Published in final edited form as:
J AAPOS. 2007 December ; 11(6): 532–540.

Plus Disease in Retinopathy of Prematurity: Pilot Study of Computer-Based and Expert Diagnosis

Rony Gelman, MD, MS[1], **Lei Jiang, BA**[1], **Yunling E. Du, PhD**[2], **M. Elena Martinez-Perez, PhD**[3], **John T. Flynn, MD**[1], and **Michael F. Chiang, MD, MA**[1,4]

1 *Department of Ophthalmology, Columbia University College of Physicians and Surgeons (New York, New York)*

2 *Department of Epidemiology and Population Health, Albert Einstein College of Medicine (New York, New York)*

3 *Department of Computer Science, Institute of Research in Applied Mathematics and Systems, National Autonomous University of Mexico (Mexico City, Mexico)*

4 *Department of Biomedical Informatics, Columbia University College of Physicians and Surgeons (New York, New York)*

Abstract

Purpose—To measure accuracy of plus disease diagnosis by recognized experts in retinopathy of prematurity (ROP), and to conduct a pilot study examining performance of a computer-based image analysis system, Retinal Image multiScale Analysis (RISA).

Methods—Twenty-two ROP experts independently interpreted a set of 34 wide-angle retinal images for presence of plus disease. A reference standard diagnosis based on expert consensus was defined for each image. Images were analyzed by the computer-based system using individual and linear combinations of 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 compared to the reference standard were determined for each expert, as well as for the computer-based system.

Results—Expert sensitivity ranged from 0.308–1.000, specificity ranged from 0.571–1.000, and AUC ranged from 0.784–1.000. Among individual computer system parameters, venular IC had highest AUC (0.853). Among all computer system parameters, the linear combination of arteriolar IC, arteriolar TI, venular IC, venular diameter, and venular TI had highest AUC (0.967), which was greater than that of 18 (81.8%) of 22 experts.

Conclusions—Accuracy of ROP experts for plus disease diagnosis is imperfect. A computer-based image analysis system has potential to diagnose plus disease with high accuracy. Further research

Corresponding author and reprints: Michael F. Chiang, MD, Columbia University Department of Ophthalmology, 635 West 165th Street, Box 92, New York, NY 10032, Tel: 212-342-3440, Fax: 212-305-5962, Email: chiang@dbmi.columbia.edu.

Institution where study conducted: Columbia University Medical Center

The authors have no commercial, proprietary, or financial interest in any of the products or companies described in this article. Dr. Chiang is an unpaid member of the Scientific Advisory Board at Clarity Medical Systems (Pleasanton, CA).

Portions of this manuscript have been presented at the Association for Research in Vision and Ophthalmology (ARVO) 2007 annual meeting, and at the American Ophthalmological Society (AOS) 2007 annual meeting.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

involving RISA system parameter cut-off values from this study are required to fully validate performance of this computer-based system compared to that of human experts.

INTRODUCTION

Retinopathy of prematurity (ROP) is a leading treatable cause of childhood blindness throughout the world.[1–3] Plus disease is a major component of the international classification for ROP,[4–5] and is characterized by arteriolar tortuosity and venous dilation within the posterior pole. The minimum amount of vascular abnormality required for plus disease is defined by a standard photograph, which was selected by expert consensus.[6] Recent major clinical trials have explicitly required ≥2 quadrants of this amount of vascular change for the diagnosis of plus disease.[7–8] Accurate assessment 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.[6,8–9]

Dilated indirect ophthalmoscopy by an experienced examiner is considered the gold standard for ROP management.[5] Because the definition of plus disease is based on a photographic standard with descriptive qualifiers, clinical diagnosis may be heavily subjective. We have previously shown imperfect agreement in plus disease diagnosis among a group of recognized ROP experts reviewing wide-angle images.[10] These factors raise concerns about diagnostic consistency, because errors in classification may result in over-treatment or under-treatment. This has important implications for ROP treatment and research.

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 diagnostic performance of image analysis algorithms compared to a reference standard of dilated ophthalmoscopy by an experienced examiner.[11–14] Yet no previous studies to our knowledge have attempted to measure the accuracy of experts for plus disease diagnosis, and to use this as a basis for analyzing performance of computer-based systems.[*]

The purposes of this paper are to measure the accuracy of plus disease diagnosis by a group of recognized ROP experts, and to describe a pilot study exploring the diagnostic performance of a computer-based image analysis system, Retinal Image multiScale Analysis (RISA).[15–16] This study utilizes the same data set from 22 graders interpreting 34 wide-angle retinal images as our previously-published work involving inter-expert agreement of plus disease diagnosis.[10] Expert and computer-based system performance were compared against a reference standard that was defined by expert consensus.

METHODS

This study was approved by the Institutional Review Board at Columbia University Medical Center, was conducted in compliance with Health Insurance Portability and Accountability Act (HIPAA) guidelines, and included waiver of consent for use of de-identified retinal images. Informed consent was obtained from all expert participants using an online click-through form before any images were displayed.

[*] PubMed search (covering 1980–2007) for “plus disease AND (accuracy OR reliability OR agreement OR reproducibility OR consistency).”

Image Interpretation by Experts

A set of 34 digital retinal images was compiled from premature infants during routine ROP care. Each image was a photograph of the posterior retina obtained using a wide-angle imaging device (RetCam-II; Clarity Medical Systems, Pleasanton, CA). Any visible peripheral ROP disease was cropped-out to decrease bias to graders.

A group of ROP experts was invited to participate. Eligible experts were defined as practicing pediatric ophthalmologists or retina specialists who met ≥1 of three criteria: having been a study center principal investigator for the CRYO-ROP or ETROP studies, having been a certified investigator for either study, or having co-authored ≥5 peer-reviewed ROP manuscripts. Each participant provided one of four mutually-exclusive diagnoses ("plus," "pre-plus," "neither," "cannot determine") for each image, using web-based software developed by the authors. For data analysis, diagnoses of "pre-plus" or "neither" were considered to be "not plus," and diagnoses of "cannot determine" were excluded. Participants were not provided with references or standards for plus disease, although it was assumed that they would be intimately familiar with these definitions. A reference standard diagnosis was defined for each image based on 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. Identities of all participating experts were anonymized by prior agreement.

Computer-Based Image Analysis

All retinal vessels were identified and classified by consensus of four co-authors (RG, LJ, JTF, MFC) as arterioles or venules, and analyzed by the computer-based system.[11] Three continuous-scale system parameters were calculated for all vessels in each image: diameter, tortuosity index (TI), and integrated curvature (IC). Mean *diameter* (pixels) is the total area of the vessel divided by its length; *TI* (unit-less) is the length of the vessel divided by the length of a line segment connecting its end points (Figure 1A); and *IC* (radians/pixel) is defined as the sum of angles along the vessel, normalized by its length (Figure 1B). For a perfectly-straight vessel, TI would have a minimum value of 1, and IC would have a minimum value of 0. The difference between TI and IC is illustrated in Figure 1C.

Data Analysis: Experts

The validity of plus disease diagnosis for each expert was measured using sensitivity and specificity compared to the reference standard. Receiver operating characteristic (ROC) curves, which illustrate discriminative ability of diagnostic tests by plotting sensitivity against (1-specificity) over a range of cutoff thresholds, were constructed for each expert.[17–18] This was done using an ordinal-scale of "plus," "pre-plus," and "neither" diagnoses. ROC areas under the curve (AUC) were calculated nonparametrically as a measure of diagnostic performance for each expert. McNemar's test was used to detect systematic tendencies of each expert to under-call or over-call 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. Median 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.[19] For each individual parameter and linear combination, sensitivity and specificity of the computer-based system was plotted as a function of threshold used to separate "plus" from "not plus." ROC curves were plotted for individual parameters and linear combinations. AUCs were calculated nonparametrically, and were statistically compared among individual parameters, linear combinations, and experts using the Delong approach.[20] Analysis was performed using statistical and computational software (Minitab version 13, Minitab Inc., State College, PA; SPSS version 14, SPSS Inc., Chicago, IL; R programming language version 2.4.0, Free Software Foundation, Boston, MA).

RESULTS

All 34 images were reviewed by 22 experts, for a total of 748 responses. A diagnosis of “cannot determine” was made in 18 (2.4%) of the 748 cases. According to the consensus reference standard, 13 (38.2%) of the 34 images represented “plus” disease, whereas 21 (61.8%) represented “not plus.”

Expert Performance

Table 1 summarizes sensitivity, specificity, and AUC of each expert compared to the reference standard. Thirteen (59.0%) of the 22 experts had sensitivity >80%, and 18 (81.8%) had specificity >80%. According to the McNemar test of bias, 2 (9.1%) experts had a statistically-significant tendency to under-call plus disease (p=0.016 for Expert #1, p=0.004 for Expert #5), and 2 (9.1%) had a statistically-significant tendency to over-call plus disease (p=0.016 for Expert #4, p=0.004 for Expert #21).

AUC for experts ranged from 0.784 to 1.000. Two (9.1%) experts had AUC between 0.701–0.800, 8 (36.4%) experts had AUC between 0.801–0.900, and 12 (54.5%) had AUC between 0.901–1.000. AUCs for all experts reflected diagnostic performance that was significantly better than chance (p<0.006 for all experts).

Computer-Based System Performance

From the 21 images without plus disease according to the reference standard, 67 arterioles (39 temporal and 28 nasal) and 83 venules (47 temporal and 36 nasal) were analyzed. From the 13 images with plus disease, 51 arterioles (22 temporal and 29 nasal) and 55 venules (28 temporal and 27 nasal) were analyzed. The mean (range) number of vessels analyzed per image was 3.47 (1–7) arterioles and 4.03 (2–6) venules.

Figure 2 shows that the median value of each individual system parameter, with the exception of arteriolar diameter, was significantly higher in “plus” images than in “not plus” images. These 5 significant system parameters were modeled with logistic regression, resulting in 26 linear combinations. The two linear combinations with highest accuracy are displayed in Figure 2: linear combination I (arteriolar IC, venular IC, venular diameter); and linear combination II (arteriolar IC, arteriolar TI, venular IC, venular diameter, venular TI).

Sensitivity and specificity curves for plus disease detection based on individual and combined system parameters are plotted in Figure 3. AUCs are displayed in Table 2, along with sensitivity and specificity values for the computer-based system at the intersection of these curves. All individual and linear combinations of parameters had AUCs significantly better than chance (p<0.05), except arteriolar diameter (p=0.559). The AUC of linear combination II was higher than that of any individual parameter (p=0.107 vs. arteriolar IC, p<0.05 vs. all other individual parameters), was higher than that of any other linear combination, and was higher than that of 18 (81.8%) of 22 experts (p<0.05 for 4 of these 18 experts). Linear combinations I and II 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 I or II (Table 1).

DISCUSSION

This is the first study to our knowledge that has systematically measured performance of plus disease diagnosis by multiple human experts and a computer-based image analysis system.[†] Three key findings from this pilot study are that: (1) Performance of plus disease diagnosis by experts is imperfect; (2) Modeling of computer-based system parameters shows diagnostic performance which has potential to be comparable to that of single human experts; and (3) Cutoff values for RISA system parameters from this study may be used for future validation studies involving independent image sets.

Accurate and reliable detection of plus disease has an increasingly critical role in identification of treatment-requiring ROP. This is particularly relevant because the ETROP trial recently determined that presence of plus disease is sufficient for meeting the definition of type-1 ROP, which benefits from early treatment.[8] As shown in Table 1, sensitivity, specificity, and AUC of recognized experts in discriminating plus disease are imperfect. Moreover, the finding that 4 of 22 study participants had statistically-significant tendencies to under- or over-diagnose plus disease supports the notion that experts may have differing subjective interpretations for vascular “dilation” and “tortuosity.” We previously showed that agreement in plus disease diagnosis among ROP experts is imperfect,[10] and the CRYO-ROP study found that 12% of eyes diagnosed as threshold disease by one certified examiner were diagnosed as less-than-threshold during confirmatory examination by a second certified examiner.[21] Taken together, these findings raise concerns about the accuracy and reliability of ROP diagnosis and treatment.

As shown in Figure 2, the computer-based RISA system appears to have strong ability to differentiate between “plus” and “not plus” disease. Based on sensitivity, specificity, and area under ROC curves, the highest measures of validity and diagnostic performance for mean individual blood vessel parameters were achieved with venular integrated curvature (IC), venular diameter, and arteriolar IC (Table 2). This is consistent with previously-published findings involving the RISA system for plus disease detection,[11] and strengthens other work involving automated image analysis for ROP diagnosis.[13–14,22] As shown in Table 2, 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.

Findings from this study suggest that it may be possible for a computer-based system to detect plus disease, with performance that appears comparable to that of individual recognized experts. We note that the sensitivity or specificity of the computer system may be improved by adjusting the cutoff threshold for any given parameter, and that an increase in one inevitably causes a decrease in the other (Figure 3). From a practical standpoint, a computer-based image analysis system could be used to provide real-time decision support for ophthalmologists managing ROP.[23–24] We emphasize that this study relied on a single set of images used by both experts and the computer-based system. Because the computer system cut-off points for individual and linear combinations of parameters were determined using this single set of images, this design biases toward optimizing system performance. For example, the pooled parameter values in Figure 2 could not necessarily be used to exclude or confirm plus disease in an individual infant’s eye. Future studies involving independent image testing sets using these established cut-off points (Figure 3), or involving cross-validation techniques, will be required for complete evaluation and rigorous comparison with expert performance.[25–26]

[†] PubMed search (covering 1980–2007) for “plus disease AND (accuracy OR reliability OR agreement OR reproducibility OR consistency).”

Our results support the notion that interpretation of the standard photographic definition of plus disease may be subjective and variable, even by recognized experts. Development of a quantifiable, objective definition of plus disease could eventually result in improved diagnostic validity and reliability. This would be analogous to widely-used methods for computer-based interpretation of electrocardiograms and Papanicolaou smears.[27–28] It would be feasible to create a quantitative definition of "plus disease," using methods described in this study, by selecting computer system parameters at cutoff thresholds that meet or exceed the performance of most experts. The current photographic standard was developed before the CRYO-ROP study, and has been successfully used in numerous clinical trials and shown to have prognostic significance, despite its potential limitations.[6,8] Creation of an entirely new quantitative definition from first principles would likely require either prospective validation in a clinical study, or retrospective validation using a library of retinal images from infants with ROP whose untreated natural histories are known. Finally, we note that different computer-based image analysis systems have been applied for detection of plus disease,[11–14,29–30] and future studies comparing these various algorithms for quantifying curvature of retinal vessels may be informative.

The reference standard diagnosis in this study was defined as majority vote among 22 ROP experts who interpreted images independently. It could be argued that this majority-vote reference standard may be influenced by small differences of expert opinion in borderline cases. To investigate this possibility by eliminating photographs that were apparently the most "borderline," we calculated results using alternative scenarios in which the reference standard required consensus by larger groups of experts. Sensitivity/specificity for diagnosis of plus disease ranged from 33.3%/57.1% to 100%/100% using a reference standard requiring agreement by ≥13/22 experts (1 borderline image excluded), and from 36.4%/57.1% to 100%/100% using a reference standard requiring agreement by ≥14/22 experts (2 borderline images excluded). We believe this suggests that our findings regarding expert validity are robust; in fact, it is precisely these borderline cases in which an objective computer-based diagnosis system might provide most added-value for clinicians. A related issue is that experts in this study were compared against a reference standard which each of them contributed toward establishing. To explore the impact of potential circularity, an alternative reference standard was defined separately for each expert as the majority-vote of all *other* experts for every image – this provided the same diagnosis as the majority-vote reference standard in all cases. Additional studies involving alternative methods for obtaining consensus reference standards, such as the Delphi technique,[31] may be informative.

While designing this study, the decision was made not to provide specific images or instructions to experts regarding definitions of "plus" or "pre-plus" disease. For example, recent ROP clinical trials have required that the minimum level of vascular change must be present in ≥2 quadrants, but participants were not explicitly instructed to conform to this guideline. It is possible that some disagreement among experts may have been prevented if we had provided representative images and definitions. However, we believe that our study design better simulates real-world diagnosis by expert ophthalmologists, who are presumably already very familiar with these definitions and typically do not refer to the standard "plus disease" photograph while making clinical decisions at the bedside. If we were to have provided instructions or example photographs to experts for comparison purposes, we felt that this would introduce a significant confounding factor – i.e. we would be testing the ability of experts to visually match study images with the instructional images, rather than testing the expert's actual diagnostic performance under simulated real-world conditions. Although we believe that the objectivity and standardization of image analysis systems may be inherently superior to that attainable by experts, we certainly acknowledge that some nuances of clinical care could be better understood by physicians than by computer systems.

Additional study limitations should be noted: (1) Each image was a wide-angle posterior pole photograph with peripheral ROP cropped-out, and without any other patient information. Although plus disease is defined only by a standard photograph of the central retina,[6] examiners might be influenced by peripheral retinal findings or patient demographics while diagnosing plus disease. (2) Study images were captured using a wide-angle camera, whereas indirect ophthalmoscopy provides a smaller field-of-view. It is not clear whether review of these wide-angle images could bias toward systematic over-diagnosis or under-diagnosis by experts, compared to ophthalmoscopic examinations.[10] Similarly, it is possible that increased intraocular pressure from a contact retinal camera could alter the appearance of retinal vessels compared to ophthalmoscopy. (3) System parameters were computed using all identifiable vessels in each image. It is possible that including all vessels could cause a wash-out 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. (4) Experts graded images using an ordinal-scale, whereas the computer-based system utilized a continuous-scale for each parameter. This decision was made because experts in real-world situations use an ordinal-scale ("plus," "pre-plus," "neither"), and there is no numeric scale available for describing vascular appearance in ROP. This may create bias in comparing AUCs between experts and the computer-based system, although previous work has suggested no significant differences in accuracies between discrete and continuous scales in image-evaluation studies.[32] ROC analysis provides an advantage over reporting only sensitivity and specificity, because the former reflects performance of the overall test whereas the latter reflects performance at a single particular cut-off point. (5) Experts were given the option of diagnosing plus disease as "cannot determine," and these responses were excluded from further analysis. However, we do not feel that this had any impact on the consensus reference standard because only 18 (2.4%) of the 748 responses from 22 experts were "cannot determine," and because none of the 34 images received more than two "cannot determine" responses.

Based on results from this study, further refinement and validation of computer-based diagnosis systems for plus disease may be performed, with comparison to expert interpretation. This may provide opportunities to improve clinical care, support new health care delivery strategies such as telemedicine,[33–36] and enhance research infrastructure.

Acknowledgements

Supported by a Career Development Award from Research to Prevent Blindness (MFC), and by grant EY13972 from the National Eye Institute of the National Institutes of Health (MFC). 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. [PubMed: 11973241]
2. Steinkuller PG, Du L, Gilbert C, et al. Childhood blindness. J AAPOS 1999;3:26–32. [PubMed: 10071898]
3. Gilbert C, Foster A. Childhood blindness in the context of VISION 2020: the right to sight. Bull World Health Organ 2001;79:227–32. [PubMed: 11285667]
4. Committee for the classification of retinopathy of prematurity. An international classification of retinopathy of prematurity. Arch Ophthalmol 1984;102:1130–1134. [PubMed: 6547831]
5. International committee for the classification of retinopathy of prematurity. The international classification of retinopathy of prematurity revisited. Arch Ophthalmol 2005;123:991–999. [PubMed: 16009843]

6. Cryotherapy for Retinopathy of Prematurity Cooperative Group. Multicenter trial of cryotherapy for retinopathy of prematurity: preliminary results. Arch Ophthalmol 1988;106:471–479. [PubMed: 2895630]

7. 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. [PubMed: 10654946]

8. Early Treatment for Retinopathy of Prematurity Cooperative Group. Revised indications for the treatment of retinopathy of prematurity: results of the early treatment for retinopathy of prematurity randomized trial. Arch Ophthalmol 2003;121:1684–1694. [PubMed: 14662586]

9. Palmer EA, Hardy RJ, Dobson V, et al. 15-year outcomes following threshold retinopathy of prematurity: final results from the multicenter trial of cryotherapy for retinopathy of prematurity. Arch Ophthalmol 2005;123:311–8. [PubMed: 15767472]

10. Chiang MF, Jiang L, Gelman R, et al. Inter-Expert Agreement of Plus Disease Diagnosis in Retinopathy of Prematurity. Arch Ophthalmol 2007;125:875–80. [PubMed: 17620564]

11. Gelman R, Martinez-Perez ME, Vanderveen DK, et al. Diagnosis of plus disease in retinopathy of prematurity using Retinal Image multiScale Analysis. Invest Ophthalmol Vis Sci 2005;46:4734–4738. [PubMed: 16303973]

12. Swanson C, Cocker KD, Parker KH, et al. Semiautomated computer analysis of vessel growth in preterm infants without and with ROP. Br J Ophthalmol 2003;87:1474–1477. [PubMed: 14660456]

13. Wallace DK, Jomier J, Aylward WR, Landers MB. Computer-automated quantification of plus disease in retinopathy of prematurity. J AAPOS 2003;7:126–130. [PubMed: 12736626]

14. Wallace DK, Zhao Z, Freedman SF. A pilot study using “ROPtool” to quantify plus disease in retinopathy of prematurity. J AAPOS 2007;11:381–7. [PubMed: 17532238]

15. Martinez-Perez ME, Hughes AD, Thom SA, et al. Segmentation of blood vessels from red-free and fluorescein retinal images . Med Image Anal 2007;11:47–61. [PubMed: 17204445]

16. 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. [PubMed: 12148830]

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. [PubMed: 16613919]

18. Sox, HC.; Blatt, MA.; Higgins, MC.; Marton, KI. Medical Decision Making. Boston: Butterworth-Heinemann; 1988.

19. Liu A, Schisterman EF, Zhu Y. On linear combinations of biomarkers to improve diagnostic accuracy. Stat Med 2005;24:37–47. [PubMed: 15515132]

20. Hanley JA, Hajian-Tilake KO. Sampling variability of nonparametric estimates of the areas under receiver operating characteristics curves: an update. Acad Radiol 1997;4:49–58. [PubMed: 9040870]

21. 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 Opthlamol 2002;120:1470–6.

22. Wallace DK, Kylstra JA, Chestnutt DA. Prognostic significance of vascular dilation and tortuosity insufficient for plus disease in retinopathy of prematurity. J AAPOS 2000;4:224–229. [PubMed: 10951298]

23. 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. [PubMed: 16221942]

24. 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. [PubMed: 15767266]

25. Rosado B, Menzies S, Harbauer A, et al. Accuracy of computer diagnosis of melanoma: a quantitative meta-analysis. Arch Dermatol 2003;139:361–367. [PubMed: 12622631]

26. Aleynikov S, Micheli-Rzanakou E. Classification of retinal damage by a neural network based system. J Med Syst 1998;22:129–36. [PubMed: 9604780]

27. Hongo RH, Goldschlager N. Status of computerized electrocardiography. Cardiol Clin 2006;24:491–504. [PubMed: 16939838]

28. Ku NN. Automated Papanicolaou smear analysis as a screening tool for female lower genital tract malignancies. Curr Opin Obstet Gynecol 1999;11:41–43. [PubMed: 10047962]

29. Capowski JJ, Kylstra JA, Freedman SF. A numeric index based on spatial frequency for the tortuosity of retinal vessels and its application to plus disease in retinopathy of prematurity. Retina 1995;15:490–500. [PubMed: 8747443]

30. Freedman SF, Kylstra JA, Capowski JJ, Realini TD, Rich C, Hunt D. Observer sensitivity to retinal vessel diameter and tortuosity in retinopathy of prematurity: a model system. J Pediatr Ophthalmol Strabismus 1996;33:248–54. [PubMed: 8827562]

31. Kors JA, Sittig AC, van Bemmel JH. The Delphi method to validate diagnostic knowledge in computerized ECG interpretation. Methods of Med 1990;29:44–50.

32. Rockette HE, Gur D, Metz CE. The use of continuous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging techniques. Invest Radiol 1992;27:169–172. [PubMed: 1601610]

33. Chiang MF, Keenan JD, Starren J, et al. Accuracy and reliability of remote retinopathy of prematurity diagnosis. Arch Ophthalmol 2006;124:322–327. [PubMed: 16534051]

34. Chiang MF, Wang L, Busuioc M, et al. Telemedical diagnosis of retinopathy of prematurity: accuracy, reliability, and image quality. Arch Ophtahlmol. In press

35. Jackson KM, Scott KE, Zivin JG, et al. Cost-utility analysis of telemedicine and ophthalmoscopy for retinopathy of prematurity management. Arch Ophthalmol. In press

36. Scott KE, Kim DY, Wang L, et al. Telemedical diagnosis of retinopathy of prematurity: inter-expert agreement between ophthalmoscopic and image-based examinations. In review

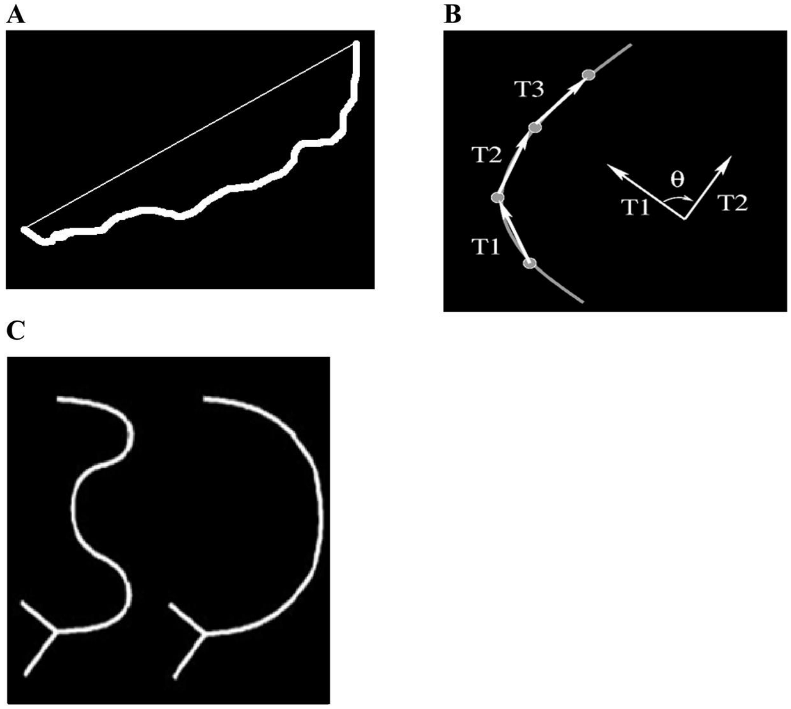

Figure 1. Parameters of computer-based Retinal Image multiScale Analysis (RISA) system

(A) Tortuosity index (TI) is defined as the length of a vessel divided by a line segment connecting the end points. (B) Integrated curvature (IC) is computed by forming vectors along the vessel, calculating cumulative sum of angles θ 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 one. (C) TI vs. IC. Departures from linearity, such as the vessel depicted on right, are captured both by IC and TI. However, multiple changes in vessel depicted on left are better captured by IC. TI is similar for the two shapes (5% difference), whereas IC is higher for the vessel depicted on left (50% difference).

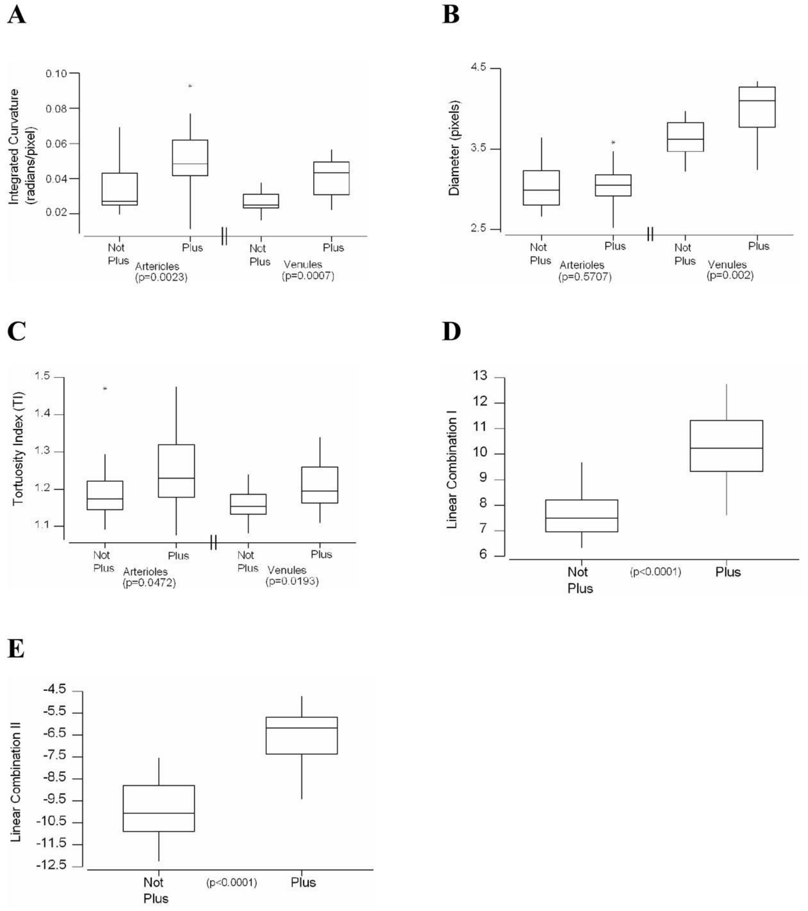

Figure 2. Box plots of computer-based system parameter values in images with “not plus” compared to “plus,” based on reference standard diagnosis

Individual parameters of (A) integrated curvature (IC), (B) diameter and (C) tortuosity index (TI) for arterioles and venules; as well as combined parameters of (D) Linear combination I (arteriolar IC, venular IC, and venular diameter), and (E) Linear Combination II (arteriolar IC, arteriolar TI, venular IC, venular diameter, and venular TI) are displayed. Boxes represent the 25th, 50th, and 75th percentile values of parameters. Whiskers represent 10th and 90th percentile values. Asterisks represent outlier values. P-values indicated are for Mann-Whitney tests between “not plus” and “plus” groups.

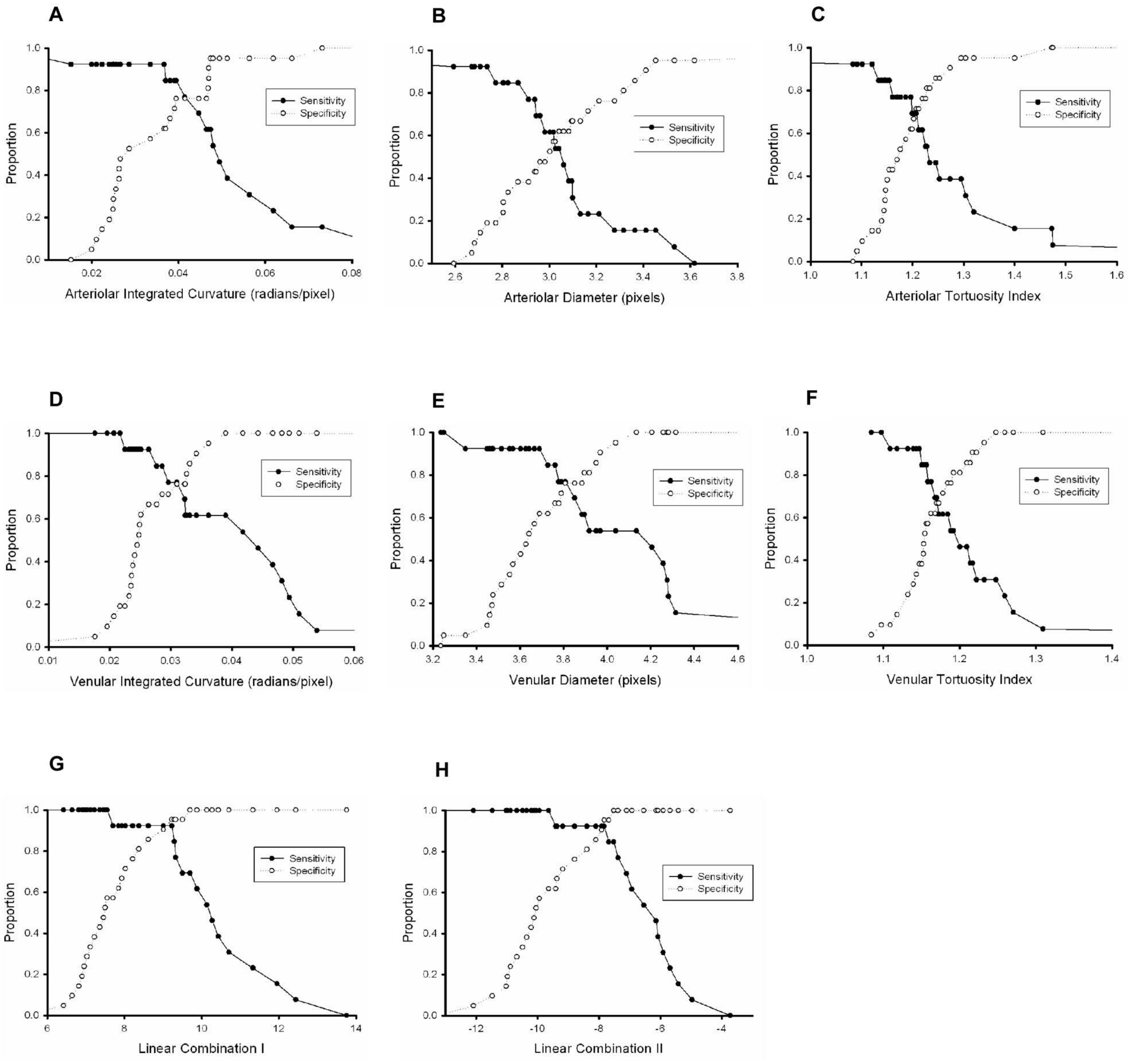

Figure 3. Sensitivity and specificity of individual computer system parameters and linear combinations of parameters for plus disease diagnosis, compared to the reference standard of majority vote among 22 recognized ROP experts

Curves are displayed as a function of parameter cutoff criteria for detection of plus disease: (A) arteriolar integrated curvature (IC), (B) arteriolar diameter, (C) arteriolar tortuosity index (TI), (D) venular IC, (E) venular diameter, (F) venular TI, (G) Linear Combination I (arteriolar IC, venular IC and venular diameter) and (H) Linear Combination II (arteriolar IC, arteriolar TI, venular IC, venular diameter, and venular TI).

Table 1

ROP expert 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.

Expert #	Sensitivity	Specificity	AUC (SE*)
1	0.462	1.000	0.859 (0.063)
2	0.615	1.000	0.982 (0.018)
3	0.615	1.000	0.890 (0.057)
4	1.000	0.667	0.833 (0.070)
5	0.308	1.000	0.951 (0.035)
6	0.846	0.714	0.784 (0.082)
7	0.846	0.952	0.921 (0.052)
8	0.923	0.905	0.921 (0.052)
9	0.923	0.905	0.921 (0.052)
10	0.778	1.000	0.918 (0.065)
11	0.538	0.857	0.885 (0.061)
12	0.769	1.000	0.956 (0.032)
13	0.692	1.000	0.934 (0.041)
14	1.000	1.000	1.000 (0.000)
15	0.750	0.952	0.899 (0.058)
16	0.923	0.810	0.879 (0.061)
17	0.846	1.000	0.967 (0.029)
18	1.000	0.750	0.875 (0.080)
19	0.923	0.857	0.896 (0.059)
20	1.000	0.952	0.976 (0.028)
21	1.000	0.571	0.786 (0.077)
22	1.000	0.952	0.976 (0.028)

*SE, Standard error.

Table 2

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 Integrated Curvature	0.760	0.760	0.817 (0.085)	0.002
Arteriolar Diameter	0.590	0.590	0.560 (0.102)	0.559
Arteriolar Tortuosity Index	0.700	0.700	0.707 (0.099)	0.045
Venular Integrated Curvature	0.760	0.760	0.853 (0.072)	0.001
Venular Diameter	0.760	0.760	0.821 (0.082)	0.002
Venular Tortuosity Index	0.640	0.640	0.744 (0.089)	0.018
Linear Combination I[§]	0.938	0.938	0.956 (0.036)	<0.001
Linear Combination II	0.938	0.938	0.967 (0.0310)	<0.001

[*] For the computer-based system, values at the intersection points of the sensitivity and specificity curves in Figure 3 are displayed in this table.

[†] SE, Standard Error.

[‡] Null hypothesis: true area = 0.5 at level of significance 0.05.

[§] Comprising arteriolar integrated curvature (IC), venular IC, and venular diameter.

[||] Comprising arteriolar integrated curvature (IC), arteriolar tortuosity index (TI), venular IC, venular diameter, and venular TI.