Plus Disease in Retinopathy of Prematurity: Diagnostic Trends in 2016 vs. 2007

Short title: Temporal trends in the diagnosis of plus disease in ROP

Chace Moleta¹, J. Peter Campbell¹, Jayashree Kalpathy-Cramer², RV Paul Chan³,
Susan Ostmo¹, Karyn Jonas³, Michael F. Chiang^{1,4}, on behalf of the Imaging &
Informatics in ROP Research Consortium

INTRODUCTION

Retinopathy of Prematurity (ROP) is a vasoproliferative disease that affects low-birth-weight infants. Major advances in screening and treatment guidelines established by the Cryotherapy for Retinopathy of Prematurity (CRYO-ROP) and Early Treatment for Retinopathy of Prematurity (ETROP) trials have decreased the risk of adverse outcomes and vision loss associated with ROP.^{1,2} Promising pharmacological treatments have also been evaluated in recent years.³ Despite this, ROP continues to be a leading cause of childhood blindness in the United States and worldwide.⁴

The International Classification of ROP (ICROP) defines the standard terminology for describing ROP and involves a number of parameters. Among these parameters, the presence of plus disease has been shown to be the most critical prognostic indicator in determining the need for potentially vision-saving treatment in infants with severe ROP.^{1,2,5} Plus disease is defined as retinal venous dilation and arterial tortuosity within the posterior pole which is greater than or equal to the vascular abnormality in a standard photograph selected by expert consensus during the 1980s.⁶ In 2005, the two-level classification scale for plus disease (plus vs. normal) was expanded to a three-level scale (plus vs. pre-plus vs. normal) by the revised-ICROP, which introduced the term “pre-plus” to describe a degree of arterial tortuosity and venous dilation that is more than normal, but less than plus disease.⁵

This dependence on visual and descriptive qualifiers, rather than on quantifiable measurements, creates potential for high subjectivity and variability in the clinical diagnosis of plus disease.⁷⁻¹⁴ Furthermore, studies have shown that there are significant inconsistencies and shortcomings in ROP training in residency and fellowship programs.¹⁵⁻¹⁹ These issues are important because accurate and reproducible diagnosis of plus disease is required to avoid under-treatment, which is associated with preventable blindness, and over-treatment, which may be associated with potential morbidities from laser photocoagulation or pharmacological therapy in infants.²⁰⁻²²

The participating clinicians in this study have high-volume ROP practices at academic referral centers throughout the United States. We have recently demonstrated that much of inter-expert variability in plus disease diagnosis is due to systematically different “cut-points” for how much dilation and tortuosity is required for plus disease between experts.^{23,24} We hypothesized that this cut-point may have a temporal trend as well. It has been our anecdotal impression that the clinical diagnosis of plus disease in ROP may be trending towards less severe cut points for plus disease, and that some ophthalmologists may be treating ROP at earlier levels of disease than before. The purpose of this study is to examine this hypothesis using a data set of 34 wide-angle retinal images that was reviewed in 2016 by 13 ROP experts, and previously in 2007 by 22 ROP experts for the presence plus disease.⁷

METHODS

This study was conducted as part of the multicenter prospective “Imaging and Informatics in ROP” (i-ROP) cohort study and received prospective approval by the Institutional Review Board at Oregon Health & Science University. Informed consent

was obtained from parents of all infants with ROP, and deidentified images were used for analysis. This study was conducted in accordance with Health Insurance Portability and Accountability Act (HIPAA) guidelines.

Study Experts and Data Sets

A data set of 34 wide-angle retinal images of the posterior pole was obtained using a commercially available device (RetCam; Clarity Medical Systems, Pleasanton, California) during routine ROP care of infants. No identifying information such as participant name, birth weight, or gestational age was associated with images, and no images were repeated.

Two groups of ROP experts were invited to participate in grading images for the study, first in 2007 and again 2016.⁷ For the objective of this study, eligible experts were defined as practicing pediatric ophthalmologists or retina specialists who satisfied at least 1 of the following criteria: having been a study center investigator for the CRYO-ROP or ETROP study or having coauthored at least 5 peer-reviewed ROP manuscripts. 22 experts participated in 2007 and 13 in 2016. Four of these experts participated in both years. In the 2007 presentation, any visible peripheral disease was cropped out of images, which affected 12 of the 34 images.⁷ In the 2016 presentation, images were not cropped.

Image Interpretation

De-identified images were uploaded to a Web server and stored in a secure database system, with a secure Web interface developed by the authors to display retinal photographs and collect responses. Each expert grader was provided with an anonymous study identifier.

Images were presented one at a time and experts were asked to provide a diagnostic classification for the image (i.e., “normal,” “pre-plus,” or “plus,” which were scored as “1,” “2,” or “3,” respectively). The order of image presentation was standardized for all participants to control for any unexpected effects from differences in image presentation.

Data Analysis

Data were exported from the study database into a spreadsheet (Excel 2011; Microsoft, Redmond, WA). A plus disease severity score was calculated for each image in 2016 and 2007 by averaging all scores (normal = 1, pre-plus = 2, plus = 3) provided for that image by respective expert cohorts. Average scores were used to compare and rank images for plus disease severity in both time periods. For the purpose of ranking images, ties of identical severity scores in each cohort were broken using an average of experts from both time periods combined.

The weighted κ statistic was used for analysis of agreement because it adjusts for relative disagreements in ordinal categories.²⁵ Interpretation of results was based on an accepted scale: 0 to 0.20 as 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, almost perfect agreement.²⁶ A mean weighted κ value was determined for each expert compared with all others in 2007 and again 2016. Weighted κ statistic was also

determined for each repeat expert (i.e., those who participated in both cohorts) in 2016 compared against themselves in 2007.

Statistical analysis was performed using IBM SPSS Statistics for Windows, Version 22.0 software (IBM Corp., Armonk, New York). Differences in grading of plus disease severity by repeat experts in 2016 vs. 2007 were analyzed using the Wilcoxon Signed Rank Test, a non-parametric alternative to the paired *t* test. Statistical significance was defined as $P < .05$.

RESULTS

Characteristics of Study Experts

Individuals were invited to participate in image grading based on the study definition of expertise in ROP. For the 2007 cohort, invitations were extended to 29 experts, of whom 22 (76%) ultimately consented to participate. All (100%) of these experts served as principal investigators or certified investigators in the CRYO-ROP or ETROP study, with 11 (50%) also having coauthored ≥ 5 peer-reviewed ROP manuscripts. For the 2016 cohort, 15 individuals were invited, of whom 13 (87%) participated. Six (46%) of these experts served as principal investigators or certified investigators in the CRYO-ROP or ETROP study, and all (100%) coauthored ≥ 5 peer-reviewed ROP manuscripts. Four experts participated in both cohorts.

Plus Disease Grading, Severity Ranking, and Inter-Expert Agreement Over Time

Figure 1 displays the comparison of average plus disease severity scores between 2016 (n=13 experts) vs. 2007 (n=22 experts). The average score for each image was higher for 30/34 (88%) images in 2016 compared to 2007. One (2.9%) image had a severity score that was unchanged, with all experts assigning it a grade of "3" (plus disease) in both cohorts. The remaining 3/34 (8.8%) images had average scores that were slightly less severe in 2016, but all 3 images were graded as plus disease by the majority of graders in both 2007 and 2016. Images with average scores in the normal range (i.e., 1.00-1.49) in 2007 were on average graded 0.599 points higher in 2016, images with scores in the pre-plus range (i.e., 1.50-2.49) in 2007 were on average graded 0.616 points higher in 2016, and images with scores in the plus disease range (i.e., 2.50-3.00) were graded 0.090 points higher in 2016.

In 2007, the mean number of images classified as normal decreased from 8.6 (range 1-19), compared to 2.5 (range 0-14) in 2016 ($P < 0.01$). The mean number classified as pre-plus was 12.4 (range 2-22), compared to 11.2 (range 4-22) in 2016 ($P = 0.52$). The mean number classified as plus increased from 12.2 (range 3-17) in 2007 to 20.2 (range 5-32) in 2016 ($P < 0.01$). Thus, there was no evidence of an increase in diagnosis of pre-plus over time, though there was evidence to suggest that both pre-plus and plus disease were being classified at lower levels of disease severity. Representative images and responses are shown in Figure 2.

Ranking of image severity

Figure 3 graphically compares the severity ranking of images based on average score by all experts in 2016 vs. 2007. Comparison of the relative ranking of images for experts between 2007 and 2016 demonstrated good correlation (Spearman's $p=0.94$) indicating near perfect agreement on relative disease severity but the tendency towards more severe classification of disease in 2016.

Inter-Expert Agreement

The mean weighted κ statistics for each expert compared with all others in 2016 ($n=13$) are shown side-by-side with that of 2007 ($n=22$) in Table 1. For example, the mean weighted κ was between 0.21 and 0.40 (fair agreement) for 10 experts (69.2%) in 2016 vs. 7 experts (31.8%) in 2007, and between 0.41 and 0.60 (moderate agreement) for no experts in 2016 vs. 15 experts (68%) in 2007. This suggests that the 2016 expert cohort had, on average, lower inter-expert agreement compared to the 2007 expert cohort.

Subgroup Analysis for Repeat Experts in 2016 vs. 2007

Comparison of average scores for the 4 experts who were in both cohorts (i.e., those who participated in both 2016 and 2007) revealed that 18 images (52.9%) demonstrated an increase in grading of plus disease severity in 2016 compared to 2007, 8 (23.5%) had no change, and 8 (23.5%) decreased. Images considered normal by these 4 repeat experts in 2007 were graded on average 0.917 points higher in 2016, pre-plus images were 0.269 points higher in 2016, and plus disease images were 0.017 points lower. Comparison of the relative ranking of images for these experts between 2007 and 2016 demonstrated good correlation (Spearman's $p=0.87$) indicating good agreement on relative disease severity over time but the tendency towards more severe classification of disease in 2016.

Each repeat expert demonstrated statistically significant differences in their grading of plus disease severity in 2016 vs. 2007 (Table 2). Confusion matrices for each expert demonstrate that for 2 of the experts, 100% (11/11 and 22/22 for Expert 1 and 3, respectively) of grading differences were in the direction of a more severe classification in 2016. For Expert 2, 83% (10/12) were in the direction of more a severe classification in 2016. Expert 4, by contrast, had 100% (17/17) of grading differences in the direction of a less severe classification. Intra-expert agreement, as determined by weighted κ statistics, varied from poor to moderate among the four repeat experts.

DISCUSSION

This study was designed to examine and compare the grading of retinal vascular abnormality in wide-angle ROP images by experts in 2016 vs. 2007. The key findings from this study are: 1) experts diagnosed plus disease at an earlier relative stage of disease in 2016 compared to 2007, and 2) the ranking of the relative severity of ROP images has remained consistent over time.

The first key study finding is that plus disease was diagnosed at an earlier relative stage of disease by study experts in 2016 compared to 2007 (Figure 1). This raises questions about the enduring relevance of the multi-center CRYO-ROP and ETROP study findings, which were published in 1988 and 2003, respectively. There are

several possible explanations for this key finding. First among them may be the influence of studies showing better outcomes with earlier treatment.¹⁴ For example, the ETROP study defined guidelines to recommend treatment for type 1 or worse disease, compared to the more severe level of “threshold disease” which had been recommended after the CRYO-ROP study.^{1,2} Slidsborg et al previously suggested that an unintended shift in treatment indication following ETROP may have at least in part resulted in a documented increase in treatment-requiring ROP in their country.¹⁴ Second, the use of intravitreal anti-VEGF agents for ROP treatment has gained popularity since publication of the multi-center BEAT-ROP study findings in 2011.^{3,28} It is possible that the ease of administration of intravitreal anti-VEGF agents compared to laser treatment may influence the decision to treat although this hypothesis warrants additional study. Third, it is possible that concerns over medicolegal liability^{29,30} from poor visual outcomes, or other medical considerations may influence physicians to treat disease earlier.³¹

Our subgroup analysis of intra-expert agreement between 2016 vs. 2007 among repeat experts provides further insights. First, the group of repeat experts were no different from the larger group in that they too demonstrated a more aggressive pattern in their grading of plus disease with consistency in relative severity ranking over time (**Table 2**). Second, though 1 of the 4 experts trended in the opposite direction versus 2007 (a higher cut-point for pre-plus and plus disease), the findings are consistent with our recent publication suggesting that inter-observer (or in this case intra-observer) variability in plus disease diagnosis may be due to systematic bias related to the cut-points between classifications, rather than random disagreement or error.²³

The second key study finding is that experts have remained consistent in their ranking of ROP images based on relative plus disease severity between 2007 and 2016 (**Figure 3**). The ability to identify relative disease severity in ROP is vital clinically for determining disease progression, recognizing improvement after treatment, and determining intervals for follow-up examinations.^{16,34-36} In a separate study, we recently showed that inter-expert agreement in the ranking of relative plus disease severity among experts is significantly higher than inter-expert agreement on plus disease classification (i.e., determining normal, pre-plus, or plus).²⁴ Although numerous previous studies have demonstrated that there is discrepancy in plus disease classification among experts,^{7,15,16,18,19,33,37-39} these studies are the first to our knowledge that highlight better agreement in the ranking of relative disease severity. This current study adds to the previous by demonstrating high agreement on relative severity ranking over time (**Figure 3**), despite persistent fair or poor inter-expert agreement on disease classification (**Table 1**).

These results emphasize the need for innovative approaches to better standardize the process of diagnosis in ROP, and suggest that the development of an ROP severity scale that reflects a larger range of disease severity may improve agreement on disease severity and standardize treatment across providers. Computer-based image analysis represents one avenue for standardizing diagnosis and developing a severity scale for ROP. Specifically, computer-based image analysis could be used to calculate a quantitative scale representing relative disease severity and generate scores as an objective measurement of disease severity to aid in clinical decision-making. Such a score could be trended over time to track changes in the

retinal vasculature and any evolution of disease progression or regression, both of which could aid in the clinical appraisal of whether or not treatment is needed.^{16,34-36,40} An additional benefit to the development of computer-based image analysis tools for ROP would be the potential role in disease screening where a quantitative severity score could help to identify infants at moderate to high risk of progression to treatment requiring disease. This kind of tool would have particular significance for a telemedicine screening program to identify children needing either closer interval telemedical or ophthalmoscopic evaluation. With 90% of ROP cases never requiring treatment,² utilization of a computer-based image analysis system could help improve efficiency of resource use by strengthening the specificity of disease screening.

Several additional study limitations should be noted. First, this study is based on expert review of wide-angle digital images, which may have been less familiar to ROP experts in 2007 compared to 2016. This difference in magnification, field of view, and perspective may have been a cause of confusion for ophthalmologists.^{32,33} However, virtually all study experts (77%) in 2007 reported that they had experience with interpretation of wide-angle RetCam, images for ROP, with 11 participants (50%) describing their previous experience as “extensive.”⁷ Second, it may be that image-based classifications do not directly correspond with bedside ophthalmoscopic evaluations. However, we studied this issue both in 2008,⁴¹ and again in 2016,⁴² showing very high levels of agreement between ophthalmoscopic and image-based diagnoses. In particular, we recently published a series of 1553 examinations with multiple image-based and ophthalmoscopic classifications and found no systematic differences between examination techniques, but significant variability between inter-expert image classifications on stage, zone, and plus disease classification. Third, any visible peripheral ROP was cropped out of study images in 2007,⁷ whereas this was not done in 2016. In principle, this difference should not affect clinical diagnosis because plus disease is defined based on findings only in the central retina.¹ That said, numerous studies have shown that a variety of features beyond this standard definition may be considered by physicians while diagnosing plus disease.^{12,33,38,43} To address this potential confounder for 12 of the 34 images, 3 experts were asked to re-classify all 34 of the original 2007 cropped images. The average score was not significantly different from their original 2016 gradings (2.60 vs. 2.67); however it was significantly more severe than the original 2007 scores (2.1). Thus, at least for these 3 experts, there were no differences in image scores related to the slight difference in presentation. Fourth, image reading conditions were not standardized for practical reasons. Although the effect of variables such as luminance and resolution of computer monitor displays has been characterized,⁴⁴ the extent to which any of these factors may have influenced our results are unclear. Finally, the extent to which our finding can be generalized to the entire field of practicing ROP ophthalmologists is unknown. Subjects in the 2016 and 2007 expert cohorts were limited to experts with the most extensive clinical and research experience to optimize the credibility of our findings. We find it reasonable to believe that there may likely be as much or more variation within the entire community of ROP physicians.

In summary, our results suggest that experts may be diagnosing plus disease at an earlier stage of disease in 2016 compared to 2007, although ranking of relative disease severity has remained constant. This has implications for patient care,

research, and teaching, and additional studies are needed to better understand possible causes for this apparently evolving diagnostic trend. Computer-based image analysis and the development of a more continuous severity scale for plus disease may represent future avenues to improve diagnostic agreement between examining clinicians and track relative disease severity over time, and ultimately improve the consistency and quality of ROP care worldwide.

1
2
3
4 ACKNOWLEDGEMENTS/DISCLOSURES:
5
6

7 Funding/Support: Jayashree Kalpathy-Cramer, Susan Ostmo, RV Paul Chan, and
8 Michael F. Chiang are supported by NIH grant EY19474 from the National Institutes of
9 Health, Bethesda, MD. J. Peter Campbell, Susan Ostmo, and Michael F. Chiang are
10 supported by NIH grant EY010572 and P30 EY010572 from the National Institutes of
11 Health, Bethesda, MD. Jayashree Kalpathy-Cramer and Michael F. Chiang are
12 supported by NIH grant EY022387 from the National Institutes of Health, Bethesda, MD.
13 Chace Moleta, J. Peter Campbell, Susan Ostmo RV Paul Chan, and Michael F. Chiang,
14 are supported by unrestricted departmental funding from Research to Prevent
15 Blindness, New York, NY. RV Paul Chan is supported by the iNsight Foundation, New
16 York, NY. J. Peter Campbell, Jayashree Kalpathy-Cramer, RV Paul Chan, Susan
17 Ostmo, and Michael F. Chiang are supported by grant 1622679 from the National
18 Science Foundation, Arlington, VA.
19
20
21

22 Financial Disclosures: Michael F. Chiang is an unpaid member of the Scientific Advisory
23 Board for Clarity Medical Systems (Pleasanton, CA) and a consultant for Novartis
24 (Basel, Switzerland). RVPC is a consultant for Visunex Medical Systems (Fremont, CA).
25 The following authors report no financial disclosures nor conflicts of interest in medicine:
26 Chace Moleta, J. Peter Campbell, Jayashree Kalpathy-Cramer, Susan Ostmo, and
27 Karyn Jonas.
28
29
30

31 Acknowledgments such as Statisticians, Medical Writers, Expert contributions. There
32 are no such additional acknowledgements.
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

REFERENCES

1. Cryotherapy for Retinopathy of Prematurity Cooperative Group. Multicenter trial of cryotherapy for retinopathy of prematurity. Preliminary results. *Arch Ophthalmol* 1988;106(4):471-479.
2. 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(12):1684-1694.
3. Mintz-Hittner HA, Kennedy KA, Chuang AZ. Efficacy of intravitreal bevacizumab for stage 3+ retinopathy of prematurity. *NEJM* 2011;364(7):603-615.
4. Sommer A, Taylor HR, Ravilla TD, et al. Challenges of ophthalmic care in the developing world. *JAMA Ophthalmol* 2014;132(5):640-644.
5. International Committee for the Classification of Retinopathy of Prematurity. The International Classification of Retinopathy of Prematurity revisited. *Arch Ophthalmol* 2005;123(7):991-999.
6. The Committee for the Classification of Retinopathy of Prematurity. An international classification of retinopathy of prematurity. *Arch Ophthalmol* 1984;102(8):1130-1134.
7. Chiang MF, Jiang L, Gelman R, Du YE, Flynn JT. Interexpert agreement of plus disease diagnosis in retinopathy of prematurity. *Arch Ophthalmol* 2007;125(7):875-880.
8. 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(4):352-356.
9. Chiang MF, Gelman R, Jiang L, Martinez-Perez ME, Du YE, Flynn JT. Plus disease in retinopathy of prematurity: an analysis of diagnostic performance. *Trans Am Ophthalmol Soc* 2007;105:73-84; discussion 84-75.
10. Gelman R, Jiang L, Du YE, Martinez-Perez ME, Flynn JT, Chiang MF. Plus disease in retinopathy of prematurity: pilot study of computer-based and expert diagnosis. *J AAPOS* 2007;11(6):532-540.
11. Gschliesser A, Stifter E, Neumayer T, et al. Inter-expert and intra-expert agreement on the diagnosis and treatment of retinopathy of prematurity. *Am J Ophthalmol* 2015;160(3):553-560.
12. Hewing NJ, Kaufman DR, Chan RV, Chiang MF. Plus disease in retinopathy of prematurity: qualitative analysis of diagnostic process by experts. *JAMA Ophthalmol* 2013;131(8):1026-1032.
13. 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(11):1470-1476.
14. 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(4):549-553.
15. Paul Chan RV, Williams SL, Yonekawa Y, Weissgold DJ, Lee TC, Chiang MF. Accuracy of retinopathy of prematurity diagnosis by retinal fellows. *Retina* 2010;30(6):958-965.

16. 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(6):573-578.
17. Wallace DK. Fellowship training in retinopathy of prematurity. *J AAPOS* 2012;16(1):1.
18. Wong RK, Ventura CV, Espiritu MJ, et al. Training fellows for retinopathy of prematurity care: a Web-based survey. *J AAPOS* 2012;16(2):177-181.
19. Nagiel A, Espiritu MJ, Wong RK, et al. Retinopathy of prematurity residency training. *Ophthalmol* 2012;119(12):2644-2645
20. Lien R, Yu MH, Hsu KH, et al. Neurodevelopmental Outcomes in Infants with Retinopathy of Prematurity and Bevacizumab Treatment. *PloS One*. 2016;11(1):e0148019.
21. Morin J, Luu TM, Superstein R, et al. Neurodevelopmental Outcomes Following Bevacizumab Injections for Retinopathy of Prematurity. *Pediatrics* 2016;137(4): e20153218.
22. Quinn GE, Darlow BA. Concerns for Development After Bevacizumab Treatment of ROP. *Pediatrics*. 2016;137(4):e20160057.
23. 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. *Ophthalmol* 2016;123(11):2338-2344..
24. Kalpathy-Cramer J, Campbell JP, Erdogmus D, et al. Plus disease in retinopathy of prematurity: improving diagnosis by ranking disease severity and using quantitative image analysis. *Ophthalmol* 2016;123(11):2345-2351.
25. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33(1):159-174.
26. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological bulletin*. 1968;70(4):213-220.
27. Darlow BA, Ells AL, Gilbert CE, Gole GA, Quinn GE. Are we there yet? Bevacizumab therapy for retinopathy of prematurity. *Arch Dis Child Fetal Neonatal Ed* 2013;98(2):F170-174.
28. Tawse KL, Jeng-Miller KW, Bauman CR. Current Practice Patterns for Treatment of Retinopathy of Prematurity. *OSLI* 2016;47(5):491-495.
29. Day S, Menke AM, Abbott RL. Retinopathy of prematurity malpractice claims: the Ophthalmic Mutual Insurance Company experience. *Arch Ophthalmol* 2009;127(6):794-798.
30. Reynolds JD. Malpractice and the quality of care in retinopathy of prematurity (an American Ophthalmological Society thesis). *Trans Am Ophthalmol Soc* 2007;105:461-480.
31. Gupta MP, Chan RV, Anzures R, Ostmo S, Jonas K, Chiang MF. Practice Patterns in Retinopathy of Prematurity Treatment for Disease Milder Than Recommended by Guidelines. *Am J Ophthalmol* 2016;163:1-10.
32. Gelman SK, Gelman R, Callahan AB, et al. Plus disease in retinopathy of prematurity: quantitative analysis of standard published photograph. *Arch Ophthalmol* 2010;128(9):1217-1220.
33. Rao R, Jonsson NJ, Ventura C, et al. Plus disease in retinopathy of prematurity: diagnostic impact of field of view. *Retina* 2012;32(6):1148-1155.

34. Wallace DK, Freedman SF, Hartnett ME, Quinn GE. Predictive value of pre-plus disease in retinopathy of prematurity. *Arch Ophthalmol* 2011;129(5):591-596.
35. Thyparampil PJ, Park Y, Martinez-Perez ME, et al. Plus disease in retinopathy of prematurity: quantitative analysis of vascular change. *Am J Ophthalmol* 2010;150(4):468-475
36. Wallace DK, Kylstra JA, Chesnutt DA. Prognostic significance of vascular dilation and tortuosity insufficient for plus disease in retinopathy of prematurity. *J AAPOS* 2000;4(4):224-229.
37. 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;134(6):651-657.
38. Keck KM, Kalpathy-Cramer J, Ataer-Cansizoglu E, You S, Erdogmus D, Chiang MF. Plus disease diagnosis in retinopathy of prematurity: vascular tortuosity as a function of distance from optic disk. *Retina* 2013;33(8):1700-1707.
39. 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(3):351-356.
40. Myung JS, Gelman R, Aaker GD, Radcliffe NM, Chan RV, Chiang MF. Evaluation of vascular disease progression in retinopathy of prematurity using static and dynamic retinal images. *Am J Ophthalmol* 2012;153(3):544-551.
41. Scott KE, Kim DY, Wang L, et al. Telemedical diagnosis of retinopathy of prematurity intraphysician agreement between ophthalmoscopic examination and image-based interpretation. *Ophthalmol* 2008;115(7):1222-1228.
42. Campbell JP, Ryan MC, Lore E, et al. Diagnostic Discrepancies in Retinopathy of Prematurity Classification. *Ophthalmol* 2016;Aug;123(8):1795-801
43. Ataer-Cansizoglu E, Kalpathy-Cramer J, You S, Keck K, Erdogmus D, Chiang MF. Analysis of underlying causes of inter-expert disagreement in retinopathy of prematurity diagnosis. Application of machine learning principles. *Methods Inf Med* 2015;54(1):93-102.
44. Herron JM, Bender TM, Campbell WL, Sumkin JH, Rockette HE, Gur D. Effects of luminance and resolution on observer performance with chest radiographs. *Radiology* 2000;215(1):169-174.

FIGURE CAPTIONS

Figure 1. Comparison of average scores for plus disease severity in 34 wide-angle retinopathy of prematurity images graded in 2016 (13 experts) versus 2007 (22 experts). Experts assigned each image a score of “1” for Normal, “2” for Pre-plus, and “3” for plus disease. Straight line indicates where equivalent average scores would fall.

Figure 2. Representative examples of retinopathy of prematurity images diagnosed by 13 experts in 2016 (vs. 22 experts in 2007). (Left) was diagnosed as normal by 0% in 2016 (vs. 54.5% in 2007), pre-plus by 46% in 2016 (vs. 41% in 2007), and plus by 54% in 2016 (vs. 4.5% in 2007). (Center) was diagnosed as normal by 0% in 2016 (vs. 45% in 2007), pre-plus by 62% in 2016 (vs. 50% in 2007), and plus by 38% in 2016 (vs. 5% in 2007). (Right) was diagnosed as normal by 45% in 2016 (vs. 50% in 2007), pre-plus by 27% in 2016 (vs. 45% in 2007), and plus by 27% in 2016 (vs. 5% in 2007).

Figure 3. Comparison of 34 wide-angle retinopathy of prematurity images ranked from least to most severe for plus disease severity in 2016 (13 experts) versus 2007 (22 experts). Ranking is based on average image score calculated after experts assigned each image a score of “1” for normal, “2” for pre-plus, and “3” for plus disease. Spearman’s ρ is 0.94,

Table 1. Comparison of inter-expert agreement in retinopathy of prematurity plus disease grading for all experts in 2016 (n=13) versus 2007 (n=22), based on mean weighted kappa values for each expert compared with all others. Each expert is categorized into a group ("poor," "fair," "moderate," or "substantial" agreement).

	2007 (n=22) # (%)	2016 (n=13) # (%)
Poor Agreement (0.00-0.20)	0 (0)	3 (15)
Fair Agreement (0.21-0.40)	7 (32)	10 (69)
Moderate Agreement (0.41-0.60)	15 (68)	0 (0%)
Substantial Agreement (0.61-0.80)	0 (0)	0 (0%)

Table 2

Table 2: Confusion matrices for intra-expert agreement in retinopathy of prematurity plus disease grading of 34 wide-angle images by 4 repeat experts, who participated in both 2016 and 2007. For each repeat expert, the p-value for plus disease grading between 2016 vs. 2007 is shown, followed by the weighted κ statistic for intra-expert agreement between 2016 vs. 2007.

		2016 Diagnosis		
		Normal	Pre-plus	Plus
Expert 1: P = 0.002, κ = 0.270				
2007 Diagnosis	Normal	0	1	2
	Pre-Plus	0	3	8
	Plus	0	0	20
Expert 2: P = 0.021, κ = 0.478				
2007 Diagnosis	Normal	4	9	0
	Pre-Plus	2	8	1
	Plus	0	0	10
Expert 3: P = <0.001, κ = 0.094				
2007 Diagnosis	Normal	0	3	9
	Pre-Plus	0	3	10
	Plus	0	0	9
Expert 4: P = <0.001, κ = 0.285				
2007 Diagnosis	Normal	7	0	0
	Pre-Plus	7	4	0
	Plus	0	10	6

Figure 1

Figure 2

Figure 3