Interexpert Agreement of Plus Disease Diagnosis in Retinopathy of Prematurity

Michael F. Chiang, MD, MA; Lei Jiang, , BA; Rony Gelman, MD; Yunling E. Du, PhD; John T. Flynn, MD

Objective: To measure agreement of plus disease diagnosis among retinopathy of prematurity (ROP) experts.

Methods: A set of 34 wide-angle retinal photographs from infants with ROP was compiled on a secure Web site and was interpreted independently by 22 recognized ROP experts. Diagnostic agreement was analyzed using 3-level (plus, pre-plus, or neither) and 2-level (plus or not plus) categorizations.

Results: In the 3-level categorization, all experts agreed on the same diagnosis in 4 of 34 images (12%), and the mean weighted κ statistic for each expert compared with all others was between 0.21 and 0.40 (fair agreement) for 7 experts (32%) and between 0.41 and 0.60 (moderate agreement) for 15 experts (68%). In the 2-level categorization, all experts who provided a diagnosis agreed in 7 of 34 images (21%), and the mean κ statistic for each expert compared with all others was between 0 and 0.20 (slight agreement) for 1 expert (5%), between 0.21 and 0.40 (fair agreement) for 3 experts (14%), between 0.41 and 0.60 (moderate agreement) for 12 experts (55%), and between 0.61 and 0.80 (substantial agreement) for 6 experts (27%).

Conclusions: Interexpert agreement of plus disease diagnosis is imperfect. This may have important implications for clinical ROP management, continued refinement of the international ROP classification system, development of computer-based diagnostic algorithms, and implementation of ROP telemedicine systems.

Arch Ophthalmol. 2007;125(7):875-880

Author Affiliations:
Departments of Ophthalmology (Drs Chiang, Gelman, and Flynn and Mr Jiang) and Biomedical Informatics (Dr Chiang), Columbia University College of Physicians and Surgeons, and Department of Epidemiology and Population Health, Albert Einstein College of Medicine (Dr Du), New York, New York.

RETINOPATHY OF PREMATURITY (ROP) is a vasoproliferative disease affecting low-birth-weight infants. Several major advances in ROP diagnosis and treatment have occurred during the past 2 decades. The international classification of ROP established a standard system for describing the location, severity, and extent of retinal findings on ophthalmoscopy,[1,2] thereby creating a foundation for clinical care and research. The Cryotherapy for Retinopathy of Prematurity (CRYO-ROP) and Early Treatment for Retinopathy of Prematurity (ETROP) trials determined specific disease criteria for which treatment with cryotherapy or laser photocoagulation has been shown to improve structural and functional outcomes.[3-5] Despite these advances, ROP continues to be a leading cause of childhood blindness throughout the world.[6-8]

For editorial comment see page 963

Plus disease is a key aspect of ROP evaluation and is characterized by retinal vascular dilation and tortuosity. The presence of plus disease is a necessary feature of threshold disease and is a sufficient feature for type 1 ROP, both of which have been shown to warrant prompt treatment.[3,4] The minimum level of vascular abnormality needed for plus disease is represented by a standard photograph, which was selected by expert consensus.[3] More recently, the following 2 modifications have been made: (1) multicenter clinical trials have explicitly stated that at least 2 quadrants of this minimum amount of vascular dilatation and tortuosity are sufficient for diagnosis of plus disease[9,10] and (2) the 2005 revised international classification of ROP formally defined an intermediate pre-plus condition as "abnormalities of the posterior pole that are insufficient for the diagnosis of plus disease but that demonstrate more arterial tortuosity and more venous dilation than normal."[10(p995)]

These definitions of plus disease and pre-plus disease may be subjective in nature because they are based on photographic standards and descriptive qualifiers rather than on quantifiable measurements. Although dilated binocular indirect ophthalmoscopy is considered the criterion standard for ROP diagnosis and classification,[10] agreement among multiple examiners in determin-

ing the presence or absence of plus disease has not been systematically studied, to our knowledge. Because inconsistent diagnosis of plus disease can lead to errors in overtreatment or undertreatment, this question has important implications for clinical management and research in ROP.

The objective of this study was to measure agreement of plus disease diagnosis among a group of 22 recognized ROP experts. To simulate real-world scenarios, expert participants used a secure Web site to review a set of retinal images captured by a commercially available wide-angle digital fundus camera. This study focuses specifically on interexpert agreement rather than on diagnostic accuracy compared with a reference standard.

METHODS

This study was approved by the Institutional Review Board at Columbia University Medical Center, New York, New York. Waiver of consent was obtained for use of deidentified retinal images.

STUDY PARTICIPANTS

A set of 34 digital retinal images was captured from premature infants using a commercially available device (RetCam II; Clarity Medical Systems, Pleasanton, California) during routine ROP care. Images were selected that in our opinion reflected a change in vasculature compared with baseline. Each photograph displayed the posterior pole of the retina, with any visible peripheral disease cropped out. Images were not annotated with any descriptive information such as name, birth weight, or gestational age. No images were repeated.

A group of ROP experts was invited to participate in the study. For the objective of this study, eligible experts were defined as practicing pediatric ophthalmologists or retina specialists who met at least 1 of the following 3 criteria: having been a study center principal investigator for the CRYO-ROP or ETROP study, having been a certified investigator for either of those studies, or having coauthored at least 5 peer-reviewed ROP manuscripts.

IMAGE INTERPRETATION

Each participant was provided with an anonymous study identifier and a password to a Web-based program developed by us to display photographs. Experts were asked to categorize each image based on the following 2 axes: (1) diagnosis (plus, pre-plus, neither, or cannot determine) and (2) quality (adequate, possibly adequate, or inadequate for diagnosis). Options in each categorization axis were mutually exclusive. Participants were allowed to revise answers by returning to previous pages in the Web-based program, but all responses were finalized and logged into a secure database (SQL 2005; Microsoft, Redmond, Washington) after the final image was categorized. Experts were asked whether their home institution had a RetCam device and whether they would describe their experience with interpreting RetCam-captured images as extensive, limited, or none.

Participants were not provided with any references or standards for plus and pre-plus disease, although it was assumed that they would be familiar with these definitions. Instead, the decision was made to rely on experts' personal experience and judgment to better simulate a real-world situation. Informed consent was obtained from each participant using a click-through Web form before images were displayed. Specific instructions included the following statement:

> We are inviting you to participate in a research study involving ROP because we believe that you are recognized as an international authority on this disease. The purpose of this study is to determine the inter-rater agreement among ROP experts for detection of plus disease from retinal photographs. We believe that there may be subjectivity in the determination of plus disease. We feel that it is important to study the inter-rater agreement for detection of plus disease because ROP diagnosis and treatment guidelines are heavily dependent on the presence or absence of plus disease. If you agree to participate in this study, you will be shown a series of approximately 30 retinal posterior pole images. You will be asked to classify each image as either "plus," "pre-plus," "neither," or "cannot determine."

DATA ANALYSIS

Data were exported from the study database into a spreadsheet (Excel 2003; Microsoft). Absolute agreement among participants was determined for each image based on 3-level (plus, pre-plus, or neither) and 2-level (plus or not plus) categorizations. Images classified as cannot determine were excluded from analysis for that participant.

In the 2-level categorization, the κ statistic was used to measure chance-adjusted agreement for the presence of plus disease between each pair of experts. In the 3-level categorization, the weighted κ statistic was used for analysis of agreement because it adjusts for small vs large disagreements.[11] An accepted scale was used to interpret results as follows: 0 to 0.20 indicated 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.[12] The mean κ values were determined for each expert compared with all others, and standard errors were calculated using the jackknife method.

Statistical analysis was performed using SPSS version 14.0 software (SPSS Inc, Chicago, Illinois). The relationship between expert characteristics and mean κ statistics compared with other examiners was analyzed using the 1-sample or independent sample t test, as appropriate. Statistical significance was defined as 2-sided $P<.05$.

RESULTS

CHARACTERISTICS OF EXPERT PARTICIPANTS

Based on the study definition of expertise that was used, 29 individuals were invited, of whom 22 (76%) consented to participate. Among these 22 experts, 18 (82%) had served as principal investigators in the CRYO-ROP or ETROP study, 4 (18%) had served as certified investigators in either study, and 11 (50%) had coauthored at least 5 peer-reviewed ROP manuscripts. Nine participants (41%) met more than 1 criteria.

Among 22 participants, 16 (73%) worked at institutions with a RetCam device, 5 (23%) worked at institutions without the device, and 1 (5%) did not report this information. Similarly, 11 participants (50%) described their previous experience with interpretation of RetCam-captured images as extensive, 6 (27%) described it as limited, 4 (18%) described it as none, and 1 (5%) did not report this information. Seventeen participants (77%) were pediatric ophthalmologists, whereas 5 participants (23%) were retina specialists.

PARTICIPANT RESPONSES

All 34 images were reviewed by 22 experts, for a total of 748 diagnosis and quality responses. A diagnosis of cannot determine was made in 18 of 748 cases (2%). Among 748 cases, image quality was scored as adequate in 656 cases (88%), possibly adequate in 72 cases (10%), and inadequate for diagnosis in 20 cases (3%).

Overall diagnostic responses are summarized in the **Table**. Three of 34 images (9%) were classified as plus by all 22 experts. In the 3-level categorization, 1 image (3%) was classified as neither plus nor pre-plus by all 22 experts, and no images were classified as pre-plus by all 22 experts. In the 2-level classification, 4 of 34 images (12%) were classified as not plus by all experts who provided a diagnosis. Representative images and responses are shown in **Figure 1**.

INTEREXPERT AGREEMENT

Figure 2 shows absolute agreement in plus disease diagnosis, based on the percentage of experts who assigned the same diagnosis to each image. For example, the same 3-level diagnosis was made by at least 90% of experts in 6 images (18%) and by at least 80% of experts in 7 images (21%). The same 2-level diagnosis was made by at least 90% of experts in 20 images (59%) and by at least 80% of experts in 24 images (71%).

The mean κ statistics for each expert compared with all others are shown in **Figure 3**. In the 3-level categorization, the mean weighted κ statistic for each expert compared with all others was between 0.21 and 0.40 (fair agreement) for 7 experts (32%) and between 0.41 and 0.60 (moderate agreement) for 15 experts (68%). In the 2-level categorization, the mean κ statistic for each expert compared with all others was between 0 and 0.20 (slight agreement) for 1 expert (5%), between 0.21 and 0.40 (fair agreement) for 3 experts (14%), between 0.41 and 0.60 (moderate agreement) for 12 experts (55%), and between 0.61 and 0.80 (substantial agreement) for 6 experts (27%).

There were no statistically significant differences in mean κ or weighted κ statistics based on the following expert characteristics: working in vs not working in an institution with a RetCam, having published at least 5 vs fewer than 5 peer-reviewed ROP manuscripts, type of ophthalmologist (pediatric vs retina specialist), self-reported level of experience interpreting RetCam images (extensive, limited, or none), status as a principal investigator vs not a principal investigator in the CRYO-ROP or ETROP study, or status as a certified investigator vs not a certified investigator in either of those studies.

Table. Absolute Agreement in Plus Disease Diagnosis Among 22 Experts Reviewing 34 Images[a]

	No. (%)				
	3-Level Categorization by 22 Experts			2-Level Categorization by 22 Experts	
Image	Plus	Pre-plus	Neither	Plus	Not Plus
1	3 (14)	15 (68)	4 (18)	3 (14)	19 (86)
2	1 (5)	16 (76)	4 (19)	1 (5)	20 (95)
3	14 (70)	6 (30)	0	14 (70)	6 (30)
4	5 (24)	12 (57)	4 (19)	5 (24)	16 (76)
5	3 (14)	9 (43)	9 (43)	3 (14)	18 (86)
6	22 (100)	0	0	22 (100)	0
7	1 (5)	9 (41)	12 (55)	1 (5)	21 (96)
8	21 (96)	1 (5)	0	21 (96)	1 (5)
9	0	9 (43)	12 (57)	0	21 (100)
10	0	0	22 (100)	0	22 (100)
11	22 (100)	0	0	22 (100)	0
12	1 (5)	11 (50)	10 (46)	1 (5)	21 (96)
13	7 (32)	15 (68)	0	7 (32)	15 (68)
14	2 (10)	11 (52)	8 (38)	2 (10)	19 (90)
15	12 (60)	8 (40)	0	12 (60)	8 (40)
16	1 (5)	10 (48)	10 (48)	1 (5)	20 (95)
17	8 (38)	11 (52)	2 (10)	8 (38)	13 (62)
18	1 (5)	10 (46)	11 (50)	1 (5)	21 (96)
19	2 (10)	14 (67)	5 (24)	2 (10)	19 (90)
20	20 (95)	1 (5)	0	20 (95)	1 (5)
21	0	8 (38)	13 (62)	0	21 (100)
22	11 (52)	10 (48)	0	11 (52)	10 (48)
23	17 (77)	5 (23)	0	17 (77)	5 (23)
24	0	5 (23)	17 (77)	0	22 (100)
25	2 (10)	9 (43)	10 (48)	2 (10)	19 (90)
26	16 (73)	6 (27)	0	16 (73)	6 (27)
27	1 (5)	8 (36)	13 (59)	1 (5)	21 (96)
28	14 (64)	8 (36)	0	14 (64)	8 (36)
29	1 (5)	15 (71)	5 (24)	1 (5)	20 (95)
30	17 (81)	4 (19)	0	17 (81)	4 (19)
31	1 (5)	8 (36)	13 (59)	1 (5)	21 (96)
32	3 (14)	14 (64)	5 (23)	3 (14)	19 (86)
33	17 (77)	5 (23)	0	17 (77)	5 (23)
34	22 (100)	0	0	22 (100)	0

[a] Number of images in each row may not add to 22 because images categorized as cannot determine were excluded for that expert.

COMMENT

This is the first study (to our knowledge) that has systematically evaluated agreement among ROP experts for plus disease diagnosis. Consistent and accurate detection of plus disease has an increasingly critical role in the identification of treatment-requiring ROP. This is particularly relevant because the multicenter ETROP trial recently determined that presence of plus disease is sufficient for meeting the definition of type 1 ROP, which benefits from early treatment regardless of the exact number of clock hours of peripheral disease.[4]

The main finding from this study is that interexpert agreement of plus disease diagnosis is imperfect. Using a 3-level categorization, all 22 experts agreed on the same diagnosis in 4 of 34 images (12%) (Figure 2), and the mean weighted κ statistic for each expert compared with all others ranged from 0.25 (fair agreement) to 0.55 (moderate agreement) (Figure 3). Using a 2-level categorization, all experts who provided a diagnosis agreed in 7 of 34 images (21%) (Figure 2), and the mean κ statistic for each expert compared with all others ranged from 0.19 (slight agreement) to 0.66 (substantial agreement) (Figure 3). This degree of variability suggests that image-based plus disease diagnosis may be heavily subjective.

Binocular indirect ophthalmoscopy by an experienced ophthalmologist is considered the criterion standard for diagnosis and classification of ROP.[10] By defi-

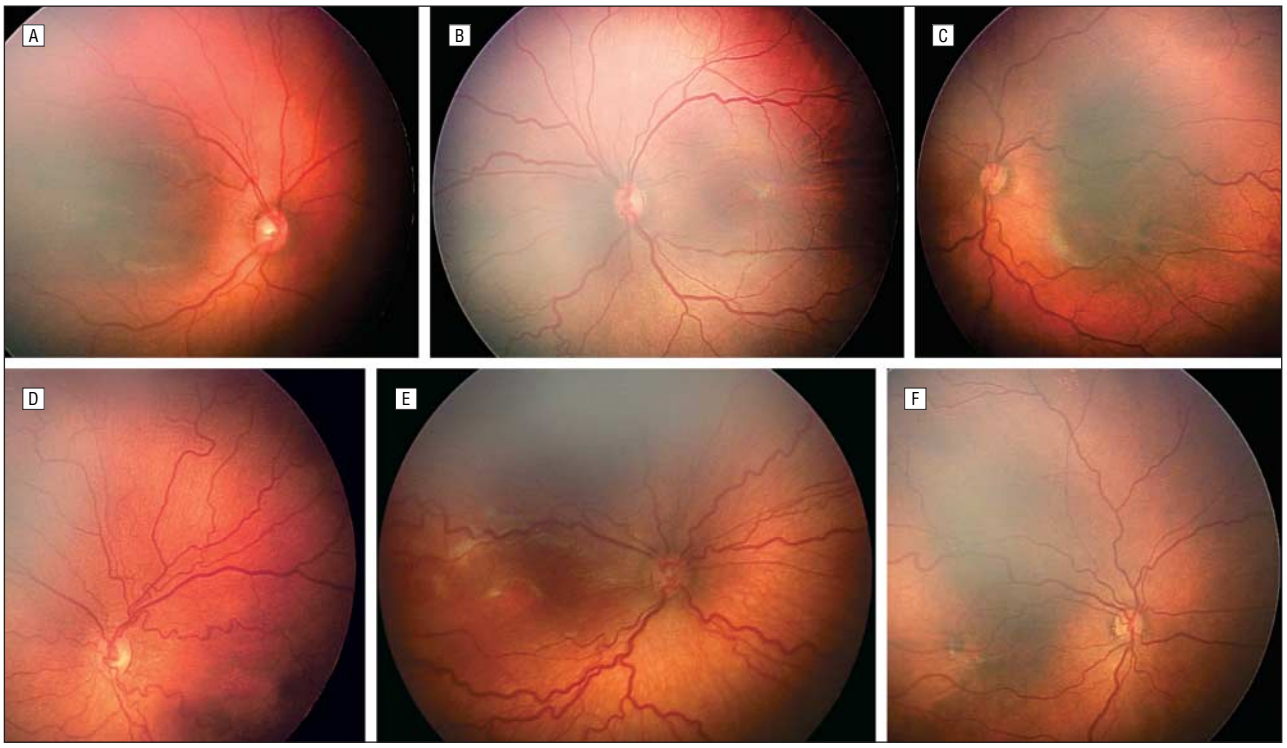

Figure 1. Representative images shown to 22 expert participants. A, Classified as neither plus nor pre-plus by all 22 experts (100%). B, Classified as plus by 2 experts (9%), pre-plus by 9 experts (41%), neither by 10 experts (46%), and cannot determine by 1 expert (5%). C, Classified as plus by 1 expert (5%), pre-plus by 16 experts (73%), neither by 4 experts (18%), and cannot determine by 1 expert (5%). D, Classified as plus by 11 experts (50%), pre-plus by 10 experts (46%), and cannot determine by 1 expert (5%). E, Classified as plus by all 22 experts (100%). F, Classified as plus by 3 experts (14%), pre-plus by 9 experts (41%), neither by 9 experts (41%), and cannot determine by 1 expert (5%).

nition, a criterion standard test should have complete accuracy and consensus.[13] The extent of disagreement in plus disease diagnosis among recognized ROP authorities in this study raises important questions about the reliability of this standard. This inconsistency presumably results from subjective differences in judgment among experts even while viewing the same images or from varying interpretations of the definition of plus disease.[3] Several studies[14-16] explored the possibility of plus disease detection using computer-based image analysis. If these automated systems can be shown to have accuracy comparable to that of human experts, the objectivity and reproducibility of computer-based techniques may offer important advantages over current methods. We emphasize that the objective of this study was to evaluate agreement in diagnosis among experts rather than to measure accuracy compared with a criterion standard. The opinion of any single participant in this study would certainly be regarded as a criterion standard for diagnosis, although we note that a criterion standard test cannot be completely accurate if it is not reproducible among multiple observers.

The recently revised international classification of ROP introduced an intermediate condition of pre-plus disease.[10] The clinical significance of pre-plus disease is not completely clear given that it was not incorporated into clinical trials such as CRYO-ROP or ETROP. Results from our current study suggest that even experts do not have complete agreement about whether a given image represents pre-plus as opposed to plus or neither (Table). If the clinical usefulness of pre-plus disease can be demonstrated,[17] future development of a more precise definition of this entity may help guide physicians in diagnosis.

The design of a study involving interexpert agreement requires an explicit definition of expertise, and the method used for this project warrants some explanation. Participants were invited for this study based on academic criteria, as evidenced by leadership roles in major multicenter clinical trials or by authorship of peer-reviewed ROP literature. This may not necessarily reflect clinical expertise in a real-world setting. However, there are numerous factors comprising medical expertise, some of which may be difficult to quantify for the purpose of study design.[18] Furthermore, it could be argued that academic ROP experts may have greater familiarity with the published photographic standard for plus disease than the overall population of ophthalmologists who perform ROP examinations. Therefore, we hypothesize that disagreement in plus disease diagnosis within the overall population of practicing clinicians may be higher than that among the academic experts in this study. This issue may warrant further study to determine the extent to which these findings are generalizable.

From a clinical perspective, it would be most useful to know the agreement of plus disease diagnosis among multiple experts performing serial indirect ophthalmoscopy on the same infant. However, that type of study would be impractical because of infant safety concerns.[19] To simulate a real-world situation for this study, images presented to participants were captured using a commercially available RetCam device. This is a contact camera with a 130° field of view and is the most well-known in-

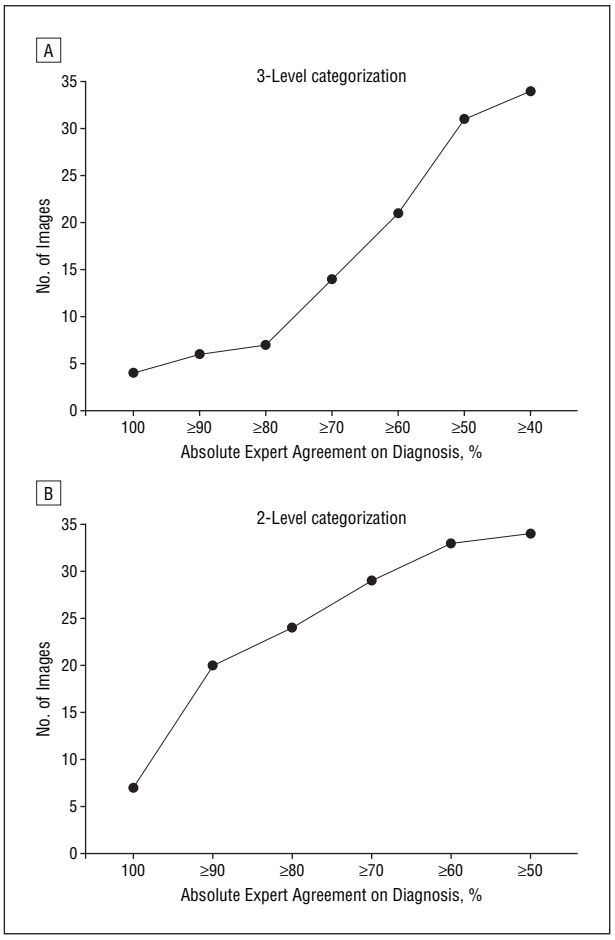

Figure 2. Absolute agreement in plus disease diagnosis, based on the percentage of 22 experts who assigned the same diagnosis to each of 34 images. A, Agreement in the 3-level categorization (plus, pre-plus, or neither). B, Agreement in the 2-level categorization (plus or not plus).

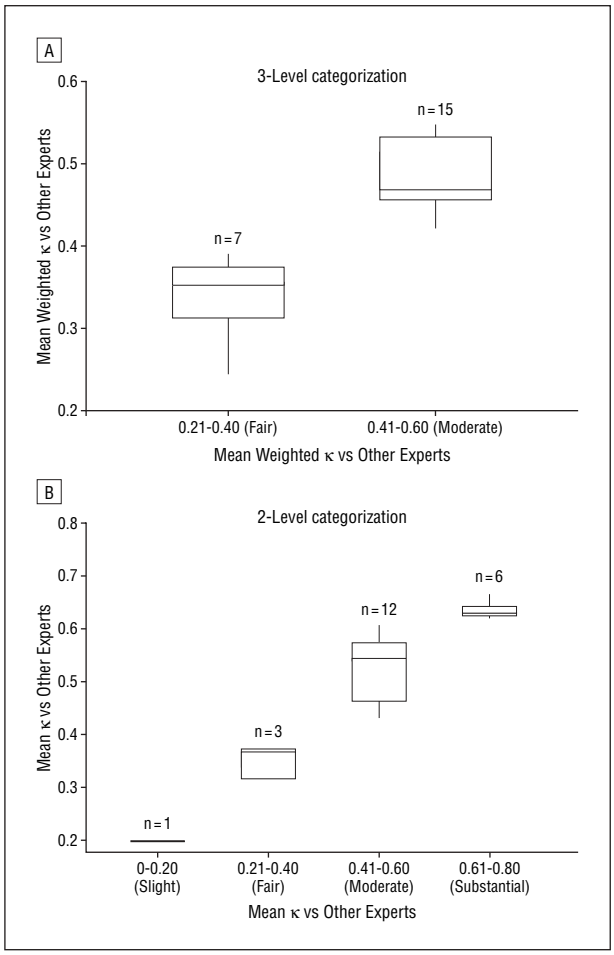

Figure 3. Agreement in plus disease diagnosis, based on box plots of the mean κ statistic for each of 22 experts compared with all others. A, The mean weighted κ statistic in the 3-level categorization (plus, pre-plus, or neither). B, The mean κ statistic in the 2-level categorization (plus or not plus). Boxes represent the 25th, 50th, and 75th percentile κ values. Whiskers represent the 10th and 90th percentile values.

strument for pediatric retinal imaging.[20-25] In contrast, standard binocular indirect ophthalmoscopy provides a 40° to 50° field of view. It is conceivable that this difference in perspective may have caused difficulty for participants, depending on their previous experience interpreting wide-angle ROP photographs. Although this study did not detect any correlation between mean κ statistics and self-reported level of RetCam experience, this question may deserve additional study with a broader spectrum of image graders. On one hand, limited experience in correlating wide-angle images with indirect ophthalmoscopy might result in systematic overdiagnosis or underdiagnosis of plus disease by some participants, thereby increasing variability. On the other hand, the fact that all participants were asked to review the exact same images in this study might produce decreased variability compared with serial ophthalmoscopy because examination quality may vary based on infant stability or cooperation.

Telemedicine strategies have been proposed as an alternative to standard ROP care involving dilated examination at the neonatal intensive care unit bedside. Findings from several limited studies[20-25] suggest that remote interpretation of retinal images may have adequate sensitivity and specificity to identify clinically significant ROP. To our knowledge, these published studies have compared accuracy of remote interpretation of RetCam images captured by ophthalmic personnel with a criterion standard of dilated ophthalmoscopy by a single examiner. The present study reveals several clinically significant disagreements among acknowledged ROP experts in plus disease diagnosis from wide-angle retinal photographs. To prevent diagnostic errors, this issue should be examined and resolved before the routine deployment of ROP telemedicine systems. However, if implemented properly, remote image interpretation at certified centers may offer advantages over dilated examination by a single ophthalmologist with regard to standardization. This is analogous to the national Fundus Photograph Reading Center on the basis of the 7-field photographic reference established by the Early Treatment for Diabetic Retinopathy Study.[26]

Four additional study limitations should be noted: (1) Images were not annotated with any clinical data. Although it is not clear that this would have affected interexpert diagnostic agreement, it may have biased against accurate photographic interpretation. (2) Study images had any visible peripheral ROP cropped out. If examiners are influenced by midperipheral changes such as vas-

cular branching while diagnosing plus disease, this may introduce a confounding factor. However, the standard plus disease photograph is a central retinal image without any peripheral details.[3] (3) For practical reasons, standardization of image reading conditions was not performed. The effect of variables such as luminance and resolution of computer monitor displays has been characterized in the radiology domain,[27] and the extent to which these factors may influence interexpert agreement is not clear. (4) Although experts were given the option to assign diagnoses of cannot determine and were asked to assess image quality, this study was not designed to analyze or correlate between those responses (eg, to explain why an expert may have provided a diagnosis despite reporting inadequate image quality). Eighteen of 748 diagnosis responses (2%) were cannot determine, and 20 quality responses (3%) were inadequate. All 34 images were assigned a diagnosis of other than cannot determine and a quality other than inadequate by at least 20 of 22 experts. One individual assigned both a diagnosis of cannot determine and a quality of inadequate to 11 images. Therefore, we do not believe that variabilities in expert impressions of image quality are of sufficient magnitude to affect the interpretation of study findings. Future studies investigating the relationship between perceived image adequacy and diagnostic performance would be informative.

In summary, the identification of retinal vascular changes, such as those reflected by plus disease, has critical significance for ROP management. Results from this study show that agreement in diagnosis of plus and preplus disease among ROP experts is imperfect. This may have important implications for clinical care, continued refinement of the international classification system, development of computer-based image analysis methods, and implementation of telemedicine systems.

Submitted for Publication: August 4, 2006; final revision received November 4, 2006; accepted November 25, 2006.
Correspondence: Michael F. Chiang, MD, MA, Department of Ophthalmology, Columbia University College of Physicians and Surgeons, 635 W 165th St, Box 92, New York, NY 10032 (chiang@dbmi.columbia.edu).
Author Contributions: Dr Chiang had full access to all of the data in the study and takes responsibility for the integrity of the data and accuracy of the data analysis.
Financial Disclosure: None reported.
Funding/Support: This study was supported by a Career Development Award from Research to Prevent Blindness (Dr Chiang) and by grant EY13972 from the National Eye Institute of the National Institutes of Health (Dr Chiang).
Role of the Sponsors: The sponsors had no role in the design or conduct of the study; in the collection, analysis, or interpretation of the data; or in the preparation, review, or approval of the manuscript.
Previous Presentation: Presented in part at the American Ophthalmological Society 143rd Annual Meeting; May 21, 2007; White Sulphur Springs, West Virginia.
Additional Contributions: We thank the 22 expert participants for their contributions to this study.

REFERENCES

1. Committee for the Classification of Retinopathy of Prematurity. An international classification of retinopathy of prematurity. *Arch Ophthalmol*. 1984;102(8):1130-1134.
2. International Committee for Classification of Late States of Retinopathy of Prematurity. An international classification of retinopathy of prematurity, II: the classification of retinal detachment [published correction appears in *Arch Ophthalmol*. 1987;105(11):1498]. *Arch Ophthalmol*. 1987;105(7):906-912.
3. Cryotherapy for Retinopathy of Prematurity Cooperative Group. Multicenter trial of cryotherapy for retinopathy of prematurity: preliminary results. *Arch Ophthalmol*. 1988;106(4):471-479.
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(12):1684-1694.
5. Palmer EA, Hardy RJ, Dobson V, et al; Cryotherapy for Retinopathy of Prematurity Cooperative Group. 15-year outcomes following threshold retinopathy of prematurity: final results from the Multicenter Trial of Cryotherapy for Retinopathy of Prematurity. *Arch Ophthalmol*. 2005;123(3):311-318.
6. Muñoz B, West SK. Blindness and visual impairment in the Americas and the Caribbean. *Br J Ophthalmol*. 2002;86(5):498-504.
7. Steinkuller PG, Du L, Gilbert C, Foster A, Collins ML, Coats DK. Childhood blindness. *J AAPOS*. 1999;3(1):26-32.
8. Gilbert C, Foster A. Childhood blindness in the context of VISION 2020: the right to sight. *Bull World Health Organ*. 2001;79(3):227-232.
9. STOP-ROP Multicenter Study Group. Supplemental Therapeutic Oxygen for Prethreshold Retinopathy of Prematurity (STOP-ROP): a randomized, controlled trial, I: primary outcomes. *Pediatrics*. 2000;105(2):295-310.
10. International Committee for the Classification of Retinopathy of Prematurity. The International Classification of Retinopathy of Prematurity revisited. *Arch Ophthalmol*. 2005;123(7):991-999.
11. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):157-174.
12. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull*. 1968;70:213-220.
13. Friedman CP, Wyatt JC. *Evaluation Methods in Medical Informatics*. New York, NY: Springer-Verlag; 1997.
14. Gelman R, Martinez-Perez ME, Vanderveen DK, Moskowitz A, Fulton AB. Diagnosis of plus disease in retinopathy of prematurity using Retinal Image multiScale Analysis. *Invest Ophthalmol Vis Sci*. 2005;46(12):4734-4738.
15. Swanson C, Cocker KD, Parker KH, Moseley MJ, Fielder AR. Semiautomated computer analysis of vessel growth in preterm infants without and with ROP. *Br J Ophthalmol*. 2003;87(12):1474-1477.
16. Wallace DK, Jomier J, Aylward WR, Landers MB III. Computer-automated quantification of plus disease in retinopathy of prematurity. *J AAPOS*. 2003;7(2): 126-130.
17. Wallace DK, Kylstra JA, Chestnutt DA. Prognostic significance of vascular dilation and tortuosity insufficient for plus disease in retinopathy of prematurity. *J AAPOS*. 2000;4(4):224-229.
18. Allen VG, Arocha JF, Patel VL. Evaluating evidence against diagnostic hypotheses in clinical decision making by students, residents and physicians. *Int J Med Inform*. 1998;51(2-3):91-105.
19. Laws DE, Morton C, Weindling M, Clark D. Systemic effects of screening for retinopathy of prematurity. *Br J Ophthalmol*. 1996;80(5):425-428.
20. Roth DB, Morales D, Feuer WJ, Hess D, Johnson RA, Flynn JT. Screening for retinopathy of prematurity employing the RetCam 120: sensitivity and specificity. *Arch Ophthalmol*. 2001;119(2):268-272.
21. Yen KG, Hess D, Burke B, Johnson RA, Feurer WJ, Flynn JT. Telephotoscreening to detect retinopathy of prematurity: preliminary study of the optimum time to employ digital fundus camera imaging to detect ROP. *J AAPOS*. 2002;6(2):64-70.
22. Ells AL, Holmes JM, Astle WF, et al. Telemedicine approach to screening for severe retinopathy of prematurity: a pilot study. *Ophthalmology*. 2003;110(11): 2113-2117.
23. Chiang MF, Keenan JD, Starren J, et al. Accuracy and reliability of remote retinopathy of prematurity diagnosis. *Arch Ophthalmol*. 2006;124(3):322-327.
24. Chiang MF, Keenan JD, Du YE, et al. Assessment of image-based technology: impact of referral cutoff on accuracy and reliability of remote retinopathy of prematurity diagnosis. *AMIA Annu Symp Proc*. 2005:126-130.
25. Chiang MF, Starren J, Du YE, et al. Remote image based retinopathy of prematurity diagnosis: a receiver operating characteristic analysis of accuracy. *Br J Ophthalmol*. 2006;90(10):1292-1296.
26. Early Treatment for Diabetic Retinopathy Study Research Group. Grading diabetic retinopathy from stereoscopic color fundus photographs: an extension of the modified Airlie House classification: ETDRS report No. 10. *Ophthalmology*. 1991;98(5)(suppl):786-806.
27. 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.