PLUS DISEASE IN RETINOPATHY OF PREMATURITY: AN ANALYSIS OF DIAGNOSTIC PERFORMANCE

BY Michael F. Chiang MD,* Rony Gelman MD, Lei Jiang BA, M. Elena Martinez-Perez PhD, Yunling E. Du PhD, AND **John T. Flynn MD**

ABSTRACT

Purpose: To measure agreement and accuracy of plus disease diagnosis among retinopathy of prematurity (ROP) experts; and to compare expert performance to that of a computer-based analysis system, Retinal Image multiScale Analysis.

Methods: Twenty-two recognized ROP experts independently interpreted a set of 34 wide-angle retinal photographs for presence of plus disease. Diagnostic agreement was analyzed. A reference standard was defined based on majority vote of experts. Images were analyzed using individual and linear combinations of computer-based system parameters for arterioles and venules: integrated curvature (IC), diameter, and tortuosity index (TI). Sensitivity, specificity, and receiver operating characteristic areas under the curve (AUC) for plus disease diagnosis were determined for each expert and for the computer-based system.

Results: Mean kappa statistic for each expert compared to all others was between 0 and 0.20 (slight agreement) in 1 expert (4.5%), 0.21 and 0.40 (fair agreement) in 3 experts (13.6%), 0.41 and 0.60 (moderate agreement) in 12 experts (54.5%), and 0.61 and 0.80 (substantial agreement) in 6 experts (27.3%). For the 22 experts, sensitivity compared to the reference standard ranged from 0.308 to 1.000, specificity from 0.571 to 1.000, and AUC from 0.784 to 1.000. Among individual computer system parameters compared to the reference standard, venular IC had highest AUC (0.853). Among linear combinations of parameters, the combination of arteriolar IC, arteriolar TI, venular IC, venular diameter, and venular TI had highest AUC (0.967).

Conclusion: Agreement and accuracy of plus disease diagnosis among ROP experts are imperfect. A computer-based system has potential to perform with comparable or better accuracy than human experts, but further validation is required.

Trans Am Ophthalmol Soc 2007;105:73-85

INTRODUCTION

Retinopathy of prematurity (ROP) is a leading cause of childhood blindness throughout the world.[1,2] Plus disease is a major component of the international classification of ROP[3,4] and is characterized by arteriolar tortuosity and venous dilation. The minimum amount of vascular abnormality required for plus disease is defined by a standard photograph, which was selected by expert consensus.[5] More recently, major clinical trials have explicitly required 2 or more quadrants of this amount of vascular change for the diagnosis of plus disease.[6,7] In addition, the 2005 revised International Classification of ROP formally defined an intermediate "pre-plus" condition as "abnormalities in the posterior pole that are insufficient for the diagnosis of plus disease but that demonstrate more arterial tortuosity and more venous dilation than normal."[4]

Accurate assessment for presence of plus disease is critical in clinical ROP management. The Cryotherapy for Retinopathy of Prematurity (CRYO-ROP) and Early Treatment for Retinopathy of Prematurity (ETROP) studies have established that plus disease is a necessary feature of threshold disease and a sufficient feature for diagnosis of type 1 ROP, both of which have been shown to warrant treatment with cryotherapy or laser photocoagulation.[5,7] Dilated binocular indirect ophthalmoscopy by an experienced examiner is considered the "gold standard" for ROP diagnosis and classification.[8] However, because the definition of plus disease is based on a photographic standard with descriptive qualifiers, its clinical diagnosis may be heavily subjective. This has important implications for ROP care because inconsistent diagnosis of plus disease can lead to errors in overtreatment or undertreatment.

Computer-based image analysis has potential to provide quantifiable, objective measurements to support the diagnosis of plus disease. Several studies have explored the possibility of automated plus disease detection by determining the accuracy of image analysis algorithms compared to a reference standard of dilated ophthalmoscopy by an experienced examiner.[9-11] Yet no published studies to our knowledge have attempted either to measure the accuracy of experts for diagnosis of plus disease or to compare the diagnostic performance of computer-based systems to that of human experts. This is an important gap in knowledge, because the utility of computer-based diagnostic systems will likely depend on the extent to which they can perform indistinguishably from, or better than, human experts.[12]

The purposes of this paper are to examine the agreement and accuracy of plus disease diagnosis among a group of 22 recognized ROP experts and to compare expert performance to that of a computer-based image analysis system, Retinal Image multiScale Analysis (RISA).[13,14] This study utilizes a set of 34 retinal images captured by a commercially available wide-angle digital fundus camera. Performance was measured by defining a reference standard based on majority vote of experts.

From the Department of Ophthalmology (Drs Chiang, Gelman, and Flynn and Mr Jiang) and Department of Biomedical Informatics (Dr Chiang), Columbia University College of Physicians and Surgeons, New York, New York; Department of Computer Science, Institute of Research in Applied Mathematics and Systems, National Autonomous University of Mexico, Mexico City (Dr Martinez-Perez); and Department of Epidemiology and Population Health, Albert Einstein College of Medicine, New York, New York (Dr Du).

*Presenter.

Bold type indicates AOS member.

METHODS

STUDY PARTICIPANTS

This study was approved by the Institutional Review Board at Columbia University Medical Center and included waiver of consent for use of de-identified retinal images. 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 the opinion of the authors, reflected a change in vasculature compared to baseline. Each photograph displayed the posterior 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 this study. For the purpose of this study, eligible experts were defined as practicing pediatric ophthalmologists or retina specialists who met at least 1 of 3 criteria: having been a study center principal investigator (PI) for either the CRYO-ROP or ETROP studies, having been a certified investigator for either of those studies, or having coauthored 5 or more peer-reviewed ROP manuscripts.

IMAGE INTERPRETATION: EXPERTS

Each participant was provided with an anonymous study identifier and password to a Web-based program developed by the authors to display photographs. Experts were asked to categorize each image based on two 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 institution had a RetCam device and whether they would describe their experience with interpreting RetCam-captured images as "extensive," "limited," or "none."

Informed consent was obtained from each participant using a click-through Web form before images were displayed. Participants were not provided with any photographs or definitions for "plus" and "pre-plus" disease, although it was assumed that they would be intimately familiar with these definitions. Instead, the decision was made to rely on experts' personal experience and judgment to better simulate a real-world situation. For each image, a reference standard diagnosis was defined as the response ("plus" or "not plus") given by the majority of experts. In cases of ties, the more severe diagnosis was selected as the reference standard.

IMAGE ANALYSIS: COMPUTER-BASED SYSTEM

All retinal vessels in each image were identified and classified by consensus of four coauthors (R.G., L.J., J.T.F., M.F.C.) as arterioles or venules and subsequently analyzed by the RISA computer-based system using previously described methods.[9] Three system parameters were calculated for all vessels in each image: integrated curvature (IC), diameter, and tortuosity index (TI). *IC* (radians/pixel) is defined as the sum of angles along the vessel, normalized by its length; *diameter* (pixels) is the total area of the vessel divided by its length; and *TI* is the length of the vessel divided by the length of a line segment connecting its end points (Figure 1). The average diameter of the optic nerve head in these images was 50.77 pixels. If the physical diameter of the optic nerve head is assumed to be 1.015 mm at <40 weeks gestation,[15] then 1 pixel in these images represents an average of 20.00 μm physical distance.

DATA ANALYSIS: EXPERTS

Data were exported from the study database into a spreadsheet (Excel 2003; Microsoft, Redmond, Washington). Absolute agreement among participants was determined for each image based on a 2-level ("plus," "not plus") categorization. Images classified as "cannot determine" were excluded from analysis for that participant. The kappa statistic was used to measure chance-adjusted agreement between each pair of experts. Mean kappa values were determined for each expert compared to all others, and standard errors were calculated using the jackknife method. A well-known scale was used to interpret results: 0 to 0.20 = slight agreement, 0.21 to 0.40 = fair agreement, 0.41 to 0.60 = moderate agreement, 0.61 to 0.80 = substantial agreement, and 0.81 to 1.00 = almost-perfect agreement.[16] The relationship between expert characteristics and mean kappa statistics compared to other examiners was analyzed using the 1-sample or independent-sample *t* test, as appropriate. Statistical significance was defined as a 2-sided *P* value <.05.

Diagnostic accuracy of experts was determined based on sensitivity, specificity, and kappa statistic compared to the reference standard. Receiver operating characteristic (ROC) curves were plotted to represent diagnostic performance of each expert,[17] and area under the curve (AUC) was calculated nonparametrically for each expert. McNemar's test was used to detect systematic tendencies of each expert to undercall or overcall plus disease compared to the reference standard.

DATA ANALYSIS: COMPUTER-BASED SYSTEM

The mean value of each individual system parameter (IC, diameter, TI) was calculated in every image. Arterioles and venules were analyzed separately. Mean values for each parameter were compared between images that were considered to be "plus" and those that were "not plus" according to the reference standard, using the Mann-Whitney test. Logistic regression was used to model 26 possible linear combinations of parameters, taken two or more at a time.[18] For each individual parameter and each linear combination, sensitivity and specificity of the computer-based system compared to the reference standard was plotted as a function of the cutoff threshold used to separate "plus" from "not plus." ROC curves were plotted for individual parameters and linear combinations, and

AUC was calculated nonparametrically. AUCs were compared between (1) linear combinations, (2) linear combinations and individual parameters, and (3) linear combinations and experts using the Delong approach.[19]

All data analysis was performed using statistical and computational software packages (Minitab version 13, Minitab Inc, State College, Pennsylvania; SPSS version 14, SPSS Inc, Chicago, Illinois; R programming language version 2.4.0, Free Software Foundation, Boston, Massachusetts). In particular, calculation of area under ROC curves was performed using SPSS software.

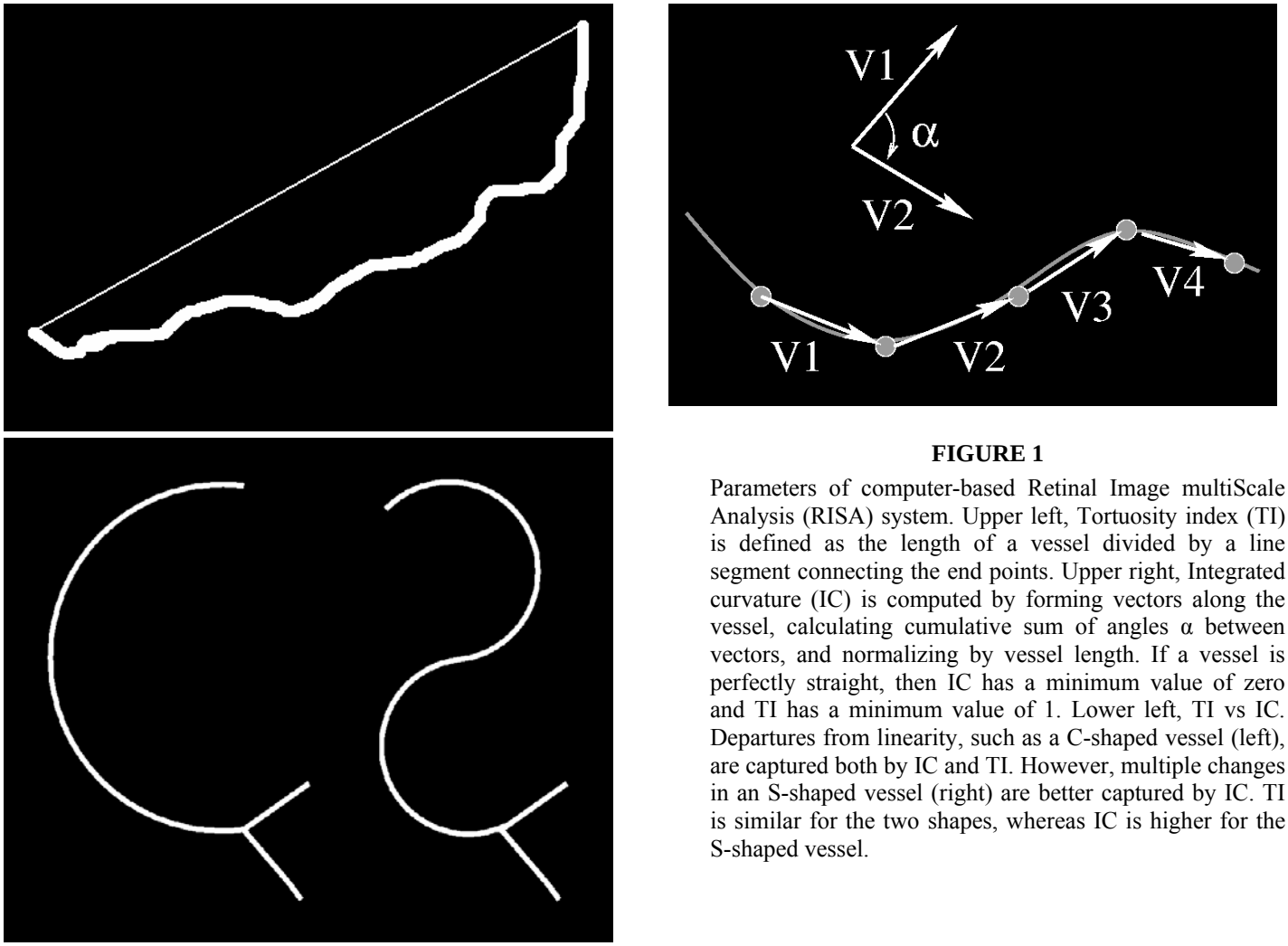


FIGURE 1

Parameters of computer-based Retinal Image multiScale Analysis (RISA) system. Upper left, Tortuosity index (TI) is defined as the length of a vessel divided by a line segment connecting the end points. Upper right, Integrated curvature (IC) is computed by forming vectors along the vessel, calculating cumulative sum of angles α between vectors, and normalizing by vessel length. If a vessel is perfectly straight, then IC has a minimum value of zero and TI has a minimum value of 1. Lower left, TI vs IC. Departures from linearity, such as a C-shaped vessel (left), are captured both by IC and TI. However, multiple changes in an S-shaped vessel (right) are better captured by IC. TI is similar for the two shapes, whereas IC is higher for the S-shaped vessel.

RESULTS

CHARACTERISTICS OF EXPERT PARTICIPANTS

Among the 22 expert participants in this study, 18 (81.8%) had served as a PI in the CRYO-ROP or ETROP studies, 4 (18.2%) had served as a certified investigator in either study, and 11 (50.0%) had coauthored 5 or more peer-reviewed ROP manuscripts. Nine participants (40.9%) met more than 1 criterion. Sixteen experts (72.7%) worked at institutions with a RetCam device, 5 (22.7%) worked at institutions without the device, and 1 (4.5%) did not report this information. Similarly, 11 participants (50.0%) described their previous experience with interpretation of RetCam-captured images as "extensive," 6 (27.3%) described it as "limited," 4 (18.2%) described it as "none," and 1 (4.5%) did not report this information. Seventeen participants (77.3%) were pediatric ophthalmologists, whereas 5 (22.7%) were retina specialists.

INTER-EXPERT AGREEMENT

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 (2.4%) of the 748 cases. Image quality was scored as "adequate" in 656 (87.7%), "possibly adequate" in 72 (9.6%), and "inadequate" for diagnosis in 20 (2.7%) of the 748 cases. Overall diagnostic responses are summarized in Table 1. Three (8.8%) of the 34 images were classified as "plus" by all 22 experts, and 3 images (8.8%) were classified as "not plus" by all experts who provided a diagnosis. Representative images and responses are shown in Figure 2.

TABLE 1. ABSOLUTE AGREEMENT IN PLUS DISEASE DIAGNOSIS AMONG 22 EXPERTS REVIEWING 34 IMAGES*

IMAGE	2-LEVEL CATEGORIZATION BY 22 EXPERTS, N (%)			
	Plus		Not plus	
1	3	(13.6%)	19	(86.4%)
2	1	(4.8%)	20	(95.2%)
3	14	(70.0%)	6	(30.0%)
4	5	(23.8%)	16	(76.2%)
5	3	(14.3%)	18	(85.7%)
6	22	(100.0%)	0	(0.0%)
7	1	(4.5%)	21	(95.5%)
8	21	(95.5%)	1	(4.5%)
9	0	(0.0%)	21	(100.0%)
10	0	(0.0%)	22	(100.0%)
11	22	(100.0%)	0	(0.0%)
12	1	(4.5%)	21	(95.5%)
13	7	(31.8%)	15	(68.2%)
14	2	(9.5%)	19	(90.5%)
15	12	(60.0%)	8	(40.0%)
16	1	(4.8%)	20	(95.2%)
17	8	(38.1%)	13	(61.9%)
18	1	(4.5%)	21	(95.5%)
19	2	(9.5%)	19	(90.5%)
20	20	(95.2%)	1	(4.8%)
21	0	(0.0%)	21	(100.0%)
22	11	(52.4%)	10	(47.6%)
23	17	(77.3%)	5	(22.7%)
24	0	(0.0%)	22	(100.0%)
25	2	(9.5%)	19	(90.5%)
26	16	(72.7%)	6	(27.3%)
27	1	(4.5%)	21	(95.5%)
28	14	(63.6%)	8	(36.4%)
29	1	(4.8%)	20	(95.2%)
30	17	(81.0%)	4	(19.0%)
31	1	(4.5%)	21	(95.5%)
32	3	(13.6%)	19	(86.4%)
33	17	(77.3%)	5	(22.7%)
34	22	(100.0%)	0	(0.0%)

*Findings are displayed based on 2-level categorization ("plus," "not plus"). Images categorized as "cannot determine" were excluded for that expert.

Figure 3, top, shows absolute agreement in plus disease diagnosis, based on percentage of experts who assigned the same diagnosis to each image. The same 2-level diagnosis was made by 90% or more of experts in 20 (58.8%) of the images, and by 80% or more of

experts in 24 images (70.6%). As shown in Figure 3, bottom, the mean kappa for each expert compared to all others was between 0 and 0.20 (slight agreement) in 1 expert (4.5%), 0.21 and 0.40 (fair agreement) in 3 experts (13.6%), 0.41 and 0.60 (moderate agreement) in 12 experts (54.5%), and 0.61 and 0.80 (substantial agreement) in 6 experts (27.3%).

There was no statistically significant difference in mean kappa or weighted kappa statistics based on working in vs not working in an institution with a RetCam; having published 5 or more vs less 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 PI vs not a PI in the CRYO-ROP or ETROP studies; or status as a certified investigator vs not a certified investigator in either of those studies.

FIGURE 2

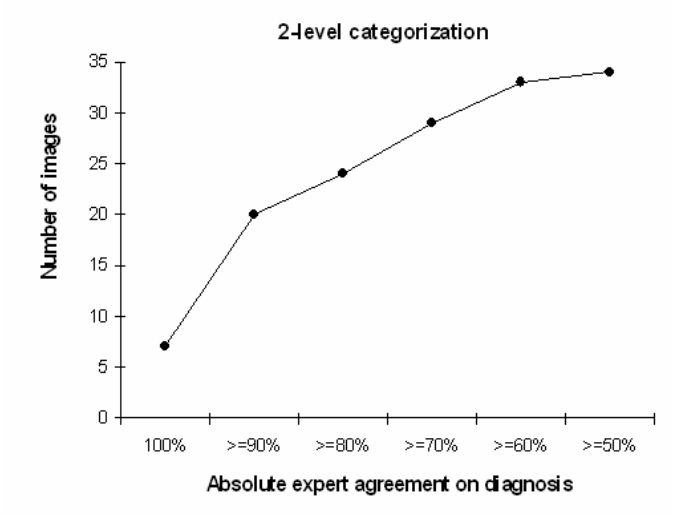
Representative images shown to 22 expert participants. Upper left, Image was classified as "neither plus nor pre-plus" by all 22 experts (100%). Upper middle, Image was classified as "plus" by 2 experts (9.5%), "pre-plus" by 9 (42.9%), "neither" by 10 (47.6%), and "cannot determine" by 1 expert. Upper right, Image was classified as "plus" by 1 expert (4.8%), "pre-plus" by 16 experts (76.2%), "neither" by 4 (19.0%), and "cannot determine" by 1 expert. Lower left, Image was classified as "plus" by 11 experts (52.4%), "pre-plus" by 10 (47.6%), and "cannot determine" by 1 expert. Lower middle, Image was classified as "plus" by all 22 experts (100%). Lower right, Image was classified as "plus" by 3 experts (14.3%), "pre-plus" by 9 (42.9%), "neither" by 9 (42.9%), and "cannot determine" by 1 expert.

EXPERT PERFORMANCE

Table 2 summarizes accuracy of each expert compared to the majority-vote reference standard. Thirteen (59.0%) of the 22 experts had sensitivity >80%, and 18 (81.8%) had specificity >80%. Nine experts (40.9%) had both sensitivity and specificity >80%, and 1 expert (4.5%) had both sensitivity and specificity of 100%. Compared to the reference standard, 1 expert (4.5%) had a kappa value of 0.21 to 0.40 (fair agreement), 4 (18.2%) had kappa of 0.41 to 0.60 (moderate agreement), 8 (36.4%) had kappa of 0.61 to 0.80 (substantial agreement), and 9 (40.9%) had kappa of 0.81 to 1.00 (near-perfect agreement). Kappa values for all experts reflected diagnostic performance that was significantly better than chance ($P < .0014$ for all experts). According to the McNemar test of bias, 2 experts

(9.1%) had a statistically significant tendency to undercall plus disease (P = .016 for expert No. 1, P = .004 for expert No. 5), and 2 (9.1%) had a statistically significant tendency to overcall plus disease (P = .016 for expert No. 4, P = .004 for expert No. 21).

Based on ROC analysis, AUC for experts ranged from 0.784 to 1.000. Two experts (9.1%) had AUC between 0.701 and 0.800, 8 experts (36.4%) had AUC between 0.801 and 0.900, and 12 (54.5%) had AUC between 0.901 and 1.000. AUC values for all experts reflected diagnostic performance that was significantly better than chance (P < .006 for all experts).



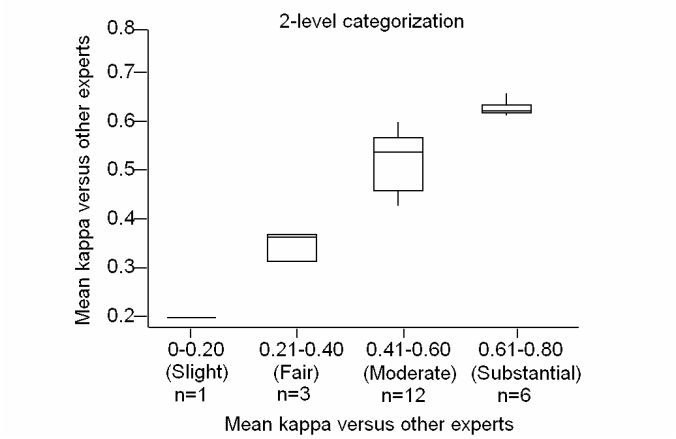


FIGURE 3

Inter-expert agreement in plus disease diagnosis among 22 experts reviewing 34 retinal images. Top, Percentage of experts who assigned the same diagnosis ("plus" or "not plus") to images. Bottom, Box plot of mean kappa statistic for each expert compared to all others. Boxes represent the 25th, 50th, and 75th percentile kappa values; whiskers represent 10th and 90th percentile values.

COMPUTER-BASED SYSTEM PERFORMANCE

From the 21 images without plus disease according to the reference standard, 66 arterioles and 83 venules were analyzed. From the 13 images with plus disease, 51 arterioles and 55 venules were analyzed. On average, 3 arterioles and 4 venules were analyzed per image.

Sensitivity and specificity curves for plus disease detection based on individual and combined system parameters are plotted in Figure 4, and AUC values are displayed in Table 3. All individual and linear combinations of parameters had AUC values that performed significantly better than chance (P < .05), with the exception of arteriolar diameter (P = .559). The AUC of linear combination III was higher than that of any individual system parameter (P = .107 vs arteriolar IC, P < .05 vs all other individual parameters), was higher than that of any other combination, and was higher than that of 18 (81.8%) of 22 experts (P < .05 vs 4 experts). Table 3 displays sample values for sensitivity and specificity of the computer-based system, by selecting points at the intersection of the curves in Figure 3. Linear combinations II and III had the highest sensitivity and specificity (0.938). In comparison,

3 (13.6%) of 22 experts had both sensitivity and specificity greater than linear combinations II or III (Table 2).

DISCUSSION

This is the first study to our knowledge that has systematically evaluated agreement and accuracy among a group of ROP experts for plus disease diagnosis and compared expert performance to that of a computer-based system. Two key findings are as follows: (1) Accuracy and agreement of plus disease diagnosis by experts are imperfect. (2) Modeling of RISA system parameters has potential to produce diagnostic accuracy that is comparable to, or better than, that of human experts.

TABLE 2. ROP EXPERT SENSITIVITY, SPECIFICITY, KAPPA STATISTIC, AND RECEIVER OPERATING CHARACTERISTIC AREA UNDER THE CURVE (AUC) FOR DETECTION OF PLUS DISEASE, COMPARED TO THE REFERENCE STANDARD OF MAJORITY VOTE AMONG 22 RECOGNIZED ROP EXPERTS

EXPERT NO.	SENSITIVITY	SPECIFICITY	KAPPA (SE)	AUC (SE)
1	0.462	1.000	0.51 (0.14)	0.859 (0.063)
2	0.615	1.000	0.66 (0.13)	0.982 (0.018)
3	0.615	1.000	0.66 (0.13)	0.890 (0.057)
4	1.000	0.667	0.61 (0.12)	0.833 (0.070)
5	0.308	1.000	0.35 (0.14)	0.951 (0.035)
6	0.846	0.714	0.53 (0.14)	0.784 (0.082)
7	0.846	0.952	0.81 (0.10)	0.921 (0.052)
8	0.923	0.905	0.82 (0.10)	0.921 (0.052)
9	0.923	0.905	0.82 (0.10)	0.921 (0.052)
10	0.778	1.000	0.83 (0.12)	0.918 (0.065)
11	0.538	0.857	0.41 (0.16)	0.885 (0.061)
12	0.769	1.000	0.81 (0.11)	0.956 (0.032)
13	0.692	1.000	0.74 (0.12)	0.934 (0.041)
14	1.000	1.000	1.00 (0.00)	1.000 (0.000)
15	0.750	0.952	0.73 (0.13)	0.899 (0.058)
16	0.923	0.810	0.70 (0.12)	0.879 (0.061)
17	0.846	1.000	0.87 (0.09)	0.967 (0.029)
18	1.000	0.750	0.74 (0.13)	0.875 (0.080)
19	0.923	0.857	0.76 (0.11)	0.896 (0.059)
20	1.000	0.952	0.94 (0.06)	0.976 (0.028)
21	1.000	0.571	0.51 (0.12)	0.786 (0.077)
22	1.000	0.952	0.94 (0.06)	0.976 (0.028)

ROP, retinopathy of prematurity; SE, standard error.

The first main finding is that consistency and accuracy of plus disease diagnosis by experts are not perfect. This is particularly relevant because the ETROP trial has determined that presence of plus disease is sufficient for meeting the definition of type 1 ROP, which benefits from early treatment.[7] As shown in Figure 3, all experts who provided a diagnosis agreed in 7 (20.6%) of the 34 images, and the mean kappa for each expert compared to all others ranged from 0.19 (slight agreement) to 0.66 (substantial agreement). Moreover, the finding that 4 of 22 study participants had statistically significant tendencies to underdiagnose or overdiagnose plus disease supports the notion that even recognized experts may have differing subjective interpretations for vascular "dilation" and "tortuosity" (Table 2). Taken together, these results raise important concerns about the accuracy and reliability of ROP diagnosis and treatment. Development of a quantifiable, objective definition of plus disease could eventually result in improved ROP management. This would be analogous to widely used methods for automated interpretation of electrocardiograms and Papanicolaou smears.[20,21]

The design of a study involving inter-expert 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 literature. This may not necessarily reflect clinical expertise in a real-world setting. However, medical expertise comprises numerous factors, some of which may be difficult to quantify for the purpose of study design.[22] 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 believe it is reasonable to hypothesize that disagreement in plus disease diagnosis within the overall population of practicing clinicians may be higher than among the academic experts in this study.

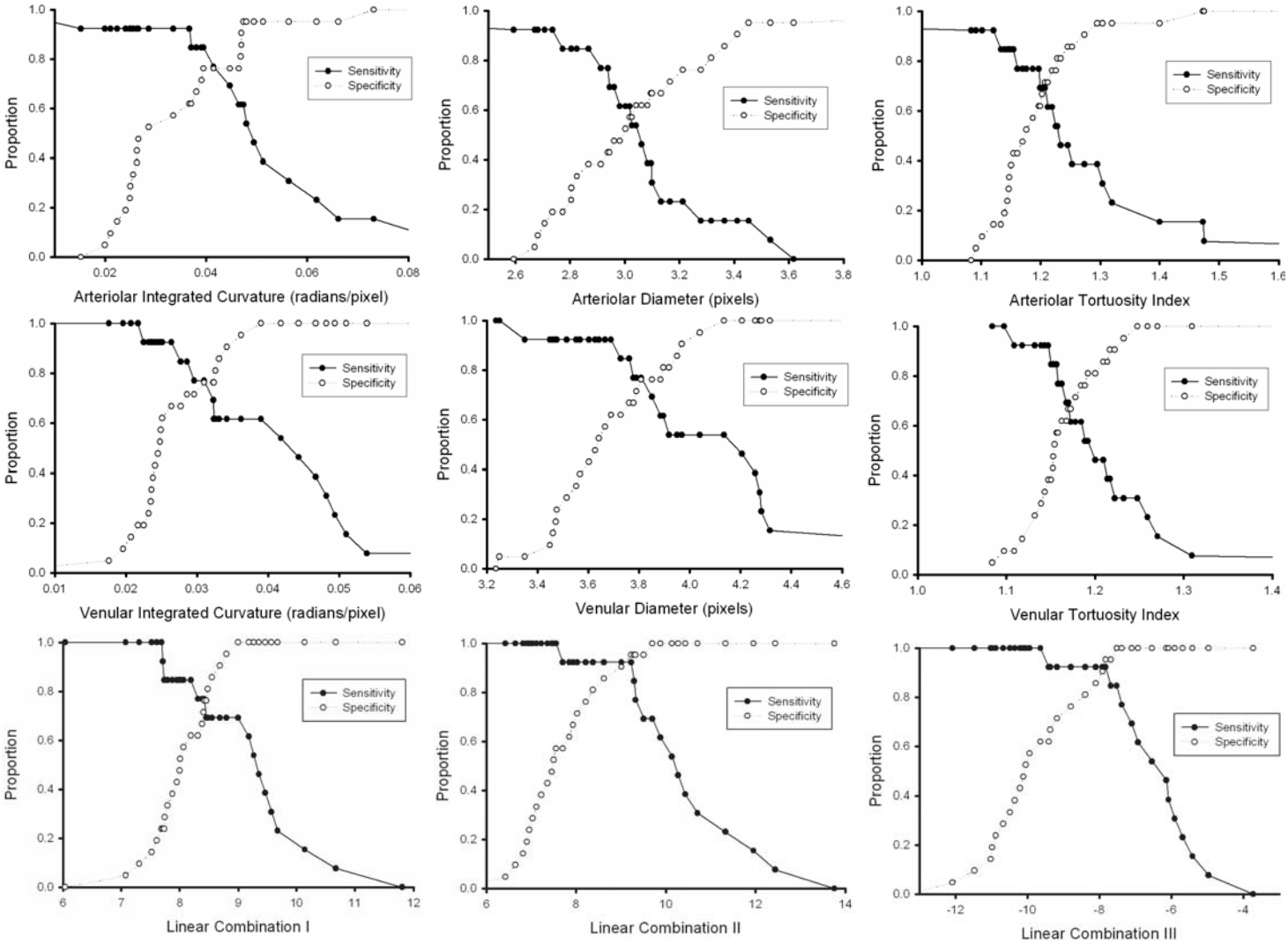


FIGURE 4

Sensitivity and specificity of individual computer system parameters and linear combinations of parameters for plus disease diagnosis, compared to the reference standard of majority vote among 22 recognized ROP experts. Curves are displayed as a function of parameter cutoff criteria for detection of plus disease: Upper left, arteriolar integrated curvature (IC). Upper middle, arteriolar diameter. Upper right, arteriolar tortuosity index (TI). Center left, venular IC. Center middle, venular diameter. Center right, venular TI. Lower left, Linear combination I (arteriolar IC and venular diameter). Lower middle, Linear combination II (arteriolar IC, venular IC, and venular diameter). Lower right, Linear combination III (arteriolar IC, arteriolar TI, venular IC, venular diameter, and venular TI).

From a clinical perspective, it would be most useful to know the accuracy and 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.[23] 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-degree field of view and is currently the most well-known

instrument for pediatric retinal imaging.[17,24-27] In contrast, standard binocular indirect ophthalmoscopy provides a 40- to 50-degree 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 kappa 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 to serial ophthalmoscopy, because examination quality may vary based on infant stability or cooperation.

TABLE 3. COMPUTER-BASED SYSTEM SENSITIVITY, SPECIFICITY, AND RECEIVER OPERATING CHARACTERISTIC AREA UNDER THE CURVE (AUC) FOR DETECTION OF PLUS DISEASE, COMPARED TO THE REFERENCE STANDARD OF MAJORITY VOTE AMONG 22 RECOGNIZED ROP EXPERTS

SYSTEM PARAMETER	SENSITIVITY*	SPECIFICITY*	AUC (SE)	AUC *P* VALUE†
Arteriolar IC	0.760	0.760	0.817 (0.085)	.002
Arteriolar diameter	0.590	0.590	0.560 (0.102)	.559
Arteriolar TI	0.700	0.700	0.707 (0.099)	.045
Venular IC	0.760	0.760	0.853 (0.072)	.001
Venular diameter	0.760	0.760	0.821 (0.082)	.002
Venular TI	0.640	0.640	0.744 (0.089)	.018
Linear combination I‡	0.766	0.766	0.835 (0.083)	.001
Linear combination II§	0.938	0.938	0.956 (0.036)	<.001
Linear combination III¶	0.938	0.938	0.967 (0.031)	<.001

IC, integrated curvature; SE, standard error; TI, tortuosity index.
*For the computer-based system, values at the intersection points of the sensitivity and specificity curves in Figure 4 are displayed in this table.
†Null hypothesis: true area under curve = 0.5 at level of significance $P < .05$.
‡Comprising arteriolar IC and venular diameter.
§Comprising arteriolar IC, venular IC, and venular diameter.
¶Comprising arteriolar IC, arteriolar TI, venular IC, venular diameter, and venular TI.

The second main study finding is that a computer-based system appears able to detect plus disease with a high degree of accuracy. Based on sensitivity, specificity, and area under ROC curves, the highest diagnostic accuracies for single blood vessel parameters were achieved with venular IC, venular diameter, and arteriolar IC (Table 3). This is consistent with previously published findings involving the RISA system for plus disease detection[9] and strengthens other work involving automated image analysis for ROP diagnosis.[10,11] As shown in Table 3, the use of linear combinations of system parameters results in better performance than the use of individual parameters. This is not surprising, given that the clinical diagnosis of plus disease is presumably based on a combination of multiple retinal vascular characteristics.

The results of this study suggest that it may be possible for a computer-based system to detect plus disease, with performance that appears to be indistinguishable from, or better than, that of recognized experts. As shown in Tables 2 and 3, area under the ROC curve is higher than that of all but 5 experts using linear combination II of computer system parameters, and higher than all but 4 experts using linear combination III. Similarly, the sensitivity and specificity of plus disease diagnosis using these computer-based parameters is comparable to, or better than, that of many experts. We note that the sensitivity or specificity of the computer system may be improved by adjusting the cutoff threshold for any given parameter, but that an increase in one inevitably causes a decrease in the other (Figure 4). Of course, determination of the optimal sensitivity-specificity operating point in a real-world automated plus disease detection system would require consideration of tradeoffs between missed cases of treatment-requiring ROP (false-negatives) and the cost of unnecessary referrals (false-positives). From a practical standpoint, a computer-based image analysis system could be used to provide real-time decision support for ophthalmologists.[28,29]

The reference standard diagnosis in this study was defined as majority vote among 22 ROP experts who interpreted images independently. We feel that this is a reasonable approach but acknowledge that it does not necessarily represent a true "gold standard." In particular, it could be argued that this majority-vote reference standard could be influenced by differences of expert opinion in borderline cases. To investigate this possibility by eliminating photographs that were apparently the most "borderline," results were calculated using alternative scenarios in which the reference standard was defined as diagnoses that were agreed upon by ≥13 of 22

experts and ≥14 of 22 experts. Sensitivity/specificity for diagnosis of plus disease ranged from 33.3%/57.1% to 100%/100% using a reference standard requiring agreement by 13 or more of 22 experts (1 borderline image excluded), and from 36.4%/57.1% to 100%/100% using a reference standard requiring agreement by 14 or more of 22 experts (2 borderline images excluded). We believe this suggests that our findings regarding expert accuracy are robust; in fact, it is precisely these borderline cases in which an objective computer-based diagnosis system may provide most value added for clinicians. A related point deserving explanation is that experts were compared against a reference standard that each of them contributed toward establishing. Although this may appear circular, we note that each expert contributed only <5% (1/22) toward this standard, and that this methodology has a small bias toward *improving* expert correlation with the reference standard. Additional studies involving alternative methods for obtaining reference standards by expert consensus, such as the Delphi technique,[30] may be informative. Future "gold standards" could also be based on other quantitative tests, such as ocular blood flow measurements.

Telemedicine strategies have been proposed as an alternative to standard ROP care involving dilated examination at the neonatal bedside. Several studies have demonstrated that remote image interpretation may have adequate sensitivity and specificity to identify clinically significant ROP.[17,24-27,31] However, this study raises some concerns about remote diagnosis because of the number of clinically significant disagreements among recognized experts reviewing wide-angle retinal photographs. To prevent diagnostic errors, this issue must be studied and resolved before the routine deployment of ROP telemedicine systems. However, if implemented properly, telemedical interpretation at certified centers could offer advantages over dilated examination by a single ophthalmologist with regard to standardization. This is analogous to the Fundus Photograph Reading Center based on the 7-field photographic reference established by the Early Treatment for Diabetic Retinopathy Study.[32]

Several additional study limitations should be noted:

1. Images were not annotated with any clinical data. The extent to which this information may have affected diagnostic accuracy and agreement among experts is not clear.
2. Study images had any visible peripheral ROP cropped out. If examiners are influenced by midperipheral changes such as vascular branching while diagnosing plus disease, then this may introduce a confounding factor. However, we note that the standard plus disease photograph is a central retinal image without any peripheral details.[5]
3. For practical reasons, standardization of image reading conditions by experts was not performed. The impact of parameters such as luminance and resolution of computer monitor displays has been characterized in the radiology domain,[33] and the extent to which these factors may influence diagnostic performance is unknown.
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. Future studies investigating the relationship between perceived image adequacy and diagnostic performance may be informative.
5. This study relied on a single data set used by both experts and the computer-based system. A main intent of the study was to characterize diagnostic accuracy of the computer-based system compared to that of human experts, using methodologies such as ROC analysis. Further studies examining expert and system performance on independent testing sets, or using cross-validation techniques, will be critical for full system evaluation.[34,35]
6. Computer-based system parameters were calculated using all identifiable vessels in each image. It is possible that including all vessels could cause a washout effect if images had a mix of "normal" and "abnormal" vessels. Although determination of computer-based system parameter values based on the two most abnormal arterioles and venules in each image showed no significant difference in diagnostic performance (data not shown), further investigation may be warranted.

In summary, this study demonstrates that accuracy and reliability of plus disease diagnosis by ROP experts are imperfect and that a computer-based system has potential to perform comparably to, or better than, human experts. Further validation of automated systems for plus disease diagnosis is required. This may have important implications for clinical care, continued refinement of the ROP classification system, development of computer-based image analysis methods, and implementation of telemedicine systems.

ACKNOWLEDGMENTS

Funding/Support: This study was supported by a Career Development Award from Research to Prevent Blindness (M.F.C.) and by grant EY13972 from the National Eye Institute of the National Institutes of Health (M.F.C.).

Financial Disclosures: None.

Author Contributions: *Design and conduct of the study* (M.F.C., R.G., J.T.F.); *Collection, management, analysis, and interpretation of the data* (R.G., M.F.C., L.J., Y.E.D., M.E.M., J.T.F.); *Preparation, review, or approval of the manuscript* (M.F.C., R.G., L.J., Y.E.D., M.E.M., J.T.F.).

Other Acknowledgments: The authors would like to thank each of the 22 expert participants for their contribution to this study.

REFERENCES

1. Munoz B, West SK. Blindness and visual impairment in the Americas and the Caribbean. *Br J Ophthalmol* 2002;86:498-504.
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.

3. Committee for the Classification of Retinopathy of Prematurity. An international classification of retinopathy of prematurity. *Arch Ophthalmol* 1984;102:1130-1134.
4. International Committee for the Classification of Retinopathy of Prematurity. The International Classification of Retinopathy of Prematurity revisited. *Arch Ophthalmol* 2005;123:991-999.
5. CRYO-ROP Cooperative Group. Multicenter trial of cryotherapy for retinopathy of prematurity: preliminary results. *Arch Ophthalmol* 1988;106:471-479.
6. STOP-ROP Multicenter Study Group. Supplemental therapeutic oxygen for prethreshold retinopathy of prematurity (STOP-ROP), a randomized, controlled trial, I: primary outcomes. *Pediatrics* 2000;105:295-310.
7. ETROP 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.
8. Section on Ophthalmology AAP, AAO, AAPOS. Screening examination of premature infants for retinopathy of prematurity. *Pediatrics* 2006;117:572-576 [erratum in: *Pediatrics* 2006;118:1324].
9. 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;46:4734-4738.
10. 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;87:1474-1477.
11. Wallace DK, Jomier J, Aylward WR, Landers MB. Computer-automated quantification of plus disease in retinopathy of prematurity. *J AAPOS* 2003;7:126-130.
12. Turing AM. Computing machinery and intelligence. *MIND* 1950;49:433-460.
13. Martinez-Perez ME. Hughes AD, Thom SA, Bharath AA, Parker KH. Segmentation of blood vessels from red-free and fluorescein retinal images. *Med Image Anal* 2007;11:47-61.
14. Martinez-Perez ME, Hughes AD, Stanton AV, et al. Retinal vascular tree morphology: a semi-automatic quantification. *IEEE Trans Biomed Eng* 2002;49:912-917.
15. Rimmer S, Keating C, Chou T, et al. Growth of the human optic disk and nerve during gestation, childhood, and early adulthood. *Am J Ophthalmol* 1993;116:748-753.
16. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:157-174.
17. Chiang MF, Starren J, Du E, et al. Remote image-based retinopathy of prematurity diagnosis: a receiver operating characteristic (ROC) analysis of accuracy. *Br J Ophthalmol* 2006;90:1292-1296.
18. Liu A, Schisterman EF, Zhu Y. On linear combinations of biomarkers to improve diagnostic accuracy. *Stat Med* 2005;24:37-47.
19. Hanley JA, Hajian-Tilake KO. Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update. *Acad Radiol* 1997;4:49-58.
20. Hongo RH, Goldschlager N. Status of computerized electrocardiography. *Cardiol Clin* 2006;24:491-504.
21. Ku NN. Automated Papanicolaou smear analysis as a screening tool for female lower genital tract malignancies. *Curr Opin Obstet Gynecol* 1999;11:41-43.
22. 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;51:91-105.
23. Laws DE, Morton C, Weindling M, Clark D. Systemic effects of screening for retinopathy of prematurity. *Br J Ophthalmol* 1996;80:425-428.
24. 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;119:268-272.
25. Yen KG, Hess D, Burke B, Johnson RA, Feuer 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;6:64-70.
26. Ells AL, Holmes JM, Astle WF, et al. Telemedicine approach to screening for severe retinopathy of prematurity: a pilot study. *Ophthalmology* 2003;110:2113-2117.
27. Chiang MF, Keenan JD, Starren J, et al. Accuracy and reliability of remote retinopathy of prematurity diagnosis. *Arch Ophthalmol* 2006;124:322-327.
28. Maviglia SM, Yoon CS, Bates DW, Kuperman G. KnowledgeLink: impact of context-sensitive information retrieval on clinicians' information needs. *J Am Med Inform Assoc* 2006;13:67-73.
29. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ* 2005;330:765-768.
30. Kors JA, Sittig AC, van Bemmel JH. The Delphi method to validate diagnostic knowledge in computerized ECG interpretation. *Methods Inf Med* 1990;29:44-50.
31. Chiang MF, Wang L, Busuioc M, et al. Telemedical retinopathy of prematurity diagnosis: accuracy, reliability, and image quality. *Arch Ophthalmol.* Forthcoming.
32. ETDRS 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.
33. 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;215:169-174.

34. Rosado B, Menzies S, Harbauer A, et al. Accuracy of computer diagnosis of melanoma: a quantitative meta-analysis. *Arch Dermatol* 2003;139:361-367.
35. Aleynikov S, Micheli-Tzanakou E. Classification of retinal damage by a neural network based system. *J Med Syst* 1998;22:129-136.

PEER DISCUSSION

DR DAVID K. COATS: I appreciate the opportunity to discuss this paper by Dr. Chiang and colleagues. They conclude that accuracy and agreement on the diagnosis of plus disease by experts is not perfect. This certainly appears to be the case based on the data presented. In only 6 of 30 eyes where at least one expert believed plus disease was present was there agreement among 80% or more of the experts. When plus disease was not present, however, agreement was much better. Expert agreement was 80% or more in 17 of 31 eyes where at least one expert did not think plus disease was present. This suggests that it is the equivocal cases that may be most problematic. Since a diagnosis of severe retinopathy of prematurity (ROP) is not made by analyzing the posterior pole vessels in isolation, it would be very interesting to know how these same experts would have performed if information about the zone, stage, and extent of ROP and general medical information about the infants were made available.

In contrast to human performance, the accuracy of a computer-based analysis was quite good. After analyzing several linear combinations of parameters for plus disease diagnosis, the authors impressively demonstrated several linear combinations that performed quite well (ie, Figure 4, lower middle and lower right), exceeding the performance of all but 3 human experts.

However, I have some concerns about the study design. The reference standard used in the study is not very stringent. I think it would have been optimal to have one group of experts set the standard and have a separate group of experts grade the eyes against this standard. More important, why did the authors choose to use a simple majority of expert opinion to establish the reference standard? A supermajority of 75% or 80% agreement would appear to be more appropriate for a subjective disease feature such as plus disease. I recognize that if 75% expert agreement had been selected as the reference standard, only 8 eyes would have been classified as having plus disease. Thus, I wonder if this decision was made prior to data collection or sometime after the data were analyzed because of sample size considerations.

Another concern I have involves a difference in the categorization of plus disease by the experts compared with how these data were managed during analysis. The experts were asked to use a *3-level scale* of "plus," "pre-plus," or "neither," with eyes designated as "cannot determine" excluded from analysis. The data were then analyzed on a *2-level scale*, namely "plus" or "not plus," lumping the "pre-plus" and "neither" together to form the "not plus" group. It is possible that the experts would have judged many borderline cases differently if they had been asked to use a forced 2-level scale, eliminating "pre-plus" as an option. The designation "pre-plus" may have been used as an "on the fence" answer. It is unclear what impact this had on the study results, but it would be interesting to know. Alternatively, this problem could have been avoided by analyzing the data on the same 3-level scale used by the experts.

Though the results of the study are impressive, I find the receiver operating characteristic area under the curve (AUC) reported for each expert confusing. The AUC is supposed to reflect how each expert performed compared to chance. Expert No. 5 had a sensitivity and specificity of 0.308 and 1.000, respectively. Yet despite this dismal sensitivity performance, the AUC was 0.951. In contrast, expert No. 6 had an AUC of only 0.784, despite substantially better overall performance with a sensitivity and specificity of 0.846 and 0.714, respectively. Some clarification on how these data were derived and how they should be interpreted would be of value.

In summary, I congratulate Dr. Chiang and his coworkers for their attempts to standardize the diagnosis of plus disease. I think that this is very important work. With refinement and when used in the proper setting, computer-assisted diagnosis of plus disease may some day show promise and may ultimately the have greatest utility in questionable cases to supplement, but not replace, human analysis.

ACKNOWLEDGMENTS

Funding/Support: None.

Financial Disclosures: None.

DR. ALLAN J. FLACH: I think the original idea of doing this study was fantastic. You described your experts as having been chosen in different ways. Some had participated in previous ROP studies, but others perhaps had only published five papers previously. It would be interesting to determine if the way you chose the expert influenced how they graded the eyes in the study. Secondly, I wonder if the different clinical experiences of some of the experts affected their decisions. For example, if some experts were a little bit more rigid in their interpretation and they had the experience that treating earlier was associated with very good results, even knowing they might be erring on the side of diagnosing a bit too soon, , then it might explain why they behaved the way they did during the testing. If you could maybe talk about each of those three aspects, this would be interesting.

DR. WILLIAM S. TASMAN: No conflict. I want to congratulate you on this presentation. It is an important study and I know it has been a lot of hard work for you. I certainly agree with your third conclusion about more quantitative work being done on plus disease. I think we all appreciate that when you see marked plus disease, it is obviously a sign you have to do something. The management and possible treatment of lesser degrees of plus disease is more controversial. My final comment is actually a suggestion. If possible, I would like you would pursue this further matter, so if a malpractice case arises, the lawyers will sue the computer and not us.

DR. MICHAEL H. GOLDBAUM: I wanted to address the gold standards. When you compare the performance of, let us say, human experts, it is good to have a reliable indicator of the effect, which is different than what is being used by either group. For example, some objective finding, such as whether the retinopathy progressed or what percentage of the eyes developed neovascularization, or something separate from the test, would be a good independent indicator of the disease state.

DR. RAND SPENCER: I would like to echo what David Coats touched on his discussion. Plus disease is a relative condition and this is particularly true regarding the degree of the vessel dilation. It is relative to what the normal vessel caliber would be at that gestational age. Perhaps when we examine the patient clinically, at least subconsciously, this may affect our interpretation of the findings. For example, a child that is 34 weeks post menstrual age will have narrower caliber vessels normally, and therefore, even if they develop plus disease, their caliber vessels maybe smaller than that of a 38 week old baby whose vessels are normally larger. I was wondering if this is a consideration for your future analysis. I congratulate you on a great study.

DR. MICHAEL F. CHIANG: I want to start by thanking everyone who asked questions and raised these points. Dr. Coats and Dr. Goldbaum asked about the reference standard we used. Is it right to use a majority vote of expert opinions? We considered this issue and thought it was fair to do so. We did explore how robust our standards were and considered alternative gold standards. The possibility of every expert having their own unique gold standard, different from everyone else, was our number one alternative. Our second alternative was analyzing as an independent panel, exactly as Dr. Coats suggests, of three people who made a diagnosis after discussion of each photograph and then used that as a gold standard. Our third alternative was to exclude the borderline cases. For example, we had twenty-two images and for a few images the voting results were twelve versus ten; if we exclude the borderline cases and just use the non-borderline images, how much of a difference would that really make? The short answer is that each of those alternative methods for determining the gold standard gave the same answer the vast majority of the time. Dr. Goldbaum makes the important point that we could use some type of outcome based gold standard, but this also raises a much more difficult question. What is the outcome of these babies who are left untreated? How many of them have poor outcomes? I believe that from a practical standpoint this question is difficult to study. After plus disease has developed and we have detect it during our examination, then we usually treat the baby with laser. Therefore, we do not know what the natural history would have been without laser treatment. Dr. Coats raised a question about two level versus three level classification. Specifically, he asked if we biased the study results by allowing graders to use a third "pre-plus" category and then analyzed those eyes as though they were in the same group as the non-plus eyes. In our manuscript, we present the results of both a three level and a two level analysis. In the three level analysis, with evaluation as either "plus," "pre-plus," or "neither" diagnoses, the responses are wider and the mean weighted kappa statistic for each grader compared to other graders ranges from .25 to .55. In response to the question if there is a bias in classification with two versus three categories, or stated differently, "will I respond differently if I am asked to grade as plus, pre-plus, or neither versus nonplus or plus," I do not know the answer. I certainly hope the answer would be "no" and that people would respond in the same manner. In other words, all the "pre-pluses" and "neithers" in the three level classification would be completely equivalent to "not-pluses" in the two-level classification, but question deserves further study. Regarding the ROC curves, an interesting point is raised. Sensitivity and specificity, by definition, refer to using a specific cutoff threshold; however, the benefit of the ROC curve is that you could evaluate performance of a diagnostic test at all levels of cutoff thresholds. For the experts in this study; however, the limitation is that we permitted only three discrete cutoff levels. Whether they categorized as plus, pre-plus, or neither, may have introduced a little bit of bias, particularly because ROC curves for the computer-based system were generated based on continuously varying the value of each parameter. We may be comparing apples versus oranges, and for that reason we chose to present sensitivity, specificity, and the areas under the ROC. I believe that if some experts demonstrate a sensitivity of 30% and a specificity of 57%, then those individuals tend to under-call or under-estimate versus overcall or overestimate the presence of plus disease, for whatever reason. Dr. Flach asked about the issue of expert performance. In other words, does performance of the experts differ depending on what type of expert they are and their experience? The short answer is that we are did not find any evidence of this effect. We queried, for example, how do the mean kappas compare among the experts? If you are a pediatric ophthalmologist versus a retina specialist, or if you were a CRYO-ROP Study Principal Investigator (PI) versus not being a PI, or if you reported extensive experience with the RetCam® photographic system that was used to obtain these images versus if you reported limited or no Retcam experience, was there a statistically significant difference in the mean kappa of that expert compared to others. There was no statistically significance difference in any of these circumstances. I do not believe that this definitively answers your questions, but that is the data that we analyzed regarding this issue. Dr. Spencer asked if we should be looking at the "delta" in vessel caliber between the study eye compared to the normal for age of any given patient rather than the absolute size of the vessels. I agree clinically that this interpretation of vessel size may be important; at the same time, I believe that we should accept the international classification, which defines plus disease at a point in time based on comparison to a standard photograph. The latter was the standard that we used to for these studies. Thank you.