Published in final edited form as:
Ophthalmology. 2016 November ; 123(11): 2345–2351. doi:10.1016/j.ophtha.2016.07.020.

Plus disease in retinopathy of prematurity: improving diagnosis by ranking disease severity and using quantitative image analysis

Jayashree Kalpathy-Cramer, PhD[1,*], **J. Peter Campbell, MD, MPH**[2,*], **Deniz Erdogmus, PhD**[3], **Peng Tian, BE**[3], **Dharanish Kedarisetti, MSc**[3], **Chace Moleta, MS**[2], **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, PhD**[11], **Ryan Swan, BS**[11], **Susan Ostmo, MS**[2], **Karyn E. Jonas, RN**[12], **R.V. Paul Chan, MD**[12], and **Michael F. Chiang, MD**[2,11] **on behalf of the i-ROP research consortium**

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

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

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. Kalpathy-Cramer and Campbell 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 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.

This paper has been submitted for consideration as a paper presentation for the 2016 AAO meeting in Chicago.

Abstract

Objective—To determine expert agreement on relative retinopathy of prematurity (ROP) disease severity, whether computer-based image analysis can model relative disease severity, and to propose consideration of a more continuous severity score for ROP.

Design—We developed two databases of clinical images of varying disease severity (100 images and 34 images) as part of the i-ROP (Imaging and Informatics in ROP) cohort study and recruited both expert physician, non-expert physician, and non-physician graders to classify and perform pairwise comparisons on both databases.

Subjects, Participants, and/or Controls—Images obtained during routine ROP screening in neonatal intensive care units. 6 participating expert ROP clinician-scientists, each with a minimum of 10 years clinical ROP experience and 5 ROP publications. 5 image graders (3 physicians and 2 non-physician graders).

Methods—Images in both databases were ranked by average disease classification (classification ranking) and by pairwise comparison using the Elo rating method (comparison ranking), and correlation with the i-ROP computer-based image analysis system.

Main Outcome Measures—Inter-expert agreement (weighted kappa statistic) compared with correlation coefficient (CC) between experts on pairwise comparisons, and correlation between expert rankings and computer-based image analysis modeling.

Results—There was variable inter-expert agreement on diagnostic classification of disease (plus, pre-plus, or normal) among the 6 experts (mean weighted kappa 0.27, range 0.06–0.63), but good correlation between experts on comparison ranking of disease severity (mean CC 0.84, range 0.74–0.93) on the set of 34 images. Comparison ranking provided a severity ranking that was in good agreement with ranking obtained by classification ranking (CC 0.92). Comparison ranking on the larger dataset by both expert and non-expert graders demonstrated good correlation (mean CC 0.97, range 0.95–0.98). The i-ROP system was able to model this continuous severity with good correlation (CC 0.86).

Conclusions—Experts diagnose plus disease on a continuum with poor absolute agreement on classification, but good relative agreement on disease severity. These results suggest that the use of pairwise rankings and a continuous severity score, such as that provided by the i-ROP system, may improve agreement on disease severity in the future.

Retinopathy of prematurity (ROP) is an important cause of childhood blindness throughout the world, despite major advances in disease screening and treatment.[1–4] The International Classification of ROP (ICROP) defined three clinical examination parameters: zone, stage, and plus disease.[5,6] Since that time, NIH-sponsored multicenter randomized controlled trials have shown that presence of plus disease is the critical feature for identifying severe treatment-requiring ROP. Therefore, accurate identification of plus disease is essential to ensure that infants at higher risk of vision loss are appropriately treated, and those who are at lower risk are not subjected to the potential risks of over-treatment.

Plus disease is defined as arterial tortuosity and venous dilation greater than a standard published photograph. In the revised ICROP, an intermediate “pre-plus” state was defined as greater arterial tortuosity and venous dilation than normal, but less than the standard photograph.[4,5] In this way, the ICROP forced separation of a continuous spectrum of disease into three discrete diagnostic categories: plus, pre-plus, and normal. With only rare exceptions (zone I, stage 3; zone III disease), infants with plus disease require treatment, whereas those with less than plus disease can be safely observed.[5] However, it has been well established that clinicians frequently disagree on diagnostic classification of plus disease. There have been multiple explanations proposed for these discrepancies including differences among "cut points" of vascular abnormality required for plus disease, differences in field of view considered, identification of different vascular parameters (e.g. venous tortuosity) by different clinicians, and inconsistencies in training and education.[7–15]

As part of the Imaging and Informatics in ROP (i-ROP) study, we have collected multiple expert diagnoses from large numbers of examinations and images. In an accompanying paper, we have demonstrated that a key explanation for diagnostic discrepancies among established ROP experts is due to different "cut-points" for the amounts of vascular abnormality required by individual experts to distinguish between plus vs. pre-plus (as well as pre-plus vs. normal).[16] In that work, we showed that much of the disagreement among different experts appears to be caused by a systematic bias for each observer to "over-call" or "under-call" disease compared to others. However, it is not known whether different experts have the same perceptions of the relative severity of retinal vascular abnormality (e.g. whether different experts would order images similarly from least to most severe, regardless of their agreement on the diagnostic classification as plus vs. pre-plus vs. normal). From a clinical perspective, this ability to identify relative disease severity in ROP is critical to determine disease progression, recognize improvement after treatment, and determine follow-up interval between examinations.[17–20] This is a significant gap in knowledge that must be understood to improve clinical ROP diagnosis, to identify retinal vascular features that are most important in clinical diagnosis, and to develop and validate computer-based image analysis systems that quantify vascular abnormality.

Our purpose is to examine the relative agreement among experts regarding disease severity in ROP. In this paper, we show that there is strong agreement among experts in relative ordering of ROP disease severity even when they disagree on diagnostic classification (plus vs. pre-plus vs. normal), demonstrate that computer-based image analysis can model this relative disease severity, and propose consideration of a more continuous severity score 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 for imaging and study participation.

Description of datasets

We developed 2 datasets of wide-angle retinal images acquired during routine clinical care and established a reference standard diagnosis (RSD) for each image using previously published methods that combine the classifications of expert ROP image graders (independent, masked classifications from two ophthalmologists and one non-physician ROP study coordinator) and the clinical diagnosis. The first dataset (A) included 100 images of varying disease severity, among which 15 had a RSD of plus disease, 31 had pre-plus disease, and 54 were normal. The second dataset (B) was designed to focus on the moderate-to-severe end of the disease spectrum, and included 34 images, of which 20 had an RSD of plus disease, 13 had pre-plus disease, and 1 was normal. Two datasets were used for this study to examine the performance and generalizability of clinical diagnosis, as well as computer-based analysis, in multiple datasets with different underlying disease prevalence.

Severity ranking methodology

We developed an open-source, web-based, image severity assessment platform through our project website (http://www.i-rop.com) which allowed us to easily collect image classification and comparison data from experts located around the world. There were two scenarios for providing expert input to determine relative disease severity among images in each dataset.

In the first scenario ("classification ranking"), images were ranked in severity based on the average diagnostic classification (plus, pre-plus, or normal) by 8 experts. Each expert was presented one image at a time and asked to provide a diagnostic classification for the image ("plus", "pre-plus", or "normal"). All experts were presented the series of images in the same order to control for any unexpected effects from differences in image presentation. Using these data we were able to develop a severity ranking based on the average disease classification of all experts (as described in the accompanying paper),[16] which we refer to as "classification rankings". This classification task was completed by 8 ROP experts, as described in the accompanying paper.[16] In this paper, we present results from 6 experts who also completed the comparison ranking described below.

In the second task ("comparison ranking"), image graders were presented a pair of images and asked to "click on the image that represents more severe disease." All users were presented the image pairs in the same order. To determine the relative ranking of images, we used pairwise comparisons to create an ordered set of images. Ranking is a commonly studied problem in many areas of research including machine learning, recommender systems, player rankings, and biology.[21] The primary goal of these analyses is to convert a set of potentially noisy, incomplete, pairwise comparisons into an ordered set. There are many algorithms to convert pairwise comparisons into rankings, which we refer to as "comparison rankings." We used the "Elo" algorithm as a simple approach that has been shown to work well.[19] In dataset A (100 images), "comparison rankings" were performed by a total of 5 expert and non-expert raters: 2 experts (among the 8 who performed classification rankings), 1 additional pediatric retina specialist, and 2 non-physician ROP study coordinators. In database B (34 images), comparison rankings were performed by a total of 6 expert raters: 6 (of the 8) original experts who performed the classification rankings. In addition to the individual comparison rankings by user, we created a consensus comparison ranking using the Elo algorithm that considered pairwise comparisons by all raters.

Computer-based image analysis development

Images in Dataset A were manually segmented and used to develop a computer-based image analysis system (the i-ROP system), using methods described previously.[9,22] The i-ROP system extracted tortuosity and dilation features from each point of the vessels, collectively analyzing arteries and veins, which yielded both tortuosity and dilation value pools for each image. Based on the value pool, each feature was represented by a Gaussian mixture model. A regression model,[22,23] establishing the relationship between image features and the image rank of experts, was employed to provide a relative severity score for each image. In particular, we used a supervised linear regression approach and 10-fold cross validation to generate an “i-ROP regression score” based on image features including tortuosity and dilation.

Data analysis

Statistical analysis was performed using Excel 2016 (Microsoft, Redmond, WA), Stata v. 11.0 (College Station, TX), and R v3.2.2.[24] Inter-expert agreement on classification ranking was assessed using the weighted kappa statistic between pairs of experts, and interpreted using a commonly accepted scale: 0–0.20, slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.00, near-perfect agreement.[25] Inter-expert agreement on comparison ranking was assessed using the correlation coefficients between the rankings of experts.

To determine the similarity of the ranking methodologies, we calculated the correlation coefficient (CC) between classification rankings and the consensus comparison rankings. A CC of 1 would imply that the orders are identical between the ranks using the classification and the ranks using the pairwise comparisons, while a CC of 0 would suggest that there is no association between the ranks. We then calculated the correlation coefficients for the individual comparison rankings between all pairs of raters. Finally, we calculated the correlation between the i-ROP regression score versus the consensus comparison rankings for Database A.

Results

Inter-Expert Agreement of Classification Ranking vs. Comparison Ranking

Six experts completed both the classification and the comparison ranking tasks for Database B (34 images), which was designed to focus on moderate to severe levels of vascular abnormality (where the inter-expert agreement is lower).[16] We compared the weighted kappa values with the equivalent correlation coefficients for the image ranks between these 6 raters. As seen in Figure 1, the inter-expert agreement using classification ranking (plus vs. pre-plus vs. normal) was slight to substantial (mean weighted kappa 0.27, range 0.06–0.63 between pairs of experts). However, the agreement using comparison ranking was consistently very high (mean CC 0.84, range 0.74–0.93 between pairs of experts). Overall, this suggests that experts are very consistent in identifying relative severity of retinal vascular abnormality (e.g. pairwise ranking in order of disease severity), but less consistent in diagnostic classification (e.g. plus vs. pre-plus vs. normal).

Relative Image Severity Determined by Comparison Ranking vs. Classification Ranking

We then compared the classification rankings[16] with the comparison rankings. As seen in Figure 2, there is high correlation between these ranks (CC 0.92), suggesting that both of these approaches for attaining consensus ranking of image severity by experts result in consistent severity gradings for ROP.

We then analyzed Dataset A, which had a broader distribution of disease severity (more normal images) using comparison rankings from three physicians and 2 non-physician graders to determine the consistency of comparison rankings between raters of varying expertise,. As seen in Figure 3, the comparison rankings by all 5 graders were highly consistent (mean CC 0.97, range 0.95–0.98) between each expert and the consensus comparison ranking. There was more disagreement at the normal end of the disease spectrum, suggesting that it may be challenging to determine which normal images are “more severe” than others. The comparison rankings by all 5 experts agreed well with the established reference standard diagnosis (Figure 3).

Computer-based Image Analysis for Ranking: i-ROP system validation

To assess feasibility of using computer-based image analysis to assess disease severity, the i-ROP severity score was measured against the consensus comparison rankings by experts. The i-ROP severity score correlated well with the comparison rankings for the images in Dataset A (CC 0.86). Figure 4 displays the i-ROP regression score for each of the 9 representative images from the accompanying paper. Images were chosen to demonstrate a continuous range of disease severity, as reflected by variable expert classification, within each reference standard diagnosis category (plus, pre-plus, or normal).

Discussion

This study compares inter-expert agreement using the ICROP disease classification (plus, pre-plus, or normal) with relative disease severity rankings in order to determine whether a more continuous severity scale for plus disease can improve diagnostic agreement. Key findings from this study are: (1) Even among experts with poor agreement on plus disease classification there is high agreement on ranking of relative disease severity, (2) using pairwise comparison rankings may have clinical utility for determining relative disease severity and assigning a quantitative severity scale, and (3) the i-ROP system performs comparably to experienced graders in determining relative disease severity using manually segmented images.

The first key finding is that inter-expert agreement on relative disease severity ranking is high (Figure 1B and Figure 3). This suggests that, although clinicians are not reliable at disease classification (plus vs. pre-plus vs. normal), they are more reliable at recognizing change in disease and identifying the more severe image. In conjunction with the accompanying paper, this strongly supports the hypothesis that inter-expert discrepancy in plus disease classification is due largely to systematic bias based on subjective cut point differences among experts regarding the amount of vascular abnormality required to distinguish between plus vs. pre-plus disease (or between pre-plus vs. normal). These results further emphasize the continuous nature of the vascular abnormalities in ROP and the continuum of expert responses along the spectrum of disease severity, even among eyes with the same reference standard diagnosis. We have demonstrated that one way to rank images is to simply average a number of responses to disease classifications.[16] Among images with a reference standard diagnosis of plus disease, images with unanimous classification appear both qualitatively and quantitatively more severe than images with less consensus, and the average score of multiple expert classification reflects this relative severity (Figure 4).

This leads to the second key finding, which is that the use of pairwise comparison rankings may have clinical utility in providing a relative disease severity score. Using a set of images similar to Figure 4, a small number of pairwise comparisons could assign a relative severity score that could then be used in several ways. First, it could complement the clinical exam and be interpreted in the context of the clinician's judgment. In the same way that a laboratory value may or may not prompt a clinical response, depending on the clinical scenario, ophthalmologists could use a relative measurement of the vascular abnormalities in ROP to aid their judgment of disease severity at a given point in time. A standardized scale would have the added benefit of improving clinician agreement on disease severity; even if experts disagree on the diagnosis of pre-plus disease – they will tend to agree that a pre-plus image is more severe than a normal image. Second, the score could be tracked over time to represent the evolution of the disease, and any signs of relative disease progression or regression, both of which could help contextualize the need for treatment.[18–20,26]

We are developing the capability through our project website (http://www.i-rop.com), by providing a feature for users to click through a set of pairwise comparisons to compare an image they are evaluating (e.g. through wide-angle imaging or indirect ophthalmoscopy) to our database of images from the i-ROP study. From a technical perspective, it would be straightforward to incorporate this functionality into existing camera software platforms, even without integrated computer-based image analysis methods, allowing users to make a few clicks and compare their image to the standard database. Similarly, this algorithm could be incorporated into existing telemedicine platforms[27–31] and could produce a severity score for any image that would not only be highly likely to agree with that same score obtained by another grader, but which could be tracked over time. Our results suggest that trained graders can perform these comparisons comparably to expert graders, however the two study coordinators included here have many years of experience with ROP and it is unclear whether these results are generalizable to graders with less experience.

The third key finding is that the i-ROP system can model these vascular abnormalities (tortuosity and dilation) in ROP and produce a regression score which correlates well with the other relative measures of disease severity. Computer-based image analysis has been previously studied as an approach toward improving ROP diagnosis both clinically and in telemedicine,[9,22,32–36] and the use of a validated automated algorithm would add an element of objectivity to the severity scoring system. In the context of telemedicine, these scores could be used to identify children at moderate to high risk of developing referral warranted or treatment requiring disease and needing either closer interval telemedical or ophthalmoscopic evaluation. These results suggest that systems such as i-ROP may improve diagnostic agreement and play a role in ROP disease diagnosis and management in the near future.

The main limitation to the applicability of computer-based image analysis findings relates to the state of validation of the current systems, such as ROPTool and the i-ROP system.[9,22,32,37] ROPTool uses a semi-automated algorithm that requires the user to identify the location of the optic disc and a number of vessels for analysis, and provides an output tortuosity scale. A recent validation study suggests that the median ROPTool tortuosity scores of images classified by experts on a 0–4 ordinal scale (no dilation, questionable, mild dilation, moderate dilation and mild tortuosity, obvious dilation and tortuosity) tended to increase with disease severity, but the differences were not statistically significant. Using the semi-automated measurement of tortuosity, the system was highly sensitive (95%) at detecting at least questionable disease with a specificity of 64%, but the correlation with the expert grading of plus disease was not reported.[37] The i-ROP data presented here are from the previously reported version of the i-ROP system, which was trained using manually segmented images.[9,22] In order for this to be utilized clinically, the challenge of designing fully automated (or semi-automated as in ROPTool) algorithms will need to be overcome.

These results demonstrate that a more continuous severity score may aid diagnostic agreement in ROP, pairwise ranking methodology may be an efficient way to assign a severity score to an image in clinical care and telemedicine, and that computer-based image analysis systems such as the i-ROP system may be able to objectively and automatically do this in the near future.

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. 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]
3. 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]
4. 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]
5. International Committee for the Classification of Retinopathy of Prematurity. The International Classification of Retinopathy of Prematurity revisited. American Medical Association. 2005; 123:991–999.
6. The Committee for the Classification of Retinopathy of Prematurity. An international classification of retinopathy of prematurity. Arch Ophthalmol. 1984; 102:1130–1134. [PubMed: 6547831]
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. 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.
9. 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
10. Nagiel A, Espiritu MJ, Wong RK, et al. Retinopathy of prematurity residency training. Ophthalmology. 2012; 119:2644–2645. e1–e2. [PubMed: 23207022]
11. 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]
12. 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]
13. 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]
14. 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.
15. 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.

16. Campbell JP, Kalpathy-Cramer J, Erdogmus D, et al. Plus disease in retinopathy of prematurity: a continuous spectrum of vascular abnormality as basis of diagnostic variability. Ophthalmology In Review.

17. Myung JS, Gelman R, Aaker GD, et al. Evaluation of vascular disease progression in retinopathy of prematurity using static and dynamic retinal images. Am J Ophthalmol. 2012; 153:544–551. e2. [PubMed: 22019222]

18. 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.

19. Wallace DK, Freedman SF, Hartnett ME, Quinn GE. Predictive value of pre-plus disease in retinopathy of prematurity. Arch Ophthalmol. 2011; 129:591–596. [PubMed: 21555612]

20. 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]

21. Saaty TL. Relative measurement and its generalization in decision making why pairwise comparisons are central in mathematics for the measurement of intangible factors the analytic hierarchy/network process. Rev R Acad Cien Serie A Mat. 2008; 102:251–318.

22. 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.

23. Bolon-Canedo, V.; Ataer-Cansizoglu, E.; Erdogmus, D., et al. A GMM-based feature extraction technique for the automated diagnosis of Retinopathy of Prematurity; 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI 2015); 2015. p. 1498-1501.

24. R Core Team. A language and environment for statistical computing. Available at: http://www.R-project.org.

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

26. Myung JS, Gelman R, Aaker GD, et al. Evaluation of Vascular Disease Progression in Retinopathy of Prematurity Using Static and Dynamic Retinal Images. Am J Ophthalmol. 2012; 153:544–551. e2. [PubMed: 22019222]

27. Graham E Quinn, e-ROP Cooperative Group. Telemedicine approaches to evaluating acute-phase retinopathy of prematurity: study design. Ophthalmic Epidemiol. 2014; 21:256–267. [PubMed: 24955738]

28. 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]

29. 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]

30. Richter GM, Williams SL, Starren J, et al. Telemedicine for retinopathy of prematurity diagnosis: evaluation and challenges. Survey of Ophthalmology. 2009; 54:671–685. [PubMed: 19665742]

31. 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]

32. 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]

33. 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 11–9–quiz 10– 20.

34. Koreen S, Gelman R, Martinez-Perez ME, et al. Evaluation of a Computer-Based System for Plus Disease Diagnosis in Retinopathy of Prematurity. Ophthalmology. 2007; 114:e59–e67. [PubMed: 18054630]

35. Chiang MF. Image analysis for retinopathy of prematurity: where are we headed? J AAPOS. 2012; 16:411–412. [PubMed: 23084374]

36. Davitt BV, Wallace DK. Plus Disease. Survey of Ophthalmology. 2009; 54:663–670. [PubMed: 19665743]

37. 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]

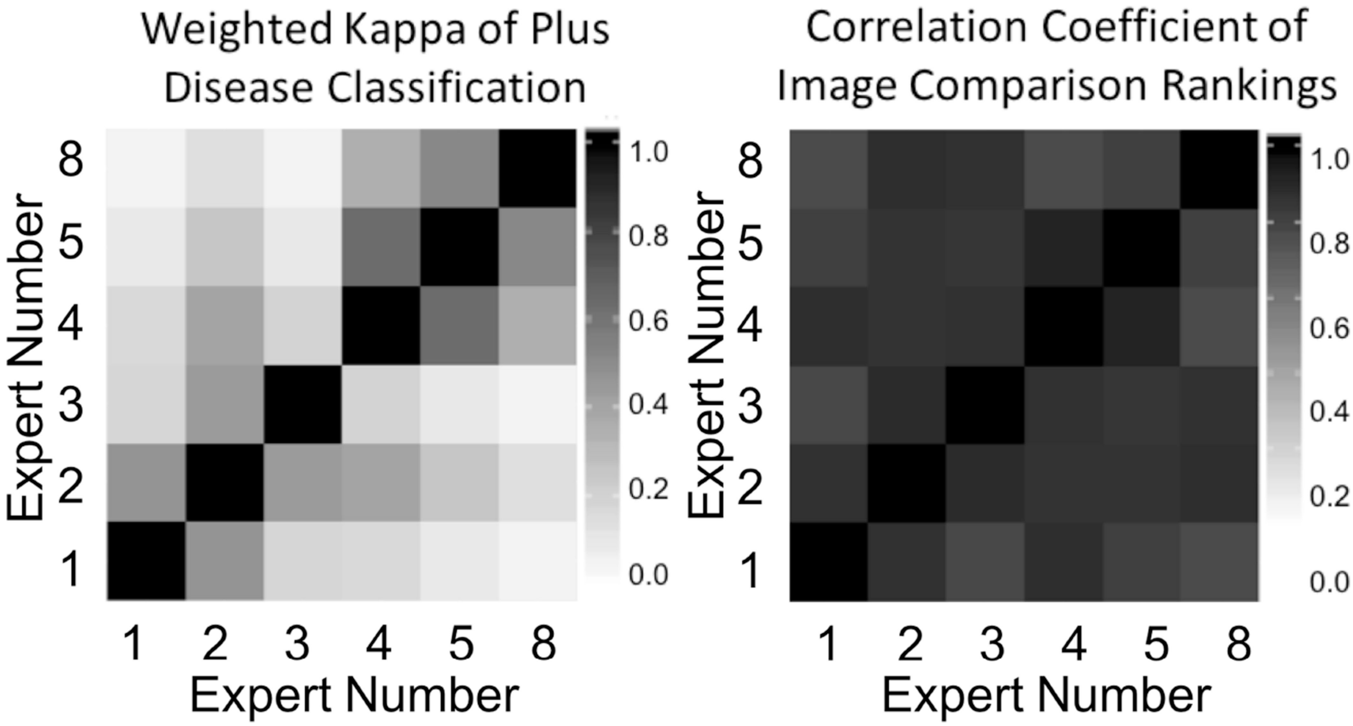

Figure 1.

Graph showing that interexpert agreement using classification ranking (plus vs. preplus vs. normal) was slight to substantial (mean weighted κ, 0.27; range, 0.06–0.63 between pairs of experts).

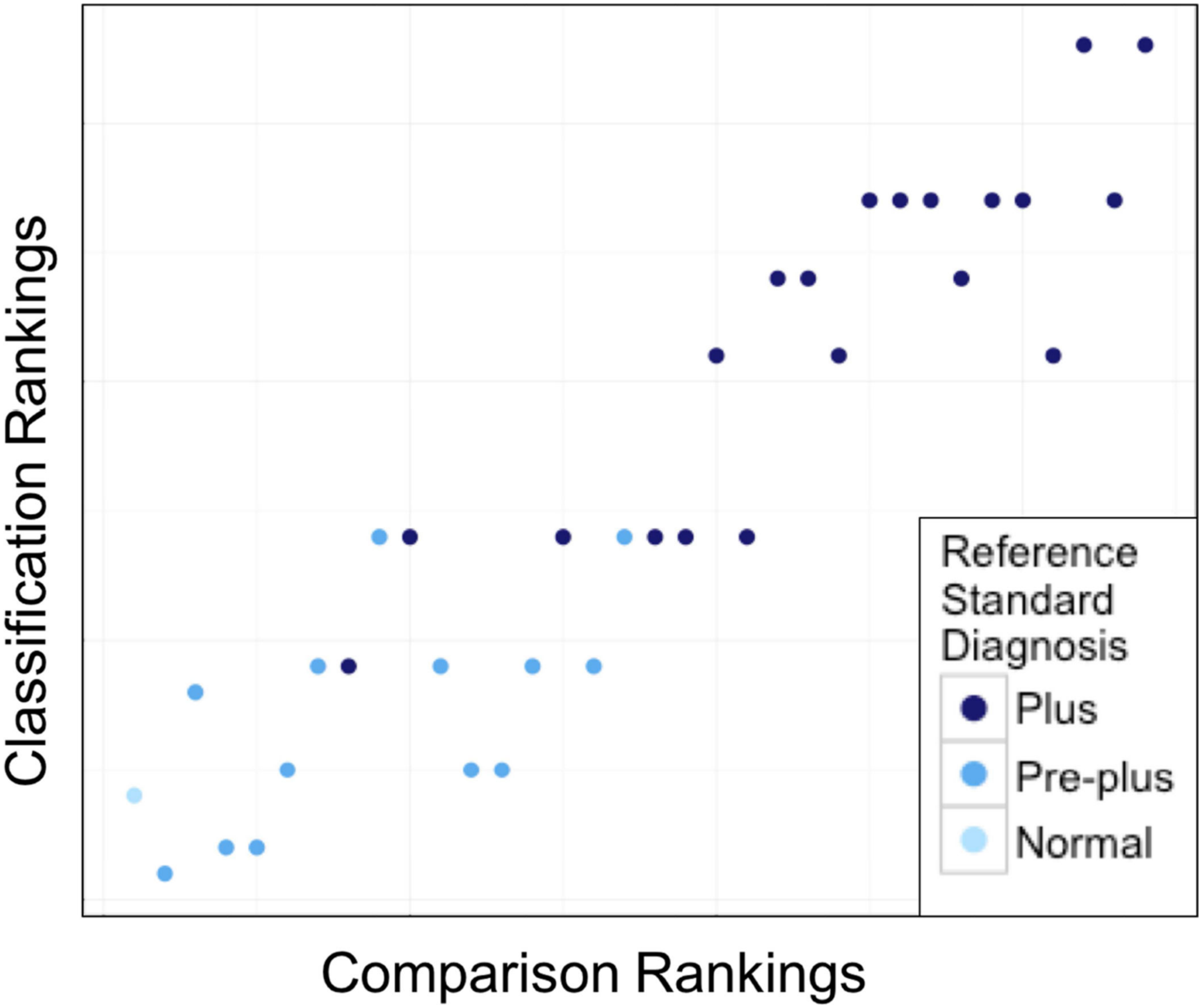

Figure 2.

Scatterplot showing the high correlation between comparison ranks (correlation coefficient, 0.92), suggesting that both of these approaches for attaining consensus ranking of image severity by experts result in consistent severity gradings for retinopathy of prematurity.

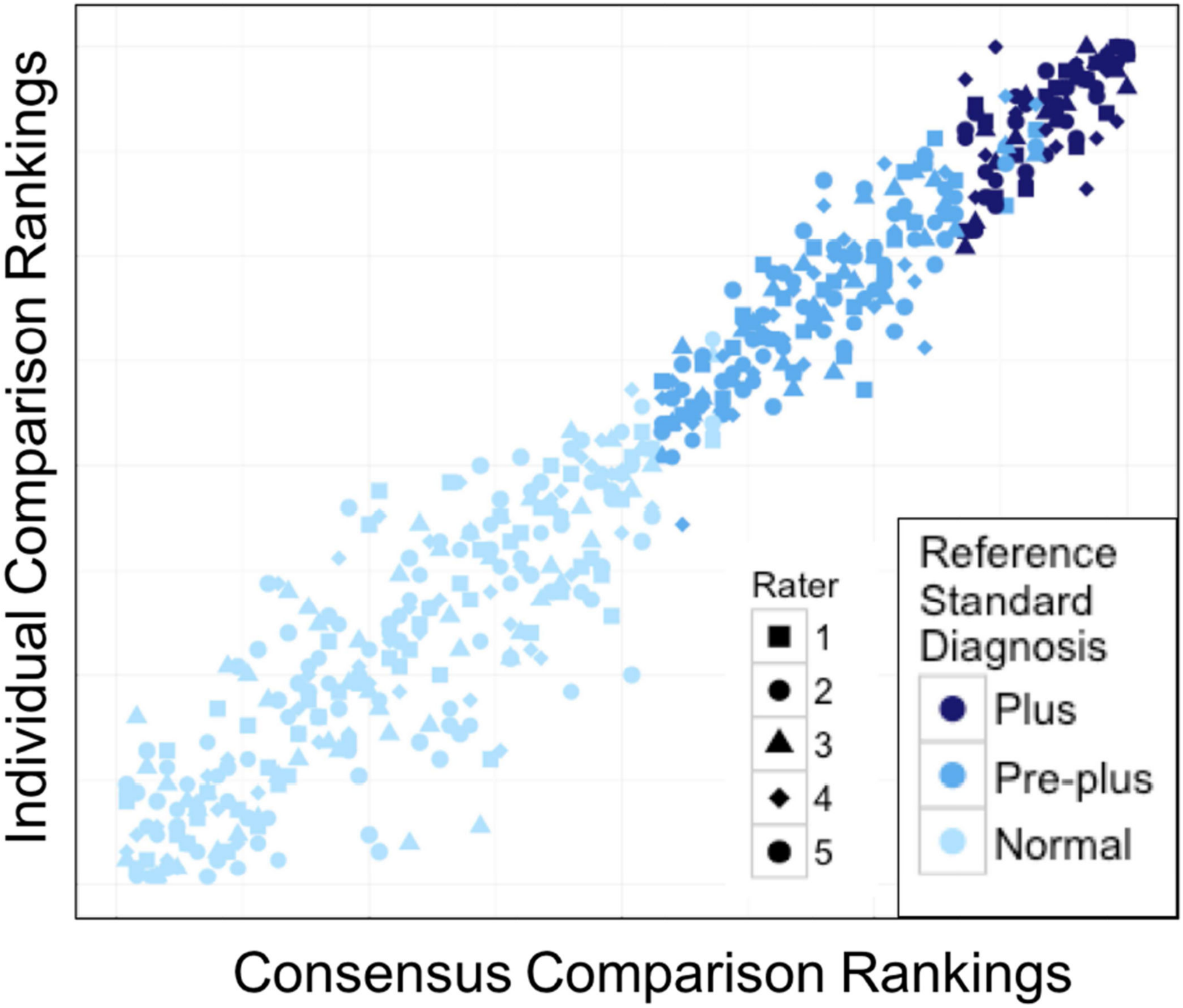

Figure 3.

Scatterplot showing that the comparison rankings by all 5 graders were highly consistent (mean correlation coefficient, 0.97; range, 0.95–0.98) between each expert and the consensus comparison ranking.

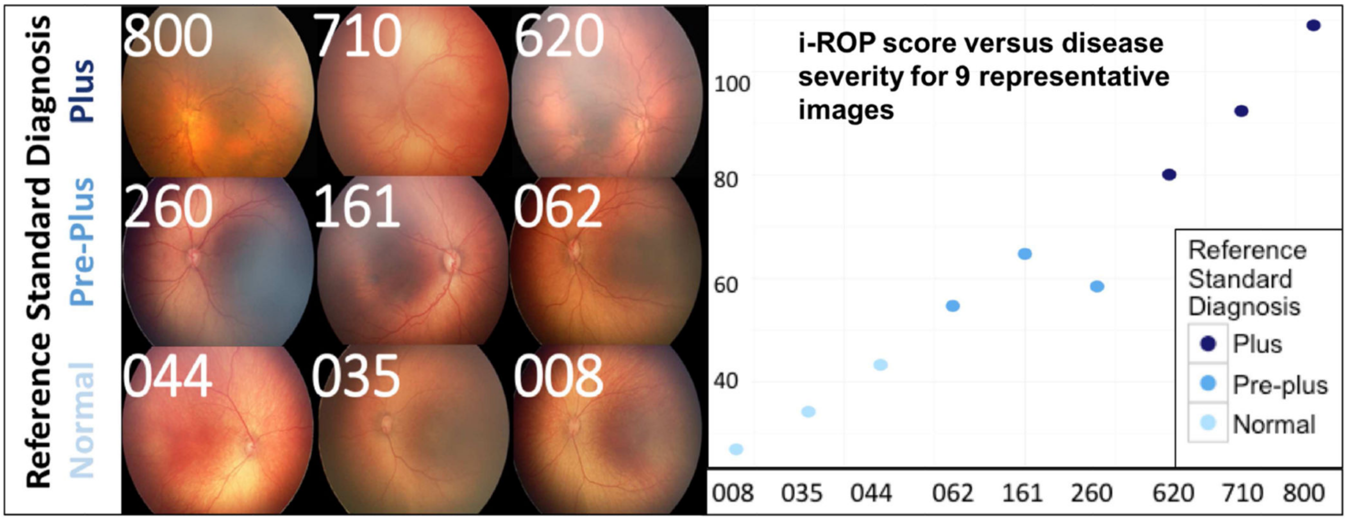

Figure 4.

The Imaging and Informatics in Retinopathy of Prematurity (i-ROP) regression score for each of the 9 representative images.