Evaluation of a novel retinopathy of prematurity severity scale applied by clinicians and deep learning

J. Peter Campbell, MD, MPH,[1*] Sang Jin Kim, MD, PhD,[1,2*] James M. Brown, PhD,[3] Susan Ostmo, MS,[1] R. V. Paul Chan, MD,[4] Jayashree Kalpathy-Cramer, PhD,[5,6†] Michael F. Chiang, MD[1,7†] on behalf of the Imaging and Informatics in Retinopathy of Prematurity (i-ROP) Consortium.

[1]Department of Ophthalmology, Casey Eye Institute, Oregon Health & Science University, Portland, OR.
[2]Department of Ophthalmology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea.
[3]School of Computer Science, University of Lincoln, Lincoln, UK.
[4]Department of Ophthalmology and Visual Sciences, Illinois Eye and Ear Infirmary, University of Illinois at Chicago, Chicago, IL.
[5]Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Charlestown, MA.
[6]Massachusetts General Hospital and Brigham and Women's Hospital Center for Clinical Data Science, Boston, MA.
[7]Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, OR.

*Drs Campbell and Kim contributed equally to this work.
†Drs. Chiang and Kalpathy-Cramer supervised this work equally.

Supported by grants R01EY19474, K12EY27720, and P30EY10572 from the National Institutes of Health (Bethesda), by grant SCH-1622679 from the National Science Foundation (Arlington, VA), and by unrestricted departmental funding and a a Career Development Award (JPC) from Research to Prevent Blindness (New York, NY).

Disclosures: Sang Jin Kim is a Consultant for Novartis (Basel, Switzerland), Curacle (Seongnam, Korea), Hanmi Pharmaceutical (Seoul, Korea), and Reyon Pharmaceutical Co., Ltd. (Seoul, Korea). R.V. Paul Chan is on the Scientific Advisory Board for Phoenix Technology Group (Pleasanton, CA), a Consultant for Novartis (Basel, Switzerland), and a Consultant for Alcon (Ft. Worth, TX). Michael F. Chiang is a Consultant for Novartis (Basel, Switzerland), and an equity owner of Inteleretina (Honolulu, HI). Michael F. Chiang, J. Peter Campbell, R.V. Paul Chan, and Jayashree Kalpathy-Cramer receive research support from Genentech. R.V. Paul Chan receives research support from Regeneron.

None of the funding agencies had any role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Word Count: 2998
Abstract Word Count: 345

Correspondence to:
Michael F. Chiang, MD
Casey Eye Institute
Oregon Health & Science University
515 SW Campus Drive
Portland, OR 97239
Tel: 503-494-3667 | Fax: 503-494-5748 | Email: chiangm@ohsu.edu

ABSTRACT

OBJECTIVE: To evaluate the clinical utility of a quantitative deep-learning derived vascular severity score for retinopathy of prematurity (ROP) by assessing its correlation with clinical ROP diagnosis and by measuring clinician agreement in applying a novel scale.

DESIGN: Analysis of existing database of posterior pole fundus images and corresponding ophthalmoscopic examinations using two methods of assigning a quantitative scale to vascular severity.

SUBJECTS AND PARTICIPANTS: Images were from clinical exams of patients in the Imaging & Informatics in ROP consortium. 4 ophthalmologists and 1 study coordinator evaluated vascular severity on a 1-9 scale.

METHODS: A quantitative vascular severity score (1-9) was applied to each image using a deep learning algorithm. A database of 499 images was developed for assessment of inter-observer agreement.

MAIN OUTCOME MEASURES: Distribution of deep learning derived vascular severity scores with the clinical assessment of zone (I,II,III), stage (0,1,2,3) and extent (<3, 3-6, >6 clock hours) of stage 3 evaluated using multivariable linear regression. Weighted kappa and Pearson correlation coefficients for inter-observer agreement on 1-9 vascular severity scale.

RESULTS: For deep learning analysis, a total of 6344 clinical examinations were analyzed. A higher deep learning derived vascular severity score was associated with more posterior disease, higher disease stage, and higher extent of stage 3 disease (P<.001 for all). For a given ROP stage, the vascular severity score was higher in zone I than zone II or III (P<.001). For a given number of clock hours of stage 3, the severity score was higher in zone I than zone II (P=.03 in zone I and P<.001 in zone II). Multivariable regression found zone, stage, and extent were all independently associated with the severity score (P<.001 for all). For inter-observer agreement, mean (±Standard Deviation [SD]) weighted kappa was 0.67 (±0.06) and Pearson Correlation coefficient (±SD) was 0.88 (±.04) on the use of a 1-9 vascular severity scale.

CONCLUSIONS: A vascular severity scale for ROP appears feasible for clinical adoption, corresponds with current international classification of ROP severity, and facilitates the use of objective technology such as deep learning to improve consistency of ROP diagnosis.

INTRODUCTION

Plus disease has been a marker of severe retinopathy of prematurity (ROP) since prior to the development of the International Classification of ROP (ICROP) in the 1980s and has been an essential component of treatment decisions since the Multicenter Trial for Cryotherapy for ROP (CRYO-ROP) study.[1-3] CRYO-ROP demonstrated improved outcomes with treatment of threshold disease, defined as 5 continuous or 8 discontinuous clock hours of stage 3 ROP with plus disease, which was defined based on a standard photograph. Subsequently, the Early Treatment for ROP (ET-ROP) study supported revised treatment criteria for any eye with stage 3 in zone 1, or any extent and stage with plus disease.[4] This had the effect of removing a quantitative variable (extent of stage 3 disease) from the assessment of disease severity in ROP and replacing treatment decisions primarily with qualitative assessment of the anterior-posterior location of stage 3 disease, and the presence or absence of plus disease.

In many domains of medicine, technological advancements have led to a transition from qualitative and subjective assessment of disease severity to quantitative and objective measures of disease. In ophthalmology, for example, the development of optical coherence tomography (OCT) has led to clinical trial and treatment paradigms that increasingly rely on objective, quantitative measures rather than qualitative examination features. In terms of ROP, it is well established that there is significant inter-observer variability in all components of clinical diagnosis (zone, stage, plus disease), and growing evidence that this leads to real-world treatment variability.[5-10] For plus disease, it has been established that systematic bias between experts is a key source of diagnostic discrepancy along the continuum of disease severity.[11,12] To this end, an objective metric of ROP disease severity might improve diagnostic agreement and facilitate future clinical trials designed to improve visual and anatomic outcomes in ROP.

Deep learning in medicine has gained prominence as an artificial intelligence methodology with potential for extremely accurate image-based disease classification. We have previously demonstrated that a deep learning approach can diagnose plus disease as well as ROP experts, and subsequent work has demonstrated that this technology may be used to develop a continuous vascular severity score to quantify disease severity objectively.[13-16] However, there is a gap in knowledge regarding how a vascular severity score may integrate into the current ROP classification schema with zone, stage, and plus disease. Moreover, it is unclear whether increasing the granularity of "plus disease" along a continuum might worsen, rather than improve, diagnostic agreement.

In this study, we aimed to evaluate the relationship between a deep learning-derived vascular severity scale with zone, stage, extent of stage 3, and plus disease, and determine whether human graders may be able to adapt and utilize such as system. We feel this approach will have significant benefits for ROP care, and that it may be generalized to other ophthalmic diseases using deep learning methods.

METHODS

This study was conducted as part of a multicenter ROP cohort study by the Imaging and Informatics in ROP (i-ROP) consortium. This study was approved by the Institutional Review Board at the coordinating center (Oregon Health & Science University) and at each of 8 study centers (Columbia University, University of Illinois at Chicago, William Beaumont Hospital, Children's Hospital Los Angeles, Cedars-Sinai Medical Center, University of Miami, Weill Cornell Medical Center, Asociacion para Evitar la Ceguera en Mexico [APEC]). This study was conducted in accordance with the Declaration of Helsinki. Written informed consent for the study was obtained from parents of all infants enrolled in this study.

Datasets

Deidentified images from clinical examinations performed between July 2011 and December 2016 were assessed. All images were obtained using a commercially available camera (RetCam; Natus Medical Incorporated, Pleasanton, CA). Each study eye examination was assigned a reference standard diagnosis (RSD) for all combinations of zone, stage, and plus disease. The RSD was determined using methods previously published. [17]In brief, the reference standard was based on a consensus diagnosis between the ophthalmoscopic grading and 3 independent image-based diagnoses on the full ICROP classification including zone, stage, and plus. The dataset (ICROP comparison dataset) also included the extent of stage 3 disease (number of clock hours) as determined by ophthalmoscopy when stage 3 was diagnosed. Images of stage 4 and higher were excluded. A subset of this dataset (499 images) was set aside for reliability analysis (inter-observer agreement dataset).

Description of the clinician-assigned vascular severity score

We defined a scale from 1-9 to represent a spectrum of vascular abnormality. The labels 1-3 were applied when the image fell into the no plus category (with 1 reflecting very thin and straight vessels and 3 reflecting some vascular abnormality but insufficient for pre-plus disease). Similarly, 4-6 broadly reflected the range of pre-plus, and 7-9 reflected the range of disease where the majority of examiners would diagnose plus disease.

Reliability Analysis

Five trained graders (4 ophthalmologists experienced in ROP and 1 non-physician experienced in review of ROP images) independently graded the 499 images as 1 to 9 using this conceptual framework. To evaluate inter-observer agreement, we calculated weighted kappa and Pearson correlation coefficients for each pair of graders. Kappa values were interpreted using a commonly-accepted scale: 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, near-perfect agreement.

Comparison of deep learning-derived score with ICROP classification

The i-ROP deep learning system was used to classify the probability of an image having an associated reference standard diagnosis of plus disease on a 3-level scale (normal, preplus, plus) for each image in the ICROP comparison dataset. An automated ROP vascular severity score was then assigned to each image, from 1 (very thin and smooth vessels) to 9 (severe plus disease) using methods previously published based on the probabilities of each disease category: (1 × probability of normal) + (5 × probability of pre-plus disease) + (9 × probability of plus disease).[14,15,18]

We compared the quantitative vascular severity score (1-9) as a function of all ICROP components as determined by the reference standard diagnosis of plus (plus, pre-plus, or no plus), stage (0, 1, 2, 3) and as a function of number of quadrants with stage 3 disease (< 3 clock hours, between 3-6 clock hours, or > 6 clock hours), in zone I, II and III. Comparisons were done using analysis of variance (ANOVA) in Stata v15 (College Station, TX). We then performed multivariable linear regression comparing the 1-9 output as a function of zone, stage, and extent as above.

RESULTS

Evaluation of a deep learning derived vascular severity score.

Using the full ICROP comparison dataset, we were able to evaluate relationships between the deep learning-derived vascular severity score and the ICROP classification for 6344 eye examinations. **Table 1** displays the demographics of the dataset and the ICROP sub-classifications for all exams in the ICROP comparison dataset.

Figure 2 demonstrates the median (interquartile range [IQR]) vascular severity score for all images by RSD for plus disease on the left panel. Images had a median value of 1.2 (1.0-2.3) for no plus, 5.1 (4.6-6.0) for pre-plus, and 8.8 (8.2 – 9.0) for plus disease (P<0.01). In the middle panel, **Figure 2** demonstrates the median and IQR for the vascular severity score as a function of stage (0, 1, 2, 3) in each zone (I, II, III). The vascular severity score as associated with increasing stage of disease in zone I (left, P<.001), zone II (middle, P<.001), and zone III (right, P<.001), and the vascular severity score for stage 1, 2 and 3 was higher in Zone I than the corresponding score for the same stage of disease in zone II (P<.001). On the right, **Figure 2** demonstrates the same relationship with the extent of stage 3 disease. The vascular severity score was associated with a higher number of clock hours of stage 3 disease in both zone I and II (P=0.03 in zone I and P<.001 in zone II), and was higher in zone I than zone II for the same number of clock hours (P<.001). Multivariable regression found zone, stage, and extent were all independently associated with the 1-9 score (P<0.001 for all dependent variables).

<u>Reliability Analysis</u>

The distribution of disease severity for the inter-observer agreement dataset is shown in **Table 1**. The mean (± standard deviation [SD]) 1-9 score applied to images with an RSD of no plus disease was 2.4 (± 0.8) for no plus disease, 4.7 (±1.1) for pre-plus, and 7.7 (±1.0) for plus disease (P<.001). **Table 2** displays the relationship between the median 1-9 score assigned to each of the 499 images by the 5 graders versus the plus disease reference standard, demonstrating the transition from no plus to pre-plus between 3 and 4, and from pre-plus to plus between 6 and 7.

Table 3 reports the weighted kappa as well as the Pearson correlation coefficient for each examiner relative to each other. Kappa statistics showed that 9 of 10 paired comparisons showed strong agreement (kappa between 0.6 and 0.8) with a mean (±SD]) weighted kappa was 0.67 (±0.06). Mean Pearson correlation coefficient (±SD) was 0.88 (±.04) with all pairs of graders demonstrating high correlation (r > 0.8).

DISCUSSION

Retinal vascular changes in retinopathy of prematurity run a continuum from very mild to very severe. In the original ICROP, these changes were grouped into two categories: plus or no plus.[19] In the ICROP revisited paper in 2005, an intermediate pre-plus category was added.[1] In this paper, we propose expanding the ordinal categories to a more granular scale from 1-9, present two different methods for developing and validating such a scale, and demonstrate the relationship between the 1-9 scale and the conventional zone, stage, and plus disease classifications in ICROP. The key findings are: 1) A higher deep learning-derived vascular severity score was associated with indicators of more severe disease in the current ICROP classification such as more posterior zone, higher maximum stage, and higher extent of stage 3 disease. 2) Expert graders agreed on both absolute and relative 1-9 scores with moderate to high agreement.

These results highlight that although ICROP defined independent classifications for zone, stage, and plus disease, these categories are not physiologically independent. Instead, the underlying disease phenotypes reflect a spectrum of disease, which is reflected in changes in the vascular severity in the posterior pole. The zone of disease represents the area of vascularized retina, which correlates with the number of capillary beds between the central retinal artery and vein, and inversely with the area of avascular retina. The stage of disease represents the degree of disrupted vasculogenesis and extraretinal neovascularization at the border, which varies both in degree and extent for up to 12 clock hours, and which presumably leads to vascular shunting that increases total retinal blood flow. It is interesting to speculate how total retinal blood flow, the role of shunt vessels and intravascular resistance in large and small blood vessels might be related these changes in the posterior pole retinal vessels; however, these parameters are difficult to measure *in vivo*. The development of better tools to quantify retinal blood flow and the micro- and macro-vascular changes of retinal blood vessels in ROP, such as OCT angiography, [20] may help better elucidate these underlying mechanisms, and improve our understanding of ROP pathophysiology.

Further, results from this study demonstrate that clinicians may be able to recognize these subtle changes in vascular abnormality that correlate with changes in overall ROP severity. In some cases, these changes in posterior pole dilation and tortuosity can be appreciated, but are not captured in the current plus disease classification (**Figure 3)**. One advantage of a quantitative 1-9 scale applied clinically is that it may improve recognition of disease progression, even in the absence of photography and image analysis. Previous work has demonstrated that this deep learning-derived scale could be used to monitor disease progression, and disease regression after treatment, over time and provide benefits with regard to prediction of disease worsening or improvement.[14,15] In other words, whether applied subjectively by a clinician, or objectively by a deep learning system, documentation of vascular severity on a more granular level may facilitate earlier recognition and referral of worsening disease.

Another advantage of a quantitative 1-9 scale is that it separates the assessment of relative vascular severity from the treatment implications of a diagnosis of plus disease. That is, assessment of “plus disease” carries the connotation of “this baby needs to be treated” given current evidence-based treatment guidelines. In contrast, the diagnosis of a “7” simply implies that the vascular severity is more severe than a “6.” Previous work has demonstrated that clinicians are much more likely to agree on relative disease severity than on labels of plus disease, perhaps in part for this reason.[12,21] Although there are published evidence-based treatment criteria based on standard photographs for plus disease, it is well recognized that subjective cognitive processes affect perception of disease severity. In particular: 1) Despite the presence of a standard photograph, in research studies experts identify widely varying degrees of vascular abnormality as plus disease, with one study demonstrating some experts diagnose up to 6 times as many babies with plus disease compared to others.[11] 2) In clinical trials, differences in diagnosis of treatment-requiring ROP have been found to be due to plus disease diagnostic differences among physicians in different geographic regions, suggesting a training bias.[10,22] 3) When asked to explain clinical reasoning, experts often cite different phenotypic features when arriving at disparate diagnoses.[23] 4) In analysis of inter-observer discrepancies, pairs of experts were more likely to disagree on the diagnosis of plus if they also differ on the diagnosis of stage, suggesting that perception of vascular severity is influenced by assessment of peripheral pathology.[5] 5) Experts are more likely to diagnose plus disease if the pre-test probability for severe disease is higher based on demographics; that is, they are more likely to see plus disease if they believe that ought to be more likely to see plus disease.[24] All of these issues could be addressed with objective assessment of vascular severity.

The therapeutic implications of this proposed vascular severity score must be evaluated prospectively and carefully. Either through clinical adoption of standard images reflecting a wider range of vascular severity or through the use of deep learning, or both, prospective evaluation of clinical trial data may help elucidate the “right” level of vascular severity to label plus disease and continue to use evidence-based criteria to guide treatment. Alternatively, it may reveal that other combinations of zone, stage, and extent are as or more important than the absolute level of vascular severity in the posterior pole. These results suggest that, on average, a zone II eye, especially in anterior zone II, would need either a higher stage or more clock hours of pathology to have the same level of "plus-ness" as a zone I eye. This may explain why multiple studies have found approximately 10% of the time clinicians document that they are treating outside published guidelines based on clinical judgment, most commonly zone II stage 3 without plus.[25,26] Clinician should be aware of this finding to minimize adverse anatomic outcomes that can occur, such as vascular straightening even in the absence of retinal detachment. Since the subjective interpretation of plus disease was a hidden bias within the ETROP study, and it has become clear that this is interpreted so widely in the real world, without prospective adoption of a more granular clinical scale, or objective assessment of vascular severity, it is not clear how to ensure consistent interpretation of evidence-based medicine over time.

There are several limitations to this analysis. First, although we have proposed two methods for the development of a vascular severity score, one objective (based on deep learning), and one subjective (based on comparison to standard images), these methods were not designed to produce identical results especially at the low and high ends of the scale. The primary reason for this is that the current deep learning system was derived from a 3-level plus disease scale and thus has the same limitation as the current system (i.e. it was not calibrated to determine differences within a given plus disease level). Development of a larger database of clinician-labeled 1-9 images would enable training of a pure deep learning model either as a classification (to identify the most likely 1-9 class label) or a regression (continuous) model. Second, the deep learning model here was trained with plus disease reference standard labels from some of the same images as presented in the ICROP comparison dataset. This means that the highly significant association with plus disease is not surprising. However, it does not affect the interpretation of the relationship between zone, stage, and extent which were not part of the training. Third, the deep learning system was trained only on RetCam images and would need to be retrained and validated on other camera systems, and across a variety of image quality. [27] Fourth, all of the images in the training set were from a North American population and thus the translatability of this scale to other populations needs to be evaluated. Fifth, the ROP graders in this study are all collaborators and may demonstrate higher inter-rater agreement than a random sample of clinicians, though it suggests that, with training, agreement on a 1-9 scale is possible.

Taken together, these findings demonstrate how a more granular vascular severity scale for ROP, such as the one proposed, may complement the existing body of knowledge that multiple clinical trials have generated using the current ICROP classification. Adopting such a scale may facilitate more precise monitoring of disease progression and enable future clinical trials that rely on objective metrics of ROP disease severity. These results further demonstrate how the rise of deep learning systems may have clinical benefits beyond image-based diagnosis for ROP. Specifically, as more of medicine is moving towards objective and quantitative diagnosis, the use of deep learning to generate objective disease severity scales may be a generalizable methodology that works in many of the diseases where deep learning is currently being applied.

References

1. International Committee for the Classification of Retinopathy of Prematurity. The International Classification of Retinopathy of Prematurity revisited. In: Vol 123. American Medical Association; 2005:991–999.

2. Cryotherapy for Retinopathy of Prematurity Cooperative Group. Multicenter trial of cryotherapy for retinopathy of prematurity. Preliminary results. Arch Ophthalmol 1988;106:471–479.

3. Owens WC, Owens EU. Retrolental Fibroplasia. Am J Public Health Nations Health 1950;40:405–408.

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

5. Campbell JP, Ryan MC, Lore E, et al. Diagnostic Discrepancies in Retinopathy of Prematurity Classification. Ophthalmology 2016;123:1795–1801.

6. 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;96:549–553.

7. Quinn GE, Ells A, Capone A, et al. Analysis of Discrepancy Between Diagnostic Clinical Examination Findings and Corresponding Evaluation of Digital Images in the Telemedicine Approaches to Evaluating Acute-Phase Retinopathy of Prematurity Study. JAMA Ophthalmol 2016;134:1263–1270.

8. Chiang MF, Thyparampil PJ, Rabinowitz D. Interexpert Agreement in the Identification of Macular Location in Infants at Risk for Retinopathy of Prematurity. Arch Ophthalmol 2010;128:1153–1159.

9. Chiang MF, Jiang L, Gelman R, et al. Interexpert agreement of plus disease diagnosis in retinopathy of prematurity. Arch Ophthalmol 2007;125:875–880.

10. 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) 2017;123:1–7.

11. 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;123:2338–2344.

12. 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;0:2345–2351.

13. 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.

14. Taylor S, Brown JM, Gupta K, et al. Monitoring Disease Progression With a Quantitative Severity Scale for Retinopathy of Prematurity Using Deep Learning. JAMA Ophthalmol 2019;137:1022–1028.

15. Gupta K, Campbell JP, Taylor S, et al. A Quantitative Severity Scale for Retinopathy of Prematurity Using Deep Learning to Monitor Disease Regression After Treatment. JAMA Ophthalmol 2019;137:1029–1036.

16. Bellsmith KN, Brown J, Kim SJ, et al. Aggressive Posterior Retinopathy of Prematurity: Clinical and Quantitative Imaging Features in a Large North American Cohort. Ophthal 2020, epublished 2/7/2020.

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.

18. 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:bjophthalmol–2018–313156.

19. The Committee for the Classification of Retinopathy of Prematurity. An international classification of retinopathy of prematurity. Arch Ophthalmol 1984;102:1130–1134.

20. Campbell JP, Nudleman E, Yang J, et al. Handheld Optical Coherence Tomography Angiography and Ultra-Wide-Field Optical Coherence Tomography in Retinopathy of Prematurity. JAMA Ophthalmol 2017;135:977–981.

21. 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. Ophthalmol 2016;123:2338–2344.

22. 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;120:1470–1476.

23. Hewing NJ, Kaufman DR, Chan RVP, Chiang MF. Plus Disease in Retinopathy of Prematurity: Qualitative Analysis of Diagnostic Process by Experts. JAMA Ophthalmol 2013;131:1026–1032.

24. Gschließer A, Stifter E, Neumayer T, et al. Effect of Patients' Clinical Information on the Diagnosis of and Decision to Treat Retinopathy of Prematurity. Retina (Philadelphia, Pa) 2017:1.

25. Gupta MP, Anzures R, Ostmo S, et al. Practice Patterns in Retinopathy of Prematurity Treatment for Disease Milder than Recommended by Guidelines. Am J Ophthalmol 2015;163:1–10.

26. Liu T, Ying G, Yang MB, Binenbaum G. Treatment of pre–type 1 disease in the postnatal growth and retinopathy of prematurity (G-ROP) Study. 2018.

27. Coyner AS, Swan R, Campbell JP, et al. Automated Fundus Image Quality Assessment in Retinopathy of Prematurity Using Deep Convolutional Neural Networks. Ophthalmology Retina 2019;3:444–450.

Figure Legends

Figure 1: Representative images from each 1-9 label. These images were selected based on the reference standard diagnosis with 1-3 having a diagnosis of no plus, 4-6 having a diagnosis of pre-plus, and 7-9 having a diagnosis of plus, but with varying degrees of vascular severity within each class.

Figure 2: Relationship between deep learning (DL) derived vascular severity score and zone, stage, extent and plus classifications. A higher vascular severity score (1-9) was associated with higher disease stage and extent of stage 3. For a given stage and extent of stage 3, the vascular severity score was higher in zone I compared with zone II or III.

Figure 3. Disease progression using current versus proposed classification. Two eyes that were included in the dataset and were noted to have disease progression over time. In both (A) and (B), disease progression is noted using the 1-9 scale that was not reflected in a change in plus disease reference standard diagnosis.

Acknowledgments

Members of the 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, Aaron Nagiel, MD, Evan Kruger, BA, Kathryn McGovern, MPH, Dilshad Contractor, Margaret Havunjian. 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.