Published in final edited form as:

Ophthalmol Retina. 2020 October ; 4(10): 1016–1021. doi:10.1016/j.oret.2020.04.022.

Real world variability in plus disease identified using a deep learning-based retinopathy of prematurity severity scale

Rene Y Choi, MD, PhD¹, James M Brown, PhD^{2,3}, Jayashree Kalpathy-Cramer, PhD², RV Paul Chan, MD⁴, Susan Ostmo, MPH¹, Michael F Chiang, MD^{1,5}, J Peter Campbell, MD, MPH¹ on behalf of the Imaging and Informatics in Retinopathy of Prematurity (i-ROP) Consortium

¹Department of Ophthalmology, Casey Eye Institute, Oregon Health & Science University (OHSU), Portland, Oregon.

²Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Charlestown, Massachusetts.

³Department of Computer Science, University of Lincoln, Lincoln, UK

⁴Department of Ophthalmology and Visual Sciences, Illinois Eye and Ear Infirmary, University of Illinois at Chicago, Chicago, Illinois.

⁵Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon.

Abstract

PURPOSE: Retinopathy of prematurity is a leading cause of childhood blindness worldwide, but clinical diagnosis is subjective, which leads to real world treatment differences. Our goal is to determine objective differences in the diagnosis of plus disease between clinicians using an automated retinopathy of prematurity vascular severity score.

DESIGN: This retrospective cohort study used data from the Imaging & Informatics in ROP consortium, which comprises 8 tertiary care centers in North America. Fundus photographs of all babies undergoing retinopathy of prematurity screening exams between July 1, 2011, to December 31, 2016 were obtained.

PARTICIPANTS: Infants meeting retinopathy of prematurity screening criteria who were diagnosed with plus disease and treatment initiated by an examining physician based on ophthalmoscopic exam.

METHODS: A retinopathy of prematurity severity score (1–9) was generated for each image using a deep learning algorithm.

Corresponding Author: J Peter Campbell, MD, Casey Eye Institute, Oregon Health & Science University, 3375 SW Terwilliger Blvd, Portland OR 97239, Phone: 503-494-3667, campbelp@ohsu.edu.

Conflict of Interest:

Michael F. Chiang is an unpaid member of the Scientific Advisory Board for Clarity Medical Systems, a Consultant for Novartis, and has equity in Intelereitina. R. V. Paul Chan is on the Scientific Advisory Board for Phoenix Technology Group, and a Consultant for Novartis and Alcon. Drs. Chiang, Kalpathy-Cramer, and Campbell receive research funding from Genentech.

MAIN OUTCOME MEASURES: The mean, median, and range of retinopathy of prematurity vascular severity scores overall and for each examiner when the diagnosis of plus disease was made.

RESULTS: A total of 5255 clinical examinations in 871 babies were analyzed. Of these, 168 eyes were diagnosed with plus disease by 11 different examiners and were included in the study. The mean ($\pm$ standard deviation) vascular severity score for patients diagnosed with plus disease was 7.4 ($\pm$ 1.9), median was 8.5 (interquartile range [IQR] 5.8 – 8.9), and range went from 1.1 – 9.0. Within some examiners, there was variability in the level of vascular severity diagnosed as plus disease, and one examiner routinely diagnosed plus disease in patients with less severe disease than the others ($P < .01$).

CONCLUSIONS: We observed variability both between and within examiners in the real-world diagnosis of plus disease using deep learning. Prospective evaluation of clinical trial data using objective measurement of vascular severity may help better define the minimum necessary level of vascular severity for the diagnosis of plus disease, or how other clinical features such as zone, stage, and extent of peripheral disease ought to be incorporated in treatment decisions.

Introduction

Retinopathy of prematurity (ROP) is a vision-threatening disease associated with abnormal retinal vascular development that occurs in premature infants.¹ The multicenter Cryotherapy for ROP (CRYO-ROP) and Early Treatment for ROP (ETROP) clinical trials determined that appropriate identification and timely treatment of ROP can decrease the risk of adverse outcomes and vision loss.^{2–4} The results of these studies demonstrated that among all parameters in the International Classification of ROP (ICROP), presence of plus disease is the most important feature for identifying infants with severe disease who require treatment with laser photocoagulation or intravitreal injection of anti-vascular endothelial growth factor (VEGF) agents to prevent blindness.⁵

There are multiple lines of evidence, however, that there is real-world treatment variability due to differences in plus disease diagnosis.^{6–9} By definition, the diagnosis of plus disease is based on comparison of a patient's retinal arterial tortuosity and venous dilation with a standard narrow-angle published photograph selected by expert consensus during the 1980s.⁴ When experts have been asked to diagnose plus disease in wide-field images, as more typically used in ROP screening, agreement has been variable.^{10–13} Even in rigorous clinical trials, there is evidence that experts treat at different levels of severity. However, this has not been evaluated quantitatively.^{7, 9}

There have been multiple explanations for this diagnostic variability in plus disease diagnosis among experts including field of view, training, magnification, and changes over time.^{6, 8, 10} Some of the variability may also be explained by subjective differences among experts in the diagnostic cut-points along a continuous spectrum from mild to severe disease (e.g., between plus disease vs. pre-plus disease, between pre-plus vs. normal).^{6–8, 10}

Deep learning (DL) has emerged as method for automated and objective image-based disease diagnosis and has shown promise throughout ophthalmology and medical imaging

applications in general. In ROP, Brown et al demonstrated that a DL system can provide expert-level plus disease classification automatically and objectively.^{14, 15} In what may represent a generalizable method for other ophthalmic diseases, the same DL classifier was then used to develop a quantitative severity score (from 1–9) reflecting the spectrum of vascular severity in ROP, which correlated not only with expert classification of plus disease, but with the overall ICROP classification based on a single fundus image. This ROP vascular severity score has further shown promise for quantitative disease monitoring over time.¹⁶ In this study, we evaluated whether objective differences in the diagnosis of plus disease between experts could be observed in real world clinical care.

Methods

This study was approved by the Institutional Review Board (IRB) at the coordinating center (Oregon Health & Science University) and by each of the 8 participating institutions in the United States (Columbia University, Cornell University, University of Illinois at Chicago, William Beaumont Hospital, Children's Hospital Los Angeles, Cedars-Sinai Medical Center, University of Miami). All institutions abided by the tenets of the Declaration of Helsinki and written informed consent was obtained from parents of all infants enrolled.

Data capture

A full clinical evaluation was performed, as part of routine care, by an experienced ophthalmologist at each study center. Wide-angle retinal images were captured after each clinical exam, unless parents declined to consent or the patient was clinically unstable. Exams performed between July 2011 and December 2016 were reviewed for this study. All images were obtained using a commercially-available camera (RetCam; Natus Medical Incorporated, Pleasanton, CA) by an ophthalmologist, nurse, or photographer.

Image grading and reference standard diagnosis

Using methods previously published, a reference standard diagnosis (RSD) that combined findings from ophthalmoscopic and image-based examinations by multiple experts was obtained for each image set.¹⁷ As part of the RSD process, each set of study images was independently graded by 3 graders for zone (I, II, III), stage (no disease, 1, 2, 3, 4, 5), and vascular severity (normal, pre-plus, plus). Each grader also assessed image quality (acceptable, possibly acceptable, or not acceptable for diagnosis).

Inclusion criteria and quantitative analysis using deep learning

For this specific analysis, our inclusion criteria were all patients diagnosed with plus disease by examining clinicians in the i-ROP cohort study.¹⁸ Patients were excluded if the images of that examination were graded “unacceptable for diagnosis” by 2 out of 3 image graders, in order to exclude images that may have been difficult to analyze using quantitative methods. Patients were also excluded if diagnosed with stage 4 or 5 ROP without prior treatment, because the purpose of this study was to analyze diagnostic discrepancies to prevent retinal detachment.

The i-ROP DL system was used to analyze each image in the dataset and classify the probability of plus disease, pre-plus disease, and no plus disease using previously published methods based on a single posterior-pole fundus image and trained using the RSD.¹⁴ Based on these three probabilities, an automated ROP vascular severity score was then assigned to each image from 1 (normal retinal vasculature) to 9 (worst retinal vasculature, e.g., severe plus disease) using methods previously described: $(1 \times \text{probability of no disease} + 5 \times \text{probability of pre-plus disease} + 9 \times \text{probability of plus disease})$.¹⁵

Data analysis

The mean, median, and range of ROP vascular severity scores were calculated for all eye exams in the study. A multi-level linear regression analysis was conducted to determine whether there were differences between examiners in the level of severity diagnosed as plus disease. A Barlett's test was used to assess for homogeneity of variances in the level of severity diagnosed as plus disease among examiners. Statistical analysis was performed using free software (R Foundation for Statistical Computing, version 3.5.3, R Core Team, Vienna, Austria). A P-value less than 0.01 was considered statistically significant.

Results

5,255 eye examinations from 871 infants were reviewed for eligibility in this study. Of these, 168 (3.2%) eyes examined by 11 i-ROP examiners were diagnosed with plus disease. All 168 eyes received treatment, which consisted of either laser photocoagulation or intravitreal anti-vascular endothelial growth factor (VEGF) injection. Thirty-seven (22%) eyes received anti-VEGF and 131 (78%) eyes received laser photocoagulation. The mean ($\pm$ standard deviation [SD]) birth weight of these patients with treatment-requiring disease was 659 [$\pm$ 166] grams, and the mean gestational age was 24.6 [$\pm$ 1.2] weeks. The mean post-menstrual age at time of treatment was 36.4 [$\pm$ 3.1] weeks. The mean birth weight, gestational age, and post-menstrual age grouped by examiner can be seen in Table 1. There was no difference for mean birth weight, gestational age, or post-menstrual age of the babies examined by each clinician.

Figure 1 shows the distribution of all the ROP vascular severity scores for the 168 eyes diagnosed with plus disease in this study by examiner. The mean [$\pm$ SD] ROP quantitative vascular severity score of eyes with plus disease was 7.4 [$\pm$ 1.9], and the median was 8.5 (interquartile range 5.8–8.9). The range of vascular severity scores was from 1.1 to 9.0. The examiners with the highest mean ($\pm$ SD) vascular severity scores were examiners J and K at 8.9 ± 0.1 and 8.9 ± 0.2 . The examiner with the lowest mean vascular severity score was examiner I at 3.6 ± 1.8 . A multi-level linear regression analysis was conducted to determine whether there were differences between examiners in the level of vascular severity score when making the diagnosis of plus disease. Examiner I had a statistically significant lower mean ROP vascular severity score than the other examiners ($P < 0.01$). Figure 2 displays an example of a baby diagnosed with plus disease by examiner I (Figure 2A) with a severity score of 1.1 compared to a baby diagnosed by examiner D with a severity score of 9.0 (Figure 2B). The variances of vascular severity among examiners were found to be statistically significantly different from each other ($P < 0.01$).

Discussion

We report differences between and within examiners in the quantitative level of vascular abnormality in patients diagnosed with plus disease and treated in routine clinical care. The data were obtained as part of the i-ROP cohort study, but treatment decisions were made in real time at the bedside by the treating physicians. The key findings to this study are: (1) Quantitative evaluation of vascular severity can be used to compare relative disease severity in plus disease diagnosis between clinicians. (2) Within each clinician, there is a wide variance of vascular severity scores when making the diagnosis of plus disease, suggesting other factors than the absolute level of vascular severity may play a role in treatment decisions. (3) Between examiners, there are different average vascular severity scores when diagnosing plus disease in ROP, which suggests that clinicians may have different “internal cut points” to initiate treatment.

The first key finding is that a quantitative ROP vascular severity score can be used to characterize differences in treatment thresholds among examiners when making the diagnosis of plus disease. Plus disease diagnosis is one example of a more general problem in ophthalmology, where many diseases are diagnosed and staged based on qualitative features and subjectively quantified levels of severity along a continuum (e.g., “1+ cell in the anterior chamber”, “2+ nuclear sclerosis”), or where despite the presence of standard photographs (e.g., 10A in diabetic retinopathy) widespread diagnostic variability exists among clinicians, even on the same diagnostic task.¹⁹ The use of deep learning to develop quantitative disease scales may thus be a generalizable method to standardize diagnostic agreement for other diseases. In the case of ROP, there is growing evidence that this is a widespread problem that leads to real world treatment differences.^{6, 7, 10, 17, 20, 21} Furthermore, this diagnostic variability limits our ability to interpret clinical trial data, since plus disease remains the most important clinical parameter determining the need for treatment and these subjective treatment differences were not accounted for in the clinical trials. A technology such as this may be used to standardize disease level for future clinical trials, as well as ensure that clinical training regarding the level of vascular severity that ought to be considered plus disease is consistent across centers.

The second key finding is that some examiners demonstrated wide variability in the level of vascular abnormality diagnosed as plus disease, whereas others were very consistent. This suggests that for some examiners, factors other than the level of vascular severity may play a role when making the decision to treat. When examining infants, many physicians describe a gestalt or sense of overall disease severity based on his/her clinical judgement.¹¹ These “other factors” may be responsible for initiating treatment with less severe vascular severity. Gupta et al previously demonstrated that examiners occasionally treat eyes that do not technically meet criteria for type I ROP, for reasons such as structural changes prompting concern for future anatomic complications (e.g., retinal traction), active ROP disease at an advanced postmenstrual age, logistical and systemic concerns (e.g., treatment in the fellow eye), and diagnostic concerns (e.g., vitreous hemorrhage).¹⁸ It is also possible that because the diagnosis of plus disease is made in conjunction with a simultaneous assessment of zone and stage (either by ophthalmoscopy or telemedicine), these other features may influence clinician perception of vascular severity. A previous analysis of diagnostic discrepancies in

ROP demonstrated that clinicians who diagnosed stage 3 were more likely to diagnose pre-plus or plus in the same image as examiners who diagnosed the same eye with stage 2. This suggests that other factors may in some cases influence clinician perception of “plus disease” beyond the quantitative degree of vascular severity.²⁰

The third key finding is that there was variability in diagnostic cutoff for plus disease between examiners (Figure 1). This provides an additional level of evidence that examiners have systematically biased “internal cut points” when making the diagnosis of plus disease. Figure 2 gives two examples of eyes diagnosed with plus disease and treated despite significant differences in the degrees of arterial tortuosity and dilation, and the vascular severity score. Based on the prior work in this area, it is not surprising that the same level of quantitative vascular severity may be treated differently by different physicians; however, it raises legitimacy questions about how to translate clinical trial data.^{2, 3, 6–9, 22, 23} Given that recent work has demonstrated that the level of vascular severity on a posterior pole photograph and the ophthalmoscopic classification are closely related,¹⁶ it is not clear whether optimizing the cutoff for type 1 ROP might be as simple as standardizing the “right” level of plus disease quantitatively, or whether factors such as the extent of stage 3 (for which there was no minimum described after ETROP), may again become relevant.^{2, 6–10, 20–22} In addition, clinicians often describe other features, such as the postmenstrual age, as well as the rapidity of clinical progression in determining the need for treatment, however these features have not been formally evaluated within a clinical trial, nor are they officially part of ROP treatment criteria. That said, prior work using this severity scale has suggested that assessing quantitative change in vascular severity may indeed better identify eyes progressing to severe ROP; however, to date it has not been possible to formally evaluate this disease feature. Quantitative disease monitoring using a vascular severity score may have clinical utility for identifying disease that is progressively worsening (rather than spontaneously regressing), as well as primary treatment failure after laser, and disease reactivation with anti-VEGF.^{16, 24}

There were a number of limitations to our study. This study was limited to 11 examiners, who work together as part of the i-ROP research consortium. Thus, the findings may not generalize to other groups of clinicians. However, we would expect the real-world variability to be at least as wide as this group that works together. In terms of the vascular severity score, it is unknown how image quality, field of view, and type of camera system influence the vascular severity score. To avoid the issue of eyes treated outside of Type I treatment guidelines, we did not examine eyes treated without the diagnosis of plus disease. The variability among all treated infants may be wider. Finally, the vascular severity score described here is a research tool that is not clinically available and thus the real-world impact of this work is theoretical, and would require prospective clinical validation prior to impacting clinical disease management in the future.

In conclusion, diseases such as plus disease that are diagnosed subjectively along a continuum can demonstrate variability in diagnostic cut points in the real world. For plus disease, this has been demonstrated multiple times in research studies and increasingly it is being recognized in clinical trials and in real world care.²⁰ Unfortunately, because subjective criteria were utilized in all existing clinical trials, many questions remain unanswered

regarding the optimum level of vascular severity and ROP phenotypes for which the risk of observation outweighs the risk of treatment. The potential use of a deep learning-based classifier for quantitative disease evaluation may lead to improved standardization of plus disease in research, teaching, and patient care. Finally, the concept that deep learning may be used to determine optimum disease thresholds for clinical intervention, and standardize clinical diagnosis and treatment, will likely generalize to other ophthalmic diseases in the future.

Acknowledgments

On Behalf of the Imaging & Informatics in Retinopathy of Prematurity (i-ROP) research consortium: Oregon Health & Science University (Portland, OR): Michael F. Chiang, MD, Susan Ostmo, MS, Sang Jin Kim, MD, PhD, Kemal Sonmez, PhD, J. Peter Campbell, MD, MPH. University of Illinois at Chicago (Chicago, IL): RV Paul Chan, MD, Karyn Jonas, RN. Columbia University (New York, NY): Jason Horowitz, MD, Osode Coki, RN, Cheryl-Ann Eccles, RN, Leora Sarna, RN. Weill Cornell Medical College (New York, NY): Anton Orlin, MD. Bascom Palmer Eye Institute (Miami, FL): Audina Berrocal, MD, Catherin Negron, BA. William Beaumont Hospital (Royal Oak, MI): Kimberly Denser, MD, Kristi Cumming, RN, Tammy Osentoski, RN, Tammy Check, RN, Mary Zajechowski, RN. Children's Hospital Los Angeles (Los Angeles, CA): Thomas Lee, MD, Evan Kruger, BA, Kathryn McGovern, MPH. Cedars Sinai Hospital (Los Angeles, CA): Charles Simmons, MD, Raghu Murthy, MD, Sharon Galvis, NNP. LA Biomedical Research Institute (Los Angeles, CA): Jerome Rotter, MD, Ida Chen, PhD, Xiaohui Li, MD, Kent Taylor, PhD, Kaye Roll, RN. Massachusetts General Hospital (Boston, MA): Jayashree Kalpathy-Cramer, PhD. Northeastern University (Boston, MA): Deniz Erdogmus, PhD, Stratis Ioannidis, PhD. Asociacion para Evitar la Ceguera en Mexico (APEC) (Mexico City): Maria Ana Martinez-Castellanos, MD, Samantha Salinas-Longoria, MD, Rafael Romero, MD, Andrea Arriola, MD, Francisco Olguin-Manriquez, MD, Miroslava Meraz-Gutierrez, MD, Carlos M. Dulanto-Reinoso, MD, Cristina Montero-Mendoza, MD.

Financial Support: This project was supported by grants R01EY19474, K12 EY027720, and P30EY10572 from the National Institutes of Health (Bethesda, MD), by grants SCH-1622679, SCH-1622542, & SCH-1622536 from the National Science Foundation (Arlington, VA), and by unrestricted departmental funding from Research to Prevent Blindness (New York, NY). The sponsors or funding organizations had no role in the design or conduct of this research.

References

1. Good WV, Hardy RJ, Dobson V et al. The incidence and course of retinopathy of prematurity: findings from the early treatment for retinopathy of prematurity study. *Pediatrics*. 2005; 116(1): 15–23. [PubMed: 15995025]
2. Multicenter trial of cryotherapy for retinopathy of prematurity. Preliminary results. Cryotherapy for Retinopathy of Prematurity Cooperative Group. *Arch Ophthalmol*. 1988; 106(4): 471–479. [PubMed: 2895630]
3. Good WV, Early TFROPCG. Final results of the Early Treatment for Retinopathy of Prematurity (ETROP) randomized trial. *Trans Am Ophthalmol Soc*. 2004; 102: 233–48; discussion 248. [PubMed: 15747762]
4. International CFTCOROP. The International Classification of Retinopathy of Prematurity revisited. *Arch Ophthalmol*. 2005; 123(7): 991–999. [PubMed: 16009843]
5. Mintz-Hittner HA, Kennedy KA, Chuang AZ, BEAT-ROP CG. Efficacy of intravitreal bevacizumab for stage 3+ retinopathy of prematurity. *N Engl J Med*. 2011; 364(7): 603–615. [PubMed: 21323540]
6. Campbell JP, Kalpathy-Cramer J, Erdogmus D et al. Plus Disease in Retinopathy of Prematurity: A Continuous Spectrum of Vascular Abnormality as a Basis of Diagnostic Variability. *Ophthalmology*. 2016; 123(11): 2338–2344. [PubMed: 27591053]
7. Fleck BW, Williams C, Juszczak E et al. An international comparison of retinopathy of prematurity grading performance within the Benefits of Oxygen Saturation Targeting II trials. *Eye (Lond)*. 2018; 32(1): 74–80. [PubMed: 28752837]

8. Kalpathy-Cramer J, Campbell JP, Erdogmus D et al. Plus Disease in Retinopathy of Prematurity: Improving Diagnosis by Ranking Disease Severity and Using Quantitative Image Analysis. *Ophthalmology*. 2016; 123(11): 2345–2351. [PubMed: 27566853]
9. Reynolds JD, Dobson V, Quinn GE et al. Evidence-based screening criteria for retinopathy of prematurity: natural history data from the CRYO-ROP and LIGHT-ROP studies. *Arch Ophthalmol*. 2002; 120(11): 1470–1476. [PubMed: 12427059]
10. Chiang MF, Jiang L, Gelman R, Du YE, Flynn JT. Interexpert agreement of plus disease diagnosis in retinopathy of prematurity. *Arch Ophthalmol*. 2007; 125(7): 875–880. [PubMed: 17620564]
11. Hewing NJ, Kaufman DR, Chan RV, Chiang MF. Plus disease in retinopathy of prematurity: qualitative analysis of diagnostic process by experts. *JAMA Ophthalmol*. 2013; 131(8): 1026–1032. [PubMed: 23702696]
12. Slidsborg C, Forman JL, Fielder AR et al. Experts do not agree when to treat retinopathy of prematurity based on plus disease. *Br J Ophthalmol*. 2012; 96(4): 549–553. [PubMed: 22174097]
13. Wallace DK, Quinn GE, Freedman SF, Chiang MF. Agreement among pediatric ophthalmologists in diagnosing plus and pre-plus disease in retinopathy of prematurity. *J AAPOS*. 2008; 12(4): 352–356. [PubMed: 18329925]
14. Brown JM, Campbell JP, Beers A et al. Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks. *JAMA Ophthalmol*. 2018; 136(7): 803–810. [PubMed: 29801159]
15. Redd TK, Campbell JP, Brown JM et al. Evaluation of a deep learning image assessment system for detecting severe retinopathy of prematurity. *Br J Ophthalmol*. 2018; Epub ahead of print.
16. Taylor S, Campbell JP, Gupta K et al. A quantitative severity scale for retinopathy of prematurity using deep learning: monitoring disease progression. *JAMA Ophthalmol*. 2019; Epub ahead of print.
17. Ryan MC, Ostmo S, Jonas K et al. Development and Evaluation of Reference Standards for Image-based Telemedicine Diagnosis and Clinical Research Studies in Ophthalmology. *AMIA Annu Symp Proc*. 2014; 2014: 1902–1910. [PubMed: 25954463]
18. Gupta MP, Chan RVP, Anzures R et al. Practice Patterns in Retinopathy of Prematurity Treatment for Disease Milder Than Recommended by Guidelines. *Am J Ophthalmol*. 2016; 163: 1–10. [PubMed: 26705094]
19. Ruamviboonsuk P, Teerasuwanajak K, Tiensuwan M, Yuttitham K, Thai SFDRSG. Interobserver agreement in the interpretation of single-field digital fundus images for diabetic retinopathy screening. *Ophthalmology*. 2006; 113(5): 826–832. [PubMed: 16650679]
20. Campbell JP, Ryan MC, Lore E et al. Diagnostic Discrepancies in Retinopathy of Prematurity Classification. *Ophthalmology*. 2016; 123(8): 1795–1801. [PubMed: 27238376]
21. Campbell JP, Ataer-Cansizoglu E, Bolon-Canedo V et al. Expert Diagnosis of Plus Disease in Retinopathy of Prematurity From Computer-Based Image Analysis. *JAMA Ophthalmol*. 2016; 134(6): 651–657. [PubMed: 27077667]
22. Early TFROPCG. Revised indications for the treatment of retinopathy of prematurity: results of the early treatment for retinopathy of prematurity randomized trial. *Arch Ophthalmol*. 2003; 121(12): 1684–1694. [PubMed: 14662586]
23. Moleta C, Campbell JP, Kalpathy-Cramer J et al. Plus Disease in Retinopathy of Prematurity: Diagnostic Trends in 2016 Versus 2007. *Am J Ophthalmol*. 2017; 176: 70–76. [PubMed: 28087400]
24. Gupta K, Campbell JP, Taylor S et al. A quantitative severity scale for retinopathy of prematurity using deep learning: monitoring disease regression after treatment. *JAMA Oph*. 2019; Epub ahead of print.

We evaluated objective levels of vascular severity in images of babies diagnosed with plus disease and treated for retinopathy of prematurity by 11 clinicians in North America and found differences both within and between examiners.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.

Boxplot representing the overall distribution of quantitative vascular severity scores (1–9 scale) of patients diagnosed with plus disease by 11 clinical examiners. Thick black line represents the median. Black dots represent outliers.

Figure 2.

A, B, Representative examples of wide-angle retinal images and corresponding quantitative vascular severity scores (1–9 scale) from deep learning. Both images correspond to infants diagnosed with plus disease and treated at the time of image acquisition. The image in (A) seems to correspond to less severe disease, which is reflected in the vascular severity scores (upper right), suggesting that the vascular severity score may be a useful quantitative metric for standardizing the diagnosis of plus disease and treatment-requiring retinopathy of prematurity.

Table 1.

Mean ($\pm$ standard deviation) birth weight, gestational age, and postmenstrual age grouped by each of 11 clinical examiners.

Examiner	Number of Eyess	Birthweight (g)	Gestational Age (weeks)	Postmenstrual Age (weeks)
A	2	802 [± 0]	24.0 [± 0]	35.0 [± 0]
B	7	615 [± 101]	23.7 [± 0.5]	35.4 [± 1.1]
C	4	681 [± 82.5]	24.5 [± 0.6]	37.5 [± 4.4]
D	17	610 [± 95.3]	24.2 [± 0.9]	34.9 [± 1.3]
E	31	620 [± 130]	24.4 [± 1.3]	36.5 [± 3.5]
F	59	741 [± 209]	25.2 [± 1.2]	37.6 [± 3.6]
G	27	568 [± 116]	24.3 [± 1.2]	35.7 [± 1.6]
H	12	624 [± 102]	24.1 [± 0.3]	35.3 [± 2.2]
I	5	638 [± 114]	24.8 [± 0.5]	34.6 [± 0.9]
J	2	710 [± 0]	24.0 [± 0]	32.0 [± 0]
K	2	694 [± 0]	26.0 [± 0]	38.0 [± 0]
Total	168	659 [± 166]	24.5 [± 1.2]	36.4 [± 3.1]