority of the discrepancies were between “none” and “pre-plus.”

Inter-expert agreement for clinically-significant levels of disease

Table 3 summarizes the inter-expert agreement for zone 1, stage 3, plus disease, and type 1 ROP. Experts 1 vs. 2 demonstrated substantial agreement with absolute agreement >92% for all clinically significant classifications, and >97% for plus disease and type 1 ROP. Similarly, compared to the BIO classification, both experts demonstrated moderate to substantial agreement and >92% accuracy for all classifications, and >95% agreement for presence of plus disease and type 1 ROP. Figure 2 shows example images of eyes with discrepancies in the diagnosis of treatment-requiring ROP

DISCUSSION

This study evaluates ICROP classification agreement between experts reading images, and between image-based classification and binocular indirect ophthalmoscopy (BIO), and analyzes the main areas of discrepancy. The first key finding of this study is that the most common area of discrepancy in ROP classification is stage, and this often leads to differences in overall disease category. The second key finding is that there were no differences between image-based grading and the clinical exam regarding ability to detect clinically-significant (Type 2 or worse) disease. The third key finding is that despite frequent discrepancies at the lower levels of disease severity, overall inter-expert agreement for clinically significant disease (Type 2 or worse) is high.

The first key finding is that two ROP experts disagreed on disease stage in over 40% of eye exams. It is well established that inter-expert agreement in ROP classification is imperfect. However, most previous work has focused on plus disease[8,9,11,12,16] and zone classification. [7,19,20] In this study, the majority of discrepancies in stage classification were at the very mild ends of the disease spectrum (stage 0 vs. 1). However, the high frequency of discrepancy merits some consideration. It would be easy to understand how two clinicians might examine an infant using BIO and come to different conclusions about the presence or absence of a demarcation line (stage 1). It is not clear whether inter-expert agreement in image-based diagnosis could be improved by standardization of computer monitor settings, background room illumination, and other factors.[26] There were also discrepancies between stage 1 vs. 2 in approximately 10% of all comparisons, and between stage 2 and 3 in approximately 5% (Table 2). In theory, BIO might have an advantage in diagnosing stage 3 disease due to the advantage of stereopsis. In this study, we cannot directly address this theoretical advantage because we did not compare each modality to an external reference standard. However, we did not observe any difference in the overall diagnosis of stage between image grading or clinical exam, and the rate of disagreements was similar between two image-based classifications, and between image-based and BIO classification.

Previous work has focused on improving ROP classification accuracy and precision through education.[27–31] However, when experts are disagreeing >40% of the time, this may suggest that additional information is required to improve agreement beyond what is apparent on a 2-dimensional color image and a BIO fundus exam – or that the current classification system is imprecise. For example, the incorporation of fluorescein angiography may be able to improve the level of agreement.[19,20] In the future, as portable and widefield optical coherence tomography becomes more commonplace, this may be expected to improve diagnostic agreement as well.[32] Both of these modalities would add time and expense to the routine evaluation of infants, and fluorescein angiography adds a small risk of allergic reaction. However, if they could improve agreement between experts (and therefore improve the standardization of care for these infants), they may have a more mainstream future role in ROP diagnosis.

The second key finding is that there were no differences between image-based vs. ophthalmsocopic diagnosis of clinically-significant (Type 2 or worse) disease. In general, the inter-expert discrepancy rate exceeded any differences between image-based and BIO diagnosis, except for classification of zone III (Table 1). Both study experts classified zone III less frequently based on image interpretation than the corresponding clinical diagnosis, which is easy to understand as due to the difficulty in confirming the presence of zone III vasculature on a digital RetCam image. While this has implications for how long a particular infant may need to be followed with serial imaging in a telemedical ROP program, it would not necessarily affect the ability of an image based system to detect clinically significant disease. The other notable difference between the image-based expert graders and the BIO classification was that the 2 image-based graders tended to classify pre-plus disease more frequently than the 8 examining clinicians as part of routine care using BIO. It is unclear whether this difference reflects a true bias between image-based and ophthalmoscopic classification or whether it is due to inter-observer variability on the tendency to or threshold to diagnose pre-plus disease in routine clinical practice, as there was no difference in the diagnosis of plus disease.[8]

The third key study finding is that overall agreement on type 1 ROP classification was 95–97% between all modalities (Table 3). This is in contrast to the original CRYO-ROP study in which non-masked experts openly disagreed in 12% of cases with clinical colleagues on the need for treatment.[33] We have previously shown that ROP experts often have difficulty describing the precise factors that weigh into their determination of plus disease, and that different experts focus on different characteristics.[11] Our group and others are exploring these image-based features to determine if they can be quantified and if by doing so we may be able to develop a quantifiable, and objective, measurement of ROP disease severity that eventually complements or replaces the current ICROP zone, stage, plus categorization.[34]

There are several limitations to this study: (1) Analysis of inter-expert agreement for image-based classifications was limited to two ROP expert image readers (MFC, RVPC). Therefore, it is not clear that study findings may be generalized to inter-expert agreement during BIO. This may limit generalizability of our study findings to a broader range of ophthalmologists. However, the two study readers (MFC, RVPC) are close collaborators who have reviewed images and clinical scenarios together for almost 10 years. For this reason, we believe that diagnostic discrepancies with other experts may be larger than what was found in this study. (2) The two expert image readers (MFC, RVPC) were also participating clinicians in recruiting this cohort of patients as part of a larger observational cohort study. As a result, in some cases the image-based interpretation and the clinical diagnoses were performed by the same person, which may be affected by recall bias and/or other biases. Overall, the number of affected observations is small, the images were presented without identifying information, and there were at least several months between clinical diagnosis and image grading, so we feel these effects should be minimal. (3) Inter-expert reliability in general is a limitation of any attempt to compare image-based grading to clinical diagnosis made by a single observer, since that observer may have more disagreement compared to the average expert than any slight systemic bias between clinical and image-based grading. (4) For this analysis, due to small numbers for comparison, we excluded 23 eyes with a clinical diagnosis of stage 4 or 5. Importantly, all of these eyes had agreement on disease category (data not shown), which would have screened “positive” in a telemedical setting. (5) Since we examined infants multiple times, and included both eyes separately, the 1553 eye examinations were not completely independent. However, our graders were shown these images without reference to prior examinations and not at the same time, so that the grades were obtained as independently as possible. We felt that the benefit of including as many eyes with various disease stages as possible outweighed the statistical disadvantage of not having fully independent samples.

In summary, this study provides additional evidence that image-based classification of ROP reliably detects clinically significant levels of ROP with high accuracy compared to the clinical exam. This suggests that efforts to incorporate telemedicine approaches using image-based grading, or computer based image analysis programs are sufficiently sensitive for clinical application to detect type 2 (referral-warranted) or type 1 (treatment-requiring) disease without the need for direct screening of every infant with BIO. Though there was frequent disagreement on stage classification, the majority of these disagreements would be clinically insignificant as they would not affect management or outcome. Novel approaches to disease classification, such as with fluorescein angiography or OCT, may warrant future exploration to improve agreement and our understanding of stage. Computer-based image analysis of these wide-field images in ROP may yield meaningful clinical data such as a scaled ROP severity score that may complement or replace our current classification system, and improve our ability to reliably (and automatically) classify and diagnose referral-warranted and treatment-requiring ROP in the future.[34,35]

Acknowledgments

Financial Support: Supported by grants R01 EY19474 and P30 EY010572 from the National Institutes of Health, Bethesda, MD (MFC), unrestricted departmental funding from Research to Prevent Blindness, New York, NY (JPC, MCR, SO, KJ, RVPC, MFC), the St. Giles Foundation (RVPC), and the iNsight Foundation (RVPC, KEJ).

References

1. Gilbert C, Foster A. Childhood blindness in the context of VISION 2020--the right to sight. Bull World Health Organ. 2001; 79:227–232. [PubMed: 11285667]
2. Sommer A, Taylor HR, Ravilla TD, et al. Challenges of ophthalmic care in the developing world. JAMA Ophthalmol. 2014; 132:640–644. [PubMed: 24604415]
3. Cryotherapy for Retinopathy of Prematurity Cooperative Group. Multicenter trial of cryotherapy for retinopathy of prematurity. Preliminary results. Arch Ophthalmol. 1988; 106:471–479. [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; 121:1684–1694. [PubMed: 14662586]
5. Fierson WM. American Academy of Pediatrics Section on Ophthalmology, American Academy of Ophthalmology, et al. Screening examination of premature infants for retinopathy of prematurity. Pediatrics. 2013; 131:189–195. [PubMed: 23277315]
6. International Committee for the Classification of Retinopathy of Prematurity. The International Classification of Retinopathy of Prematurity revisited. Arch Ophthlamol. 2005; 123:991–999.
7. Chiang MF, Thyparampil PJ, Rabinowitz D. Interexpert agreement in the identification of macular location in infants at risk for retinopathy of prematurity. Arch Ophthalmol. 2010; 128:1153–1159. [PubMed: 20837799]
8. Chiang MF, Jiang L, Gelman R, et al. Interexpert agreement of plus disease diagnosis in retinopathy of prematurity. Arch Ophthalmol. 2007; 125:875–880. [PubMed: 17620564]
9. Rao R, Jonsson NJ, Ventura C, et al. Plus disease in retinopathy of prematurity: diagnostic impact of field of view. Retina. 2012; 32:1148–1155. [PubMed: 22466473]
10. Gelman SK, Gelman R, Callahan AB, et al. Plus disease in retinopathy of prematurity: quantitative analysis of standard published photograph. Arch Ophthalmol. 2010; 128:1217–1220. [PubMed: 20837812]
11. Chiang MF, Gelman R, Williams SL, et al. Plus disease in retinopathy of prematurity: development of composite images by quantification of expert opinion. Invest Ophthalmol Vis Sci. 2008; 49:4064–4070. [PubMed: 18408188]
12. Keck KM, Kalpathy-Cramer J, Ataer-Cansizoglu E, et al. Plus disease diagnosis in retinopathy of prematurity: vascular tortuosity as a function of distance from optic disk. Retina. 2013; 33:1700–1707. [PubMed: 23538582]
13. Hewing NJ, Kaufman DR, Chan RVP, Chiang MF. Plus Disease in Retinopathy of Prematurity: Qualitative Analysis of Diagnostic Process by Experts. JAMA Ophthalmol. 2013; 131:1026–1032. [PubMed: 23702696]

14. Gelman R, Jiang L, Du YE, et al. Plus disease in retinopathy of prematurity: Pilot study of computer-based and expert diagnosis. Journal of American Association for Pediatric Ophthalmology and Strabismus. 2007; 11:532–540. [PubMed: 18029210]

15. Ataer-Cansizoglu E, Kalpathy-Cramer J, You S, et al. Analysis of Underlying Causes of Inter-expert Disagreement in Retinopathy of Prematurity Diagnosis. Application of Machine Learning Principles. Methods Inf Med. 2015; 54:93–102. [PubMed: 25434784]

16. Chiang MF, Gelman R, Jiang L, et al. Plus disease in retinopathy of prematurity: an analysis of diagnostic performance. Trans Am Ophthalmol Soc. 2007; 105:73–84. discussion 84–5. [PubMed: 18427596]

17. Slidsborg C, Forman JL, Fielder AR, et al. Experts do not agree when to treat retinopathy of prematurity based on plus disease. Br J Ophthalmol. 2012; 96:549–553. [PubMed: 22174097]

18. Bolon-Canedo V, Ataer-Cansizoglu E, Erdogmus D, et al. Dealing with inter-expert variability in retinopathy of prematurity: A machine learning approach. Comput Methods Programs Biomed. 2015; 122:1–15. [PubMed: 26120072]

19. Patel SN, Klufas MA, Ryan MC, et al. Color Fundus Photography Versus Fluorescein Angiography in Identification of the Macular Center and Zone in Retinopathy of Prematurity. Am J Ophthalmol. In press.

20. Klufas MA, Patel SN, Ryan MC, et al. Influence of Fluorescein Angiography on the Diagnosis and Management of Retinopathy of Prematurity. Ophthalmology. 2015; 122:1601–1608. [PubMed: 26028345]

21. Chiang MF, Gelman R, Martinez-Perez ME, et al. Image analysis for retinopathy of prematurity diagnosis. Journal of American Association for Pediatric Ophthalmology and Strabismus. 2009; 13:438– 445. [PubMed: 19840720]

22. Wallace DK, Quinn GE, Freedman SF, Chiang MF. Agreement among pediatric ophthalmologists in diagnosing plus and pre-plus disease in retinopathy of prematurity. J AAPOS. 2008; 12:352–356. [PubMed: 18329925]

23. Scott KE, Kim DY, Wang L, et al. Telemedical diagnosis of retinopathy of prematurity intraphysician agreement between ophthalmoscopic examination and image-based interpretation. Ophthalmology. 2008; 115:1222–1228. e3. [PubMed: 18456337]

24. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005; 37:360–363. [PubMed: 15883903]

25. Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. J Biomed Inform. 2002; 35:99–110. [PubMed: 12474424]

26. Wang J, Langer S. A brief review of human perception factors in digital displays for picture archiving and communications systems. J Digit Imaging. 1997; 10:158–168. [PubMed: 9399169]

27. Myung JS, Paul Chan RV, Espiritu MJ, et al. Accuracy of retinopathy of prematurity image-based diagnosis by pediatric ophthalmology fellows: implications for training. J AAPOS. 2011; 15:573–578. [PubMed: 22153403]

28. Paul Chan RV, Williams SL, Yonekawa Y, et al. Accuracy of retinopathy of prematurity diagnosis by retinal fellows. Retina. 2010; 30:958–965. [PubMed: 20168274]

29. Wong RK, Ventura CV, Espiritu MJ, et al. Training fellows for retinopathy of prematurity care: a Web-based survey. J AAPOS. 2012; 16:177–181. [PubMed: 22525176]

30. Paul Chan RV, Patel SN, Ryan MC, et al. The Global Education Network for Retinopathy of Prematurity (Gen-Rop): Development, Implementation, and Evaluation of A Novel Tele-Education System (An American Ophthalmological Society Thesis). Trans Am Ophthalmol Soc. 2015; 113:T21–T226. [PubMed: 26538772]

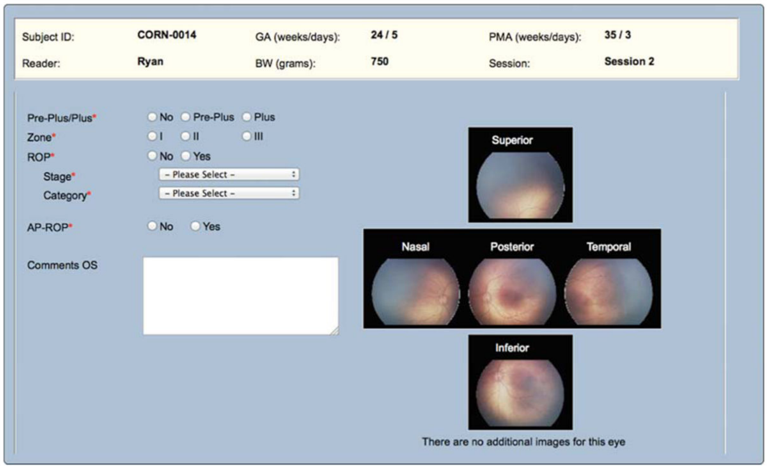
31. Campbell JP, Swan R, Ostmo S, et al. Implementation and evaluation of a tele-education system for the diagnosis of ophthalmic disease by international trainees. AMIA Annu Symp Proc. In press.

32. Maldonado RS, Toth CA. Optical coherence tomography in retinopathy of prematurity: looking beyond the vessels. Clin Perinatol. 2013; 40:271–296. [PubMed: 23719310]

33. Reynolds JD, Dobson V, Quinn GE, et al. Evidence-based screening criteria for retinopathy of prematurity: natural history data from the CRYO-ROP and LIGHT-ROP studies. Arch Ophthalmol. 2002; 120:1470–1476. [PubMed: 12427059]

34. Ataer-Cansizoglu E, Bolon-Canedo V, Campbell JP, et al. Computer-Based Image Analysis for Plus Disease Diagnosis in Retinopathy of Prematurity: Performance of the “i-ROP” System and Image Features Associated With Expert Diagnosis. Transl Vis Sci Technol. In press.

35. Campbell JP, Ataer-Cansizoglu E, Bolon-Canedo V, et al. Expert diagnosis of plus disease in retinopathy of prematurity from computer-based image analysis. JAMA Ophthalmol. In Press.



Subject ID:	CORN-0014	GA (weeks/days):	24 / 5	PMA (weeks/days):	35 / 3
Reader:	Ryan	BW (grams):	750	Session:	Session 2

Pre-Plus/Plus* ○ No ○ Pre-Plus ○ Plus

Zone* ○ I ○ II ○ III

ROP* ○ No ○ Yes

Stage* – Please Select –

Category* – Please Select –

AP-ROP* ○ No ○ Yes

Comments OS

There are no additional images for this eye

Figure 1.

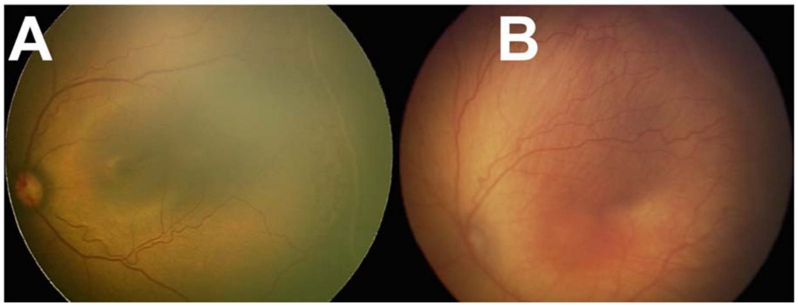


Figure 2.

Table 1

Distribution of retinopathy of prematurity (ROP) classification for image-based and ophthalmoscopic diagnosis by experts.

ROP classification	Image-based (expert 1)	Image-based (expert 2)	Ophthalmoscopic (examining clinician)
	No (%) of classifications	No (%) of classifications	No (%) of classifications
Zone			
I	170 (11)	146 (9)	132 (9)
II	1382 (89)	1403 (90)	1204 (78)
III	1 (<1)	4 (<1)	217 (14)
Stage			
0	533 (34)	891 (57)	713 (46)
1	551 (35)	317 (20)	387 (25)
2	368 (24)	224 (14)	321 (21)
3	98 (6)	120 (8)	132 (9)
Plus			
None	1259 (81)	1161 (75)	1390 (90)
Pre-plus	229 (15)	351 (23)	104 (7)
Plus	65 (4)	41 (3)	59 (4)
Category			
None	531 (34)	849 (55)	714 (46)
Mild	642 (41)	297 (19)	575 (37)
Type II and/or Pre-plus	296 (19)	350 (23)	179 (12)
Type I	84 (5)	57 (4)	85 (5)

*P < 0.05 for all comparisons of distributions between experts and the ophthalmoscopic exam for zone, stage, plus, and category.

Table 2

Discrepancies in retinopathy of prematurity (ROP) classification

Area of Discrepancy	Image-based (expert 1 versus expert 2) Number (%) of discrepancies	Image-based (expert 1) vs. BIO Number (%) of discrepancies	Image-based (expert 2) vs. BIO Number (%) of discrepancies
Zone	117 (8)	322 (21)*	329 (21)*
I vs. II	114 (7)	106 (7)	112 (7)
II vs. III	3 (0.2)	216 (14)*	217 (14)*
Stage	620 (40)	614 (40)	512 (33)*
0 vs. 1	356 (23)	332 (21)	237 (15)*
1 vs. 2	148 (10)	162 (10)	139 (9)
2 vs. 3	60 (4)	64 (4)	61 (4)
Plus	287 (18)	200 (13)*	342 (22)*
None vs. Pre-Plus	255 (16)	146 (10)*	300 (19)*
Pre-Plus vs. Plus	27 (2)	37 (2)*	27 (2)
Category	618 (40)	582 (37)*	529 (34)*
None vs. Mild	320 (21)	317 (20)	212 (14)*
Mild vs. Type II	186 (12)	144 (9)*	200 (13)*
Type II vs. Treatment-requiring	35 (2)	47 (3)*	42 (3)*

N = 1553 study eye examinations. BIO = binocular indirect ophthalmoscopy.

*P<0.05 compared to the image-based inter-expert discrepancy frequency (column 1).

Table 3

Agreement in retinopathy of prematurity (ROP) for clinically significant disease level.

Clinically significant thresholds	Image-based classification (expert 1 versus expert 2)		Image-based classification (expert 1) vs. BIO		Image-based classification (expert 2) vs. BIO	
	Weighted Kappa (95% CI)	Agreement (95% CI)	Weighted Kappa (95% CI)	Agreement (95% CI)	Weighted Kappa (95% CI)	Agreement (95% CI)
Zone I disease	0.6 (0.5–0.7)	93 (91–94)	0.6 (0.5–0.7)	93 (92–94)	0.6 (0.5–0.6)	93 (92–94)
≥ Stage 3	0.6 (0.6–0.7)	95 (94–96)	0.6 (0.5–0.7)	95 (94–96)	0.6 (0.5–0.7)	94* (93–95)
Plus disease	0.7 (0.6–0.8)	98 (97–99)	0.5 (0.4–0.7)	97* (96–97)	0.6 (0.4–0.7)	97 (97–98)
Treatment-Requiring (Type I ROP)	0.7 (0.6–0.8)	97 (97–98)	0.6 (0.5–0.7)	96* (95–97)	0.6 (0.5–0.7)	96 (95–97)

N = 1553 study eye examinations. BIO = binocular indirect ophthalmoscopy.

*P<0.05 compared to the image-based inter-expert discrepancy frequency (column 1).