Published in final edited form as:
Ophthalmology. 2016 November ; 123(11): 2338–2344. doi:10.1016/j.ophtha.2016.07.026.

Plus disease in retinopathy of prematurity: a continuous spectrum of vascular abnormality as basis of diagnostic variability

J. Peter Campbell, MD, MPH[1,*], **Jayashree Kalpathy-Cramer, PhD**[2,*], **Deniz Erdogmus, PhD**[3], **Peng Tian, BE**[3], **Dharanish Kedarisetti, MSc**[3], **Chace Moleta, MS**[1], **James D. Reynolds, MD**[4], **Kelly Hutcheson, MD**[5], **Michael J. Shapiro, MD**[6], **Michael X. Repka, MD, MBA**[7], **Philip Ferrone, MD**[8], **Kimberly Drenser, MD**[9], **Jason Horowitz, MD**[10], **Kemal Sonmez**[11], **Ryan Swan**[11], **Susan Ostmo, MPH**[1], **Karyn E. Jonas, RN**[12], **R.V. Paul Chan, MD**[12], and **Michael F. Chiang, MD**[1,11] **on behalf of the i-ROP research consortium**

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

[2]Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Charlestown, MA, USA

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

[4]Department of Ophthalmology, Ross Eye Institute, State University of New York at Buffalo, Buffalo, NY, USA

[5]Department of Ophthalmology, Sidra Medical & Research Center, Doha, Qatar

[6]Retina Consultants, Chicago, IL

[7]Wilmer Institute, Johns Hopkins University School of Medicine, Baltimore, MD

[8]Long Island Vitreoretinal Consultants, Great Neck, NY

[9]Associated Retinal Consultants, Oakland University, Royal Oak, MI

[10]Columbia University, New York, NY

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

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

Abstract

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.

[*]Drs. Campbell and Kalpathy-Cramer contributed equally to the development of this manuscript.

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.

Objective—To identify patterns of inter-expert discrepancy in plus disease diagnosis in retinopathy of prematurity (ROP).

Design—We developed two datasets of clinical images of varying disease severity (100 images and 34 images) as part of the Imaging and Informatics in ROP study, and determined a consensus reference standard diagnosis (RSD) for each image, based on 3 independent image graders and the clinical exam. We recruited 8 expert ROP clinicians to classify these images and compared the distribution of classifications between experts and the RSD.

Subjects, Participants, and/or Controls—Images obtained during routine ROP screening in neonatal intensive care units. 8 participating experts with >10 years of clinical ROP experience and >5 peer-reviewed ROP publications.

Methods, Intervention, or Testing—Expert classification of images of plus disease in ROP.

Main Outcome Measures—Inter-expert agreement (weighted kappa statistic), and agreement and bias on ordinal classification between experts (ANOVA) and the RSD (percent agreement).

Results—There was variable inter-expert agreement on diagnostic classifications between the 8 experts and the RSD (weighted kappa 0 – 0.75, mean 0.30). RSD agreement ranged from 80 – 94% agreement for the dataset of 100 images, and 29 – 79% for the dataset of 34 images. However, when images were ranked in order of disease severity (by average expert classification), the pattern of expert classification revealed a consistent systematic bias for each expert consistent with unique cut points for the diagnosis of plus disease and pre-plus disease. The two-way ANOVA model suggested a highly significant effect of both image and user on the average score ($P<0.05$, adjusted $R^2=0.82$ for dataset A, and $P< 0.05$ and adjusted $R^2 =0.6615$ for dataset B).

Conclusions and Relevance—There is wide variability in the classification of plus disease by ROP experts, which occurs because experts have different “cut-points” for the amounts of vascular abnormality required for presence of plus and pre-plus disease. This has important implications for research, teaching and patient care for ROP, and suggests that a continuous ROP plus disease severity score may more accurately reflect the behavior of expert ROP clinicians, and may better standardize classification in the future.

INTRODUCTION

Retinopathy of prematurity (ROP) is a leading cause of childhood blindness in the United States and throughout the world.[1] The Cryotherapy for ROP (CRYO-ROP) and Early Treatment for ROP (ETROP) studies demonstrated that appropriate identification and timely treatment of ROP can reduce the risk of adverse outcomes and vision loss.[2–4] Those investigations also showed that, among all parameters in the International Classification of ROP (ICROP), presence of plus disease is the most critical feature for identifying infants with severe disease who require treatment to prevent blindness. Thus, accurate and consistent identification of plus disease is essential for ROP management.

Plus disease is defined by a standard published photograph, which was selected by expert consensus during the 1980’s.[3,5] This standard photograph represents the minimum amount of arterial tortuosity and venous dilation in the central retina that is required for the presence of plus disease. In 2005, an intermediate classification called “pre-plus” disease was introduced and was defined as more venous dilation and arterial tortuosity than normal, but less than the standard photograph.[2] Importantly, it is well established that there is significant inter-expert variability in diagnosis.[6–9] Several possible explanations exist for this inter-expert variability. Available evidence suggests that some clinicians may focus on wider fields of view than shown in the standard photograph,[10–12] that clinicians may focus on different vascular features (such as venous tortuosity)[13,14] which were not included in the standard definition of plus disease, that experts may be unable to readily identify which vascular features are most important to them in diagnosis,[15] and that experts may have different "cut points" for vascular abnormality required for diagnosis of plus disease. Overall, the underlying basis for variability in plus disease diagnosis is not well understood. Addressing this gap in knowledge will allow creation of better diagnostic and educational methods, which will lead to improved ROP management and prevention of visual impairment in children.

The current method of clinical plus disease diagnosis is based on delineating "cut points" in vascular abnormality between either a two-level (plus or normal) or three-level (plus, pre-plus, or normal) classification system.[2,5] Since 2010, our group has conducted the Imaging & Informatics in ROP (i-ROP) study, which includes a goal of developing and validating computer-based methods for quantitative ROP diagnosis. Through this work, we have collected multiple expert diagnoses from large numbers of ROP examinations and images. The purpose of this paper is to analyze these data to determine the underlying basis of inter-expert variability in ROP diagnosis. We show that there is a continuous spectrum of vascular abnormality in ROP from very normal to very abnormal vessels, show that ROP experts have different cut points between categories on this continuous spectrum of disease, and propose consideration of a more continuous quantitative severity scale for vascular abnormalities in ROP.

METHODS

This study was approved by the Institutional Review Board at Oregon Health & Science University, and followed the tenets of the Declaration of Helsinki. Written informed consent was obtained from parents of all infants in the "Imaging and Informatics in Retinopathy of Prematurity" (i-ROP) study

Description of datasets

We developed 2 data sets of wide-angle retinal images acquired during routine clinical care. For each image, we established a reference standard diagnosis (RSD: plus, pre-plus, or normal), using previously published methods that combines the classifications of three expert ROP image graders (independent, masked classifications from two ophthalmologists and one non-physician ROP study coordinator) and the actual clinical diagnosis.[16] The first dataset (A) was designed to represent the full range of disease severity and included 100 images, of which 15 had a RSD of plus disease, 31 had pre-plus disease, and 54 were normal. The second dataset (B) was designed to represent infants with more clinically-significant disease and included 34 images, of which 20 had an RSD of plus disease, 13 had pre-plus disease, and 1 was normal.

Expert classification

Each of the images was reviewed by 8 ROP experts and classified as plus, pre-plus, or normal, which was coded as “3,” “2,” or “1,” respectively in the database. Participating experts were all practicing clinicians with a minimum of 10 years experience in ROP screening. Five experts served as Principal Investigators at ETROP study centers, and one served as a certified ETROP Investigator. All 8 experts had published a minimum of 5 peer-reviewed ROP journal publications.

Data analysis

Statistical analysis was performed using Excel 2016 (Microsoft, Redmond, WA), Stata v. 11.0 (College Station, TX), and R v3.2.2.[17] The average score for each image was calculated, and the images were ranked from most severe to least severe based on these scores. The average score of all the images was calculated for each expert, and experts were ranked from highest average score to lowest. The “bias” of each expert was calculated using the average difference between each expert’s classification and the RSD, and inter-expert agreement was calculated using a weighted kappa function, and interpreted using 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.[18,19] A two-way ANOVA model (image, expert) without replication was used to model the effect of the user and the image without interactions. Pearson’s correlation coefficient was calculated to determine relative ordering of experts between datasets.

RESULTS

Table 1 displays the distribution of plus disease classification (plus, pre-plus, or normal) for all 8 experts (labeled 1–8) and the RSD, ranked from least severe average grade (top) to most severe average grade (bottom) for dataset A (100 images). The 8 experts are displayed in the same order for dataset B (34 images). The average percent RSD agreement was higher in dataset A (82%, range 77–94%) than dataset B (65% range 29–91%) due to the large number of normal images with good agreement, demonstrating the effect of the population studied on the percent agreement as an outcome.

There was a systematic tendency to relatively over-call or under-call for each expert, which was consistent between datasets A and B (Pearson’s correlation coefficient 0.90 for correlation of average image score for each expert between datasets). For example, Expert #1 diagnosed plus disease in 6/100 (6%) images in dataset A, whereas Expert #7 diagnosed plus disease in 29/100 (29%) images in the same dataset. Similarly, Expert #1 diagnosed plus disease in 6/34 (18%) images in dataset B, whereas Expert #7 diagnosed plus disease in 30/34 (88%) images in the same dataset. Weighted kappa statistics were calculated between each pair of experts. For dataset A the mean weighted kappa was 0.67 (range 0.48 – 0.88), whereas for dataset B mean weighted kappa was 0.30 (range 0.00 – 0.75). The two-way ANOVA model suggested a highly significant effect of both image and user on the average score ($P<0.05$, adjusted $R^2=0.82$ for dataset A, and $P< 0.05$ and adjusted $R^2 =0.6615$ for dataset B).

Figure 1 displays the range of diagnoses for individual images, ordered by the average classification from most severe (left) to least severe (right) for both datasets. In the dataset of 100, 3/100 images (3%) were classified as plus disease by all 8 experts (dark blue), and 33/100 (33%) images were classified as normal by all 8 experts (light blue, 32 were excluded from right side of the Figure 1A for space limitations). In 64/100 (64%) images, there was disagreement among experts as to the disease classification. In the set of 34, 2/34 (6%) were classified as plus by all experts, and in 32/34 (94%) there was disagreement. The distribution of the classifications of 8 experts reveals a wide transition zone between majority vote agreement for both the classification of pre-plus and plus disease for both datasets, suggesting that experts have systematic differences in “cut points” for the border between pre-plus vs. plus and between normal vs. pre-plus. As shown in Figure 1, the RSD cut points between transitions are in the middle of the range of the 8 experts, and differences in percent RSD “agreement” from Table 1 reflect systematic differences in cut-points (bias) rather than random error, with good overall agreement on the relative severity of images.

Figure 2 displays a representative range of images within each category of disease (plus, pre-plus, normal), and the range of expert diagnostic classifications for each image. This graphically depicts the continuous spectrum of severity of vascular abnormality severity within each discrete ICROP diagnostic category (plus, pre-plus, or normal), from most severe (left) to least (right). In addition to demonstrating the spectrum of vascular abnormality within each ordinal classification, this shows that different experts appear to have different cut-offs for the transitions between diagnostic classifications.

DISCUSSION

This study analyzes the classification of plus disease by ROP experts, with the goal of examining the pattern of diagnostic discrepancies. Key findings from this study are: (1) Even among ROP experts, there is limited agreement on diagnostic classification of plus disease, (2) diagnostic discrepancy in plus disease reflects consistent systematic biases for each expert as to the appropriate cut points for plus and pre-plus disease, (3) a continuous severity score, instead of discrete classifications of plus, pre-plus, and normal, may more accurately model the real world behavior of experts and the range of vascular abnormalities in ROP than the current discrete ICROP classification (plus disease, pre-plus disease, or normal).

The first key finding is that that even among experts in ROP there is poor agreement on what constitutes plus disease (Table 1 and Figure 1). This suggests that there may be variation in treatment recommendations for the same infant between different examiners, despite an international standard for the ROP classification. Thus, despite the evidence from CRYO-ROP and ETROP on evidence-based thresholds for intervention, there may be infants who are undertreated and at higher than necessary risk of blindness. Similarly, other infants may be over-treated and subjected to unnecessary treatment with associated morbidity of laser and/or anti-vascular endothelial growth factor treatment. This is consistent with previously-published studies that have shown diagnostic inconsistency among experts,[6,7,15,20–23] and reinforces the importance of developing methods to improve the accuracy and consistency of plus disease diagnosis. We feel that computer-based image analysis methods[14,24–26] and tele-education[27,28] are two promising approaches to address this issue.

The second key finding is that these diagnostic discrepancies among experts can be largely explained by systematic biases, i.e. some experts tend to "over-call" compared to the overall group, whereas others tend to systemically "under-call" (Table 1 and Figure 1), compared to the RSD. The results of ANOVA modeling suggest that the response of a given expert is almost completely predictable given the response of another expert, because of these systematic differences. The clinical significance of this variability is unknown, as all of the pivotal clinical trials[3,4] would have the same inherent limitation: inter-expert variation in the diagnostic classification of plus disease. Therefore, it is unclear whether the less aggressive experts are "under-treating," or whether the more aggressive experts are "over-treating." This would require some understanding of the prognostic significance of these various thresholds, which is beyond the scope of this study.

In this study, the use of two datasets to explore these trends strengthens the findings in several ways. First, we were able to demonstrate consistent diagnostic trends supporting the concept of systematic bias in two different populations, with similar relative ordering between experts (Figure 1 and Table 1). Second, as computer-based image analysis is developed as a diagnostic tool, it will require validation across populations with different underlying disease prevalence, as the predictive value of any diagnostic test is dependent on the underlying population. Third, in the accompanying paper we use both datasets to demonstrate the utility of pairwise comparison testing and computer-based image analysis to produce a relative ordering of disease severity that works in both datasets over the full range of disease, as well as over the more clinically relevant severe disease end of the spectrum.[26]

Virtually all published clinical, telemedicine, and computer-based image analysis studies in ROP rely on findings from a single clinical expert as a "gold standard".[7,25,29–34] This raises important questions about the external validity of those results because of the significant variations in cut points for plus disease and pre-plus disease (e.g. Figure 2) between experts. To address this problem, our group has developed a methodology for determining consensus reference standard diagnoses in ROP that utilizes the classifications of three expert image graders (two physicians and one study coordinator), together with the clinical ophthalmoscopic exam diagnosis.[16] At the very least, this process assures a "regression to the mean," as the average score from a randomly selected group of these 8 experts would be very close to our RSD classification, as demonstrated in a recent publication comparing expert diagnosis with computer-based image analysis (computer-based image analysis) diagnosis.[14] We advocate for the incorporation of a similar consensus RSD for future ROP research.

Our third key finding is that diagnostic behavior of experts in this study suggests that vascular abnormality in plus disease runs on a continuum (Figure 1 and Figure 2). For that reason, we believe that development of a more continuous quantitative severity score for ROP may better model the nature of vascular abnormality, more closely reflect the behavior of experts in the real world, and improve standardization of clinical plus disease diagnosis in the future. Such a scale would create improved opportunities for identifying clinical change in vascular abnormality beyond the current 3-level discrete scale (plus, pre-plus, normal)[35,36] which could have profound implications for clinical care, prospective outcomes research, and telemedicine validation.[26] This also has implications for the design of clinical trials in ROP that may require determination of whether patients are “improving” or “worsening”. For example, images such as those in Figure 2 may be used in multi-center ROP trials to standardize the clinical diagnoses of study investigators, and measure clinical outcomes, beyond the granularity available in a 3-level scale (plus, pre-plus, or normal).[37]

Computer-based image analysis systems may be applied to sort images into disease severity, and may be validated by constructing receiver operating characteristic (ROC) curves against a reference standard diagnosis. Two such systems, ROPTool and the i-ROP system, have been used to effectively generate a continuous “plus” scale to complement the clinical exam.[24,38] Based on findings from this current study, we believe that future computer-based image analysis systems may be used to generate quantitative, reproducible, continuous scales representing vascular abnormality, rather than attempting to classify disease into ordinal categories (e.g. using ROC curves against one expert’s standard with cut points that may not reflect the behavior of other experts). This might require development and calibration of a more continuous severity scale (e.g. 1–9 from least severe to most severe) to reflect the phenotypic continuum of plus disease and provide a more objective measure of disease severity. We are developing tools on our website (http://www.i-rop.com) to facilitate determination of relative disease severity using both computer-based image analysis and pairwise comparison with study images.[26]

There are some limitations to this study. First, as mentioned, these results are based on expert interpretation of fundus images, and may not reflect what that expert would have classified using binocular indirect ophthalmoscopy (BIO), the gold standard for ROP diagnosis, for a number of reasons including image quality, incomplete field of view, and magnification differences. However, for one study expert (MFC) with available comparison data from the i-ROP study, the intra-grader agreement between BIO and image classification was 563/616 (91%, kappa 0.76, unpublished data). Moreover, since fundus photography is being increasingly employed for telemedicine[30,39–43] and computer-based image analysis development,[14,24,25,32,44] an understanding of the range of agreement of experts using this modality is critical. Second, these analyses did not explore how previously hypothesized reasons for inter-expert variability (e.g. field of view, attention to non-ICROP features such as venous tortuosity, etc.)[10–12,15] may relate to the apparently different cut points between experts. Computer-based image analysis systems may be able to help us better understand the relationships of all of these factors to expert behavior.[11,14,45,46] Third, the implications of these findings are limited to the diagnostic discrepancy in plus disease diagnosis in a single posterior pole image, and do not address other ICROP categories (e.g. zone and stage), or peripheral vascular features, which can influence overall disease severity in the current ICROP classification scheme.[2,9] The RSD was developed using a standard set of images including the posterior pole and peripheral sweeps, and it is unclear whether providing access to these peripheral images would impact these results. Fourth, the generalizability of these findings to the entire community of practicing ROP clinicians is unknown. We limited these analyses to 8 experts with extensive clinical and research experience to minimize any criticism of the credibility of our findings. We suspect that, on average, there may be equal or higher variation within the overall population of ROP clinicians. We hope that both the qualitative description of the variation (Figure 2) as well as the quantitative analysis of the disease severity (through computer-based image analysis systems) will improve this variability in the future. In that sense, these findings have important implications for ROP education, as we have previously demonstrated that trainees often perform poorly in clinical diagnosis of ROP.[47–50] We have developed a website (http://www.i-rop.com) to provide reference images along with common areas of disagreement and reference standard classifications for zone, stage, and plus disease diagnosis to improve trainee education and hopefully provide a reference for practicing clinicians that may improve standardization of plus disease diagnosis in the future.

These findings have important implications for ROP research, education, and clinical care. From the research perspective, we demonstrate that given the wide variability in plus disease diagnosis, careful consideration must be made of any “gold standard” for clinical research (e.g. evaluation of ROP telemedicine or image analysis programs), and that continued research using computer-based image analysis systems may yield an automated severity scale that reflects the continuous nature of the disease phenotype. In terms of education, Figure 2 and http://www.i-rop.com may be used to help trainees better understand the range of disease severity within each ordinal ICROP category. With regard to clinical care, these results may help standardize the management of infants with plus disease, to ensure a common community standard as we prospectively evaluate the clinical significance of the practice pattern variation among experts.

Acknowledgments

Financial Support:

Supported by grants R01 EY19474 (JKC, DE, SO, KEJ, RVPC, MFC), P30 EY010572 (JPC, SO, MFC), R21 EY022387 (JKC, DE, MFC), T32 EY23211 (MFC, RS), and an Oregon Clinical and Translational Research Institute (OCTRI) grant (TL1TR000129) from the National Center for Advancing Translational Sciences (NCATS) at the National Institutes of Health (NIH), Bethesda, MD; unrestricted departmental funding from Research to Prevent Blindness, New York, NY (JPC, SNP, JDR, MXR, SO, KEJ, RVPC, MFC); and the iNsight Foundation (RVPC, KEJ); The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. No funding organizations had any role in the design or conduct of this research

Conflict of interest:

MFC is an unpaid member of the Scientific Advisory Board for Clarity Medical Systems (Pleasanton, CA), and is a Consultant for Novartis (Basel, Switzerland). JDR is a Consultant for Novartis (Basel, Switzerland). RVPC is a Consultant for Visunex Medical Systems (Fremont, CA)

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

Members of the i-ROP research consortium

Oregon Health & Science University (Portland, OR): Michael F. Chiang, MD, Susan Ostmo, MS, Kemal Sonmez, PhD, J. Peter Campbell, MD, MPH. University of Illinois at Chicago (Chicago, IL): RV Paul Chan, MD, Karyn Jonas, RN. Columbia University (New York, NY): Jason Horowitz, MD, Osode Coki, RN, Cheryl-Ann Eccles, RN, Leora Sarna, RN. Bascom Palmer Eye Institute (Miami, FL): Audina Berrocal, MD, Catherin Negron, BA. William Beaumont Hospital (Royal Oak, MI): Kimberly Denser, MD, Kristi Cumming, RN, Tammy Osentoski, RN, Tammy Check, RN, Mary Zajechowski, RN. Children’s Hospital Los Angeles (Los Angeles, CA): Thomas Lee, MD, Evan Kruger, BA, Kathryn McGovern, MPH. Cedars Sinai Hospital (Los Angeles, CA): Charles Simmons, MD, Raghu Murthy, MD, Sharon Galvis, NNP. LA Biomedical Research Institute (Los Angeles, CA): Jerome Rotter, MD, Ida Chen, PhD, Xiaohui Li, MD, Kent Taylor, PhD, Kaye Roll, RN. Massachusetts General Hospital (Boston, MA): Jayashree Kalpathy-Cramer, PhD. Northeastern University (Boston, MA): Deniz Erdogmus, PhD. Asociacion para Evitar la Ceguera en Mexico (APEC) (Mexico City): Maria Ana Martinez-Castellanos, MD, Samantha Salinas-Longoria, MD, Rafael Romero, MD, Andrea Arriola, MD, Francisco Olguin-Manriquez, MD, Miroslava Meraz-Gutierrez, MD, Carlos M. Dulanto-Reinoso, MD, Cristina Montero-Mendoza, MD.

REFERENCES

1. Sommer A, Taylor HR, Ravilla TD, et al. Challenges of ophthalmic care in the developing world. JAMA Ophthalmol. 2014; 132:640–644. [PubMed: 24604415]
2. International Committee for the Classification of Retinopathy of Prematurity. The International Classification of Retinopathy of Prematurity revisited. Vol. 123. American Medical Association; 2005. p. 991-999.
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. The Committee for the Classification of Retinopathy of Prematurity. An international classification of retinopathy of prematurity. Arch Ophthalmol. 1984; 102:1130–1134. [PubMed: 6547831]
6. 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]
7. 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]
8. 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]
9. Campbell JP, Ryan MC, Lore E, et al. Diagnostic Discrepancies in Retinopathy of Prematurity Classification. Ophthalmology. 2016
10. Rao R, Jonsson NJ, Ventura C, et al. Plus disease in retinopathy of prematurity: diagnostic impact of field of view. Retina (Philadelphia, Pa). 2012; 32:1148–1155.
11. Ataer-Cansizoglu E, Bolon-Canedo V, Campbell JP, et al. Computer-based image analysis for plus disease diagnosis in retinopathy of prematurity: image analysis features associated with expert diagnosis. Transl Vis Sci Technol. 2015 In Press.
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 (Philadelphia, Pa). 2013; 33:1700–1707.
13. 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]
14. 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. 2016
15. Hewing NJ, Kaufman DR, Chan RVP, Chiang MF. Plus Disease in Retinopathy of Prematurity. JAMA Ophthalmol. 2013; 131:1026–1027. [PubMed: 23702696]
16. Ryan MC, Ostmo S, Jonas K, et al. Development and Evaluation of Reference Standards for Image-based Telemedicine Diagnosis and Clinical Research Studies in Ophthalmology. AMIA Annu Symp Proc. 2014; 2014:1902–1910. [PubMed: 25954463]

17. Team RC. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2013. Available online at: http.www.R-project.org
18. Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. J Biomed Inform. 2002; 35:99–110. [PubMed: 12474424]
19. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005; 37:360–363. [PubMed: 15883903]
20. 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]
21. 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]
22. 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]
23. 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. [PubMed: 18427596]
24. 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. 2015; 4:5.
25. Wittenberg LA, Jonsson NJ, Chan RVP, Chiang MF. Computer-based image analysis for plus disease diagnosis in retinopathy of prematurity. J Pediatr Ophthalmol Strabismus. 2012; 49:11–19. quiz 10– 20. [PubMed: 21366159]
26. Kalpathy-Cramer J, Campbell JP, Erdogmus D, et al. Plus disease in retinopathy of prematurity: improving diagnostic variability by ranking disease severity and using quantitative image analysis. Ophthalmology In Review.
27. Chan RP, Patel SN, Ryan MC. The global education network for retinopathy of prematurity (Gen-rop). Trans Ophthalmol Soc UK. 2015
28. Campbell JP, Swan R, Jonas K, et al. Implementation and evaluation of a tele-education system for the diagnosis of ophthalmic disease by international trainees. AMIA Annu Symp Proc. 2015; 2015:366–375. [PubMed: 26958168]
29. Quinn, Graham E. e-ROP Cooperative Group. Telemedicine approaches to evaluating acute-phase retinopathy of prematurity: study design. Ophthalmic Epidemiol. 2014; 21:256–267. [PubMed: 24955738]
30. Vinekar A, Gilbert C, Dogra M, et al. The KIDROP model of combining strategies for providing retinopathy of prematurity screening in underserved areas in India using wide-field imaging, tele-medicine, non-physician graders and smart phone reporting. Indian J Ophthalmol. 2014; 62:41–49. [PubMed: 24492500]
31. Williams SL, Wang L, Kane SA, et al. Telemedical diagnosis of retinopathy of prematurity: accuracy of expert versus non-expert graders. Br J Ophthalmol. 2010; 94:351–356. [PubMed: 19955195]
32. Abbey AM, Besirli CG, Musch DC, et al. Evaluation of Screening for Retinopathy of Prematurity by ROPtool or a Lay Reader. Ophthalmology. 2015; 123:385–390. [PubMed: 26681393]
33. Cabrera MT, Freedman SF, Kiely AE, et al. Combining ROPtool measurements of vascular tortuosity and width to quantify plus disease in retinopathy of prematurity. J AAPOS. 2011; 15:40–44. [PubMed: 21397804]
34. Fijalkowski N, Zheng LL, Henderson MT, et al. Stanford University Network for Diagnosis of Retinopathy of Prematurity (SUNDROP): five years of screening with telemedicine. Ophthalmic Surg Lasers Imaging Retina. 2014; 45:106–113. [PubMed: 24444469]
35. Thyparampil PJ, Park Y, Martinez-Perez M, et al. Plus Disease in Retinopathy of Prematurity (ROP): Quantitative Analysis of Vascular Change. Invest Ophthalmol Vis Sci. 2009; 50:5725–5725.
36. Wallace DK, Kylstra JA, Chesnutt DA. Prognostic significance of vascular dilation and tortuosity insufficient for plus disease in retinopathy of prematurity. Journal of American Association for Pediatric Ophthalmology and Strabismus. 2000; 4:224–229. [PubMed: 10951298]

37. RAINBOW Study: RAnibizumab Compared With Laser Therapy for the Treatment of INfants BOrn Prematurely With Retinopathy of Prematurity (RAINBOW). Available at: https://clinicaltrials.gov/ct2/show/NCT02375971

38. Abbey AM, Besirli CG, Musch DC, et al. Evaluation of Screening for Retinopathy of Prematurity by ROPtool or a Lay Reader. Ophthalmology. 2015; 123:385–390. [PubMed: 26681393]

39. Chiang MF, Melia M, Buffenn AN, et al. Detection of Clinically Significant Retinopathy of Prematurity Using Wide-angle Digital Retinal Photography. OPHTHA. 2012; 119:1272–1280.

40. Quinn GE, Ying G-S, Daniel E, et al. Validity of a Telemedicine System for the Evaluation of Acute-Phase Retinopathy of Prematurity. JAMA Ophthalmol. 2014; 132:1178–1177. [PubMed: 24970095]

41. Fierson WM, Capone A. the AMERICAN ACADEMY OF PEDIATRICS SECTION ON OPHTHALMOLOGY, AMERICAN ACADEMY OF OPHTHALMOLOGY, and AMERICAN ASSOCIATION OF CERTIFIED ORTHOPTISTS. Telemedicine for Evaluation of Retinopathy of Prematurity. Pediatrics. 2015; 135:e238–e254. [PubMed: 25548330]

42. Wang SK, Callaway NF, Wallenstein MB, et al. SUNDROP: six years of screening for retinopathy of prematurity with telemedicine. Can J Ophthalmol. 2015; 50:101–106. [PubMed: 25863848]

43. Chiang MF, Wang L, Busuioc M, et al. Telemedical retinopathy of prematurity diagnosis: accuracy, reliability, and image quality. Arch Ophthalmol. 2007; 125:1531–1538. [PubMed: 17998515]

44. Wallace DK. Computer-assisted quantification of vascular tortuosity in retinopathy of prematurity (an American Ophthalmological Society thesis). Trans Am Ophthalmol Soc. 2007; 105:594–615. [PubMed: 18427631]

45. 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]

46. 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]

47. Paul Chan RV, Williams SL, Yonekawa Y, et al. Accuracy of retinopathy of prematurity diagnosis by retinal fellows. Retina (Philadelphia, Pa). 2010; 30:958–965.

48. 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]

49. 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]

50. 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]

Variability in expert classification of plus disease in retinopathy of prematurity (ROP) is related to systematic differences in the individual cut-points for defining plus disease along the spectrum of vascular abnormalities in ROP.

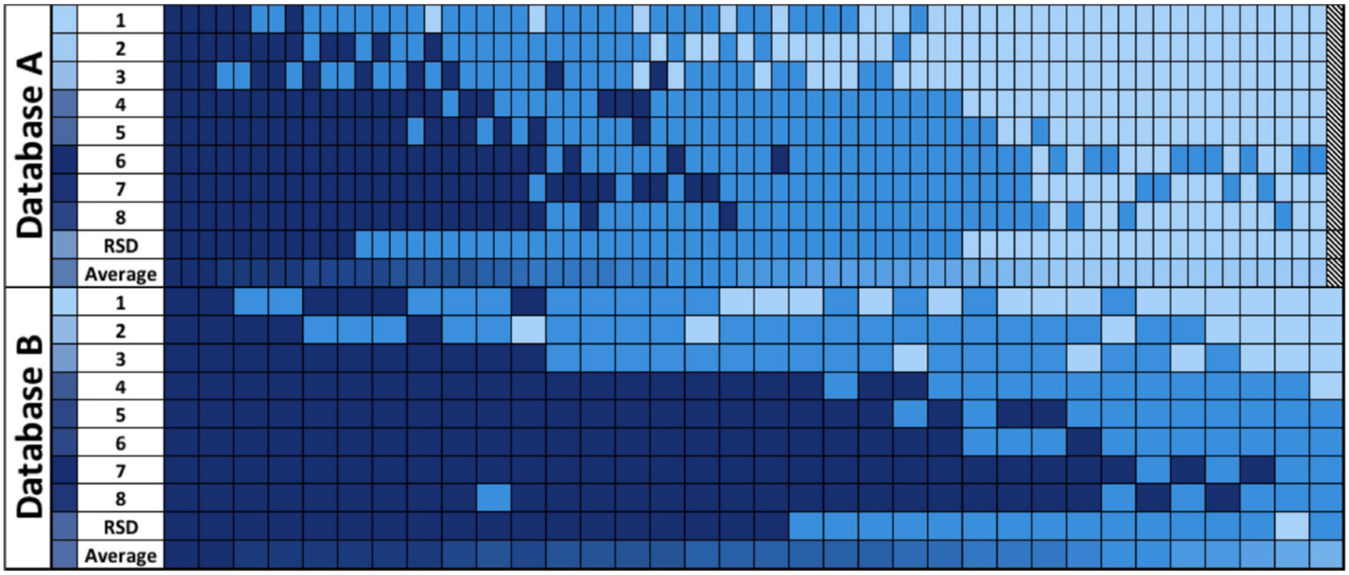
Database A
1
2
3
4
5
6
7
8
RSD
Average
Database B
1
2
3
4
5
6
7
8
RSD
Average

Figure 1.

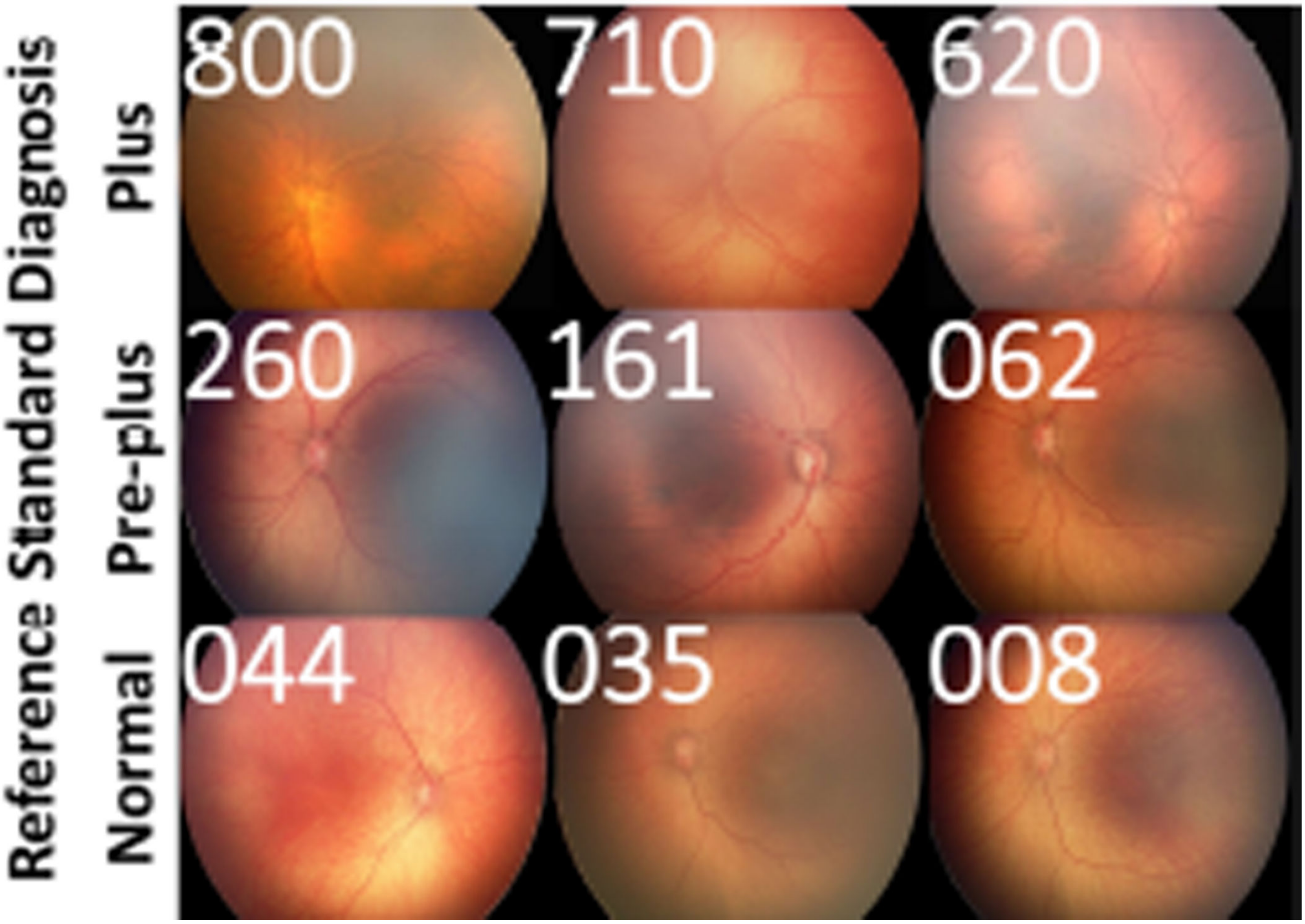
Reference Standard Diagnosis
Plus
Pre-plus
Normal
800
710
620
260
161
062
044
035
008

Figure 2.

Table 1

Distribution of disease classifications (plus, pre-plus, or normal) of 100 wide-angle images (A) and 34 wide-angle images (B) by 8 retinopathy of prematurity (ROP) experts.

	Database A						Database B					
Expert	**Plus**	**Pre-plus**	**Normal**	**Average**	**Bias**	**RSD Agreement (%)**	**Plus**	**Pre-plus**	**Normal**	**Average**	**Bias**	**RSD Agreement (%)**
1	6	30	64	1.4	−0.19	81%	6	14	14	1.8	−0.79	29%
2	12	20	68	1.4	−0.17	81%	5	22	7	2.1	−0.41	62%
3	11	25	64	1.5	−0.14	80%	11	17	6	1.9	−0.62	38%
4	21	25	54	1.7	0.06	94%	21	12	1	2.6	0.03	91%
5	20	29	51	1.7	0.08	90%	24	10	0	2.7	0.15	79%
6	24	30	46	1.8	0.17	83%	28	6	0	2.8	0.15	85%
7	29	25	46	1.8	0.22	78%	30	4	0	2.9	0.32	71%
8	25	34	41	1.8	0.23	77%	24	10	0	2.7	0.26	68%
RSD	15	31	54	1.6	N/A	N/A	20	13	1	2.6	N/A	N/A

RSD = reference standard diagnosis. Bias = mean deviation from RSD. Average = average severity score for the database for each expert (plus = 3, pre-plus = 2, normal = 1).