AMERICAN ACADEMY OF OPHTHALMOLOGY®

Artificial Intelligence for Retinopathy of Prematurity

Validation of a Vascular Severity Scale against International Expert Diagnosis

J. Peter Campbell, MD, MPH,[1] Michael F. Chiang, MD, MA,[2] Jimmy S. Chen, MD,[1] Darius M. Moshfeghi, MD,[3] Eric Nudleman, MD, PhD,[4] Paisan Ruambivoonsuk, MD,[5] Hunter Cherwek, MD,[6] Carol Y. Cheung, PhD,[7] Praveer Singh, PhD,[8,9] Jayashree Kalpathy-Cramer, PhD,[8,9] Susan Ostmo, MS,[1] Malvina Eydelman, MD,[10] R.V. Paul Chan, MD,[11] Antonio Capone Jr., MD,[12] for the Collaborative Community in Ophthalmic Imaging Executive Committee and the Collaborative Community in Ophthalmic Imaging Retinopathy of Prematurity Workgroup

Purpose: To validate a vascular severity score as an appropriate output for artificial intelligence (AI) Software as a Medical Device (SaMD) for retinopathy of prematurity (ROP) through comparison with ordinal disease severity labels for stage and plus disease assigned by the International Classification of Retinopathy of Prematurity, Third Edition (ICROP3), committee.

Design: Validation study of an AI-based ROP vascular severity score.

Participants: A total of 34 ROP experts from the ICROP3 committee.

Methods: Two separate datasets of 30 fundus photographs each for stage (0−5) and plus disease (plus, preplus, neither) were labeled by members of the ICROP3 committee using an open-source platform. Averaging these results produced a continuous label for plus (1−9) and stage (1−3) for each image. Experts were also asked to compare each image to each other in terms of relative severity for plus disease. Each image was also labeled with a vascular severity score from the Imaging and Informatics in ROP deep learning system, which was compared with each grader's diagnostic labels for correlation, as well as the ophthalmoscopic diagnosis of stage.

Main Outcome Measures: Weighted kappa and Pearson correlation coefficients (CCs) were calculated between each pair of grader classification labels for stage and plus disease. The Elo algorithm was also used to convert pairwise comparisons for each expert into an ordered set of images from least to most severe.

Results: The mean weighted kappa and CC for all interobserver pairs for plus disease image comparison were 0.67 and 0.88, respectively. The vascular severity score was found to be highly correlated with both the average plus disease classification (CC = 0.90, $P < 0.001$) and the ophthalmoscopic diagnosis of stage ($P < 0.001$ by analysis of variance) among all experts.

Conclusions: The ROP vascular severity score correlates well with the International Classification of Retinopathy of Prematurity committee member's labels for plus disease and stage, which had significant intergrader variability. Generation of a consensus for a validated scoring system for ROP SaMD can facilitate global innovation and regulatory authorization of these technologies. *Ophthalmology 2022;129:e69-76 © 2022 Published by Elsevier Inc. on behalf of the American Academy of Ophthalmology*

The Collaborative Community in Ophthalmic Imaging (CCOI) was organized to clarify challenges and determine best practices, strategies, and standards for advancing innovation in ophthalmic imaging, including artificial intelligence (AI) Software as a Medical Device (SaMD) solutions.[1−3] Retinopathy of prematurity (ROP) is 1 of the 4 current vertical work groups focused on specific clinical indications: ROP was chosen by the organizing committee because of its epidemiological importance within the field of ophthalmology,[4,5] and extensive efforts are under way to use AI to screen for and diagnose the disease.[6−16]

Retinopathy of prematurity is a leading cause of childhood blindness worldwide, disproportionately affecting low- and middle-income countries.[4,5] Yet, blindness from ROP is nearly always preventable with eye examinations and access to treatment in the at-risk population. With increasing availability of digital fundus photography, there has been significant effort in the development of AI to provide an objective and quantitative framework for evaluation of patients with ROP to extend the reach of providers and improve interobserver agreement.[6−16] Although this may have benefits in any ROP screening population, it may have the greatest impact in lower- and middle-income

https://doi.org/10.1016/j.ophtha.2022.02.008
ISSN 0161-6420/22

countries where the disease prevalence is highest and available human resources are lowest.[15] Thus, a compelling equity case can be made for improved access to expert-level care through AI SaMD globally.[4,5,17]

The International Classification of ROP (ICROP) was first established in 1984 before the availability of modern imaging technologies and relies on subjective and qualitative assessment of disease features. The ICROP defined the subclassifications of zone (I–III), stage (0–5), and plus disease (present or not).[18] This classification schema provided the framework current treatment guidelines, which are based on the 2003 Early Treatment for ROP study that established that presence of plus disease, based on the original standard photograph(s)[18,19] with one rare exception (zone I stage 3 without plus), is the determining factor for which babies need to be treated.[20] The ICROP was revisited in 2005 and defined "preplus" disease as an intermediate category, reflecting that the spectrum of vascular dilation and tortuosity could not be adequately represented as a binary category.[19] In 2021, the ICROP, Third Edition (ICROP3), was published and extended that observation to acknowledge that preplus and plus are part of a disease continuum that ranges from immature retinal vasculature to severe ROP requiring treatment, and although there is agreement at the ends of the spectrum, significant interobserver agreement is a problem in the middle.[21,22]

While this new classification system is helpful in that it better reflects patterns of clinical diagnosis[22] and the underlying spectrum of disease, it does not necessarily help improve clinical diagnosis. Complicating matters, clinicians often disagree on all components of the ICROP classification, including zone, stage,[23–25] and plus disease.[22–26] More specifically, in some cases they disagree on the diagnosis itself, for example, what level of disease constitutes plus versus preplus disease or stage 2 versus stage 1.[22] In other cases, they disagree on the diagnostic process, that is, determining which clinical features (e.g., number of quadrants, dilation vs. tortuosity) are relevant to the diagnostic label (e.g., "there isn't enough vascular branching to call it stage 2").[27,28] Despite the presence of standard photographs both from the original and revised ICROP documents, there is evidence of geographic and temporal diagnostic drift (differences in plus disease diagnosis due to geographic or training differences, and over time),[29] a key consideration in the development and evaluation of reference standards for SaMD,[1] and evidence of treatment variability due to these differences.[14,30]

The 2020 CCOI workshop reviewed key questions related to clinical evaluation of ROP SaMDs within the context of the International Medical Device Regulators Forum framework.[1–3] The CCOI ROP working group concluded that additional work is needed to establish valid clinical association and reference standards for SaMD aimed at screening for or diagnosing ROP. Valid clinical associations, as defined by the International Medical Device Regulators Forum,[1] refer to the extent to which the SaMD's output is clinically accepted and corresponds accurately in practice to the SaMD's targeted clinical condition. The working group believes a vascular severity score that highly correlates with the clinical diagnosis of plus disease, and overall severity of disease may be one acceptable output. Consensus on such a score would allow future ROP SaMDs to use this score as the output of these devices informing health care providers in a consistent manner. As discussed by Abramoff et al[1] for diagnostic AI SaMD, a reference standard is needed for output comparison. Such reference standards can be categorical or continuous. An ROP vascular severity score with corresponding images that are accepted by the ROP community could become the reference standard for future diagnostic or screening AI ROP systems.

Recent work evaluating the Imaging and Informatics in ROP deep learning (i-ROP DL) AI system has demonstrated that the output of the system can be directly converted to a continuous vascular severity score, based on the algorithmic probability of preplus or plus disease diagnosis.[13,16] The original i-ROP DL algorithm produced a 3 class output (no plus, preplus, or plus), with underlying probabilities for each output, which are used to generate an output from 1 (100% probability of no plus) to 9 (100% probability of plus).[13,16] The committee decided to explore the validity of this vascular severity score as an output of an ROP SAMD against an expert group of clinicians, as there is a gap in knowledge as to whether the current algorithm, which was trained on the basis of a reference standard diagnosis from a small group of experts,[31,32] reflects the broader community of ROP clinicians' assessment of plus disease and overall ROP severity, as suggested by prior work.[10,12,13] In this article, we compare clinical diagnostic labels for stage 1 to 3 and plus disease assigned by the 34 members of the ICROP3 committee with the output of the i-ROP to investigate the clinical validity, or clinical meaningfulness, of the vascular severity score.

Methods

Description of Datasets

This study is based on data obtained during the third ICROP committee process, using images from prior ICROP publications and the Imaging and Informatics in ROP (i-ROP) cohort study.[21] The i-ROP study was approved by an institutional review board at Oregon Health and Science University and at each center with written informed consent obtained from parents of all infants whose images were included in the study. This study adhered to the Declaration of Helsinki. Thirty-four experts were asked to label 2 separate sets of images, 1 for plus and 1 for stage. The plus dataset consisted of 30 wide-angle retinal images of the posterior retina obtained by the Retcam Camera (Natus Medical Inc), as part of the i-ROP cohort study. Independently, experts were asked to label each image: plus, preplus, or neither. Full ophthalmoscopic classification of zone and stage was also available for 29/30 of the eyes with images in the plus disease dataset.

The stage dataset consisted of 33 images of various fields of view and stages of ROP, some of which came from the i-ROP study and were Retcam images and some of which were submitted by members of the ICROP committee or included in previous ICROP versions (including Kowa camera images) as typical examples of the representative stage. Thirty of 34 experts labeled the images as representing stage 0 (incomplete vascularization), 1, 2, 3, 4a, 4b, or 5, with the label representing the highest stage present in

the image. A total of 28 images with a mode diagnosis of stage 0 to 3 were used for this analysis.

Classification Labels: Ordinal and Continuous Labels for Plus and Stage

Following the ICROP expert labeling process, images were assigned ordinal class labels (i.e., plus, preplus, neither) based on the mode (most common) label of the group of experts for plus and stage. Images were then ranked from least severe to most severe based on expert classification in each dataset and assigned an "average" plus disease severity based on the arithmetic mean for 34 grades of no plus (1), preplus (5), and plus (9), resulting in a score of 1 to 9 for each image in the plus dataset. A similar process was performed to calculate an average stage classification for images with a mode diagnosis of stage 0 to 3, excluding examples of stage 4 and 5 retinal detachments. Figure 1 provides 2 examples of images from each dataset, with the expert labels for those images, and the average score.

Classification versus Comparison Labels: Agreement and Rankings of Plus Disease

For the plus disease dataset, in addition to providing a "class label" (plus, preplus, neither) for each image, experts were asked to perform "comparisons" for each pair of images in the dataset. To do this, we applied a similar methodology to Kalpathy-Cramer et al[33] using an open-source, web-based image severity assessment platform. Each expert was provided a unique access weblink and then presented with sequential pairs of images. For each pair, they were asked to "click on the image that represents more severe disease." To avoid confounding by the zone or stage of disease, no images showed the location or extent of stage. We used the Elo algorithm to convert the pairwise comparisons into an ordered set of images from least to most severe.[33]

We calculated weighted kappa and Pearson correlation coefficients (CCs) for each pair of graders and the mode class label. For evaluation of comparison ranking agreement, we compared the Pearson CC for each expert's ordered set of ranked images with the final ordered set as determined by the Elo algorithm. Finally, we compared the CC for the ordered set of images determined by averaging expert class labels with the pairwise comparison ranking. Agreement analysis was performed in Stata version 15 (StataCorp).

Relationship between Deep Learning–Derived Severity Score and Expert Classifications

Each image was labeled with a deep learning–derived score from the i-ROP DL algorithm. From the output of the algorithm, an automated ROP vascular severity score was assigned to each image from 1 (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

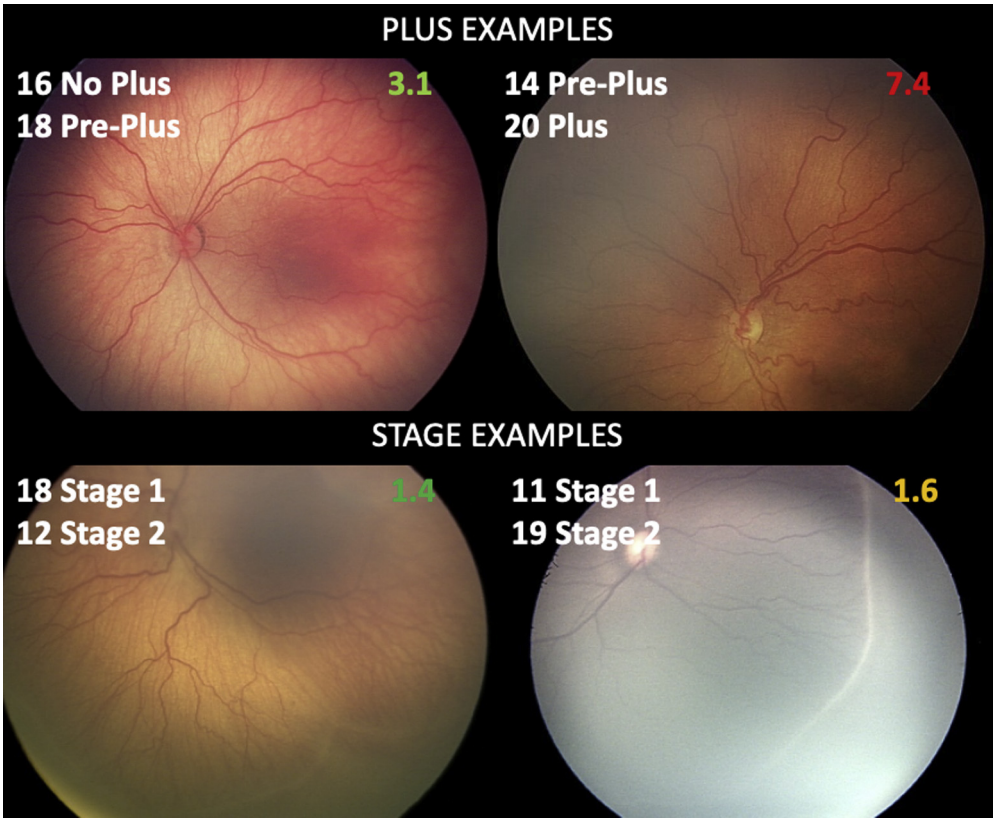

Figure 1. Example images for determination of plus disease and stage by members of the International Classification of Retinopathy of Prematurity (ICROP) committee. The 2 images at the top are from the plus disease database, and the bottom 2 images are from the stage dataset. The white numbers represent the number of committee experts who assigned each label for plus or stage to each image, respectively. The colored numbers in the upper right of each image represent the averaged expert classification on a scale of 1 to 3 for stage and 1 to 9 for plus. Both plus and stage seem to present on a continuum, which can be measured by comparing expert labels. Note that the lower right quadrant image is one of the standard images of "stage 1" disease published in the 2005 ICROP revisited paper,[19] which may be an example of temporal diagnostic drift.

preplus disease) + (9 × probability of plus disease). We compared the quantitative vascular severity score (1–9) as a function of mode ICROP committee plus disease diagnosis (no, pre, plus), as well as to the ophthalmoscopic diagnosis of stage in those eyes (which was not shown on the photograph but was available for 29/30 images). Comparisons were done using analysis of variance in Stata version 15 (StataCorp).

Results

Expert Classification of Plus and Stage

Figure 2 demonstrates the individual expert class labels, the ordinal (mode) class label, and the average label for the plus and stage image datasets. The plus dataset contained 8, 13, and 9 images with a mode diagnosis of no plus, preplus, and plus, respectively. The stage dataset contained 1, 6, 8, and 13 images with a mode diagnosis of stage 0, 1, 2, and 3, respectively. There was significant variability observed in plus disease labels by experts among the ICROP group, with 100% agreement only at extreme ends of the spectrum (least and most severe). The mean (and standard deviation) number of images labeled plus by experts was 9.9 (4.8) with a range of 3/30 (10%) to 22/30 (73%). For the stage image dataset, a similar spectrum of clinical variability seemed to be present for stage 1 and 2 ROP.

Classification versus Comparison Labels: Agreement and Rankings of Plus Disease

Figure 3 displays the results of the interobserver agreement for class labels (weighted kappa) and comparison labels (CC) for each pair of 34 images. The mean weighted kappa and CC for all interobserver pairs for plus disease image comparison were 0.67 and 0.88, respectively. Figure 3 also shows the relationship between ranking of plus disease images using the method of averaging class labels (as shown in Fig 2) versus Elo ranking (CC = 0.96).

Relationship between Deep Learning–Derived Severity Score and Expert Classifications

We compared the deep learning–derived vascular severity score across a number of ordinal groupings, as shown in Figure 4. The vascular severity score was significantly associated with mode plus label ($P < 0.001$) and with the ophthalmoscopic diagnosis of stage in the same eyes ($P < 0.001$) and correlated well with the average disease severity (CC = 0.90).

Discussion

In this article, we use 2 datasets of expert-annotated images and a deep learning–derived vascular severity score to demonstrate the following key findings. First, a deep learning–derived quantitative vascular severity score correlates well with the spectrum of expert labels plus disease in posterior pole images and the ophthalmoscopic diagnosis of peripheral stage. Second, although the ROP categories of stage and plus are defined ordinally, experts diagnose both along a continuum. Together, these findings suggest that a quantitative vascular severity score may be an appropriate output for an ROP SaMD.

To some degree, it is useful and necessary for a clinical classification schema to collapse a disease spectrum into ordinal categories, as has been done for multiple ophthalmic diseases, including diabetic retinopathy, cataract, and macular degeneration. These classification systems are necessary for clinical trials to ensure consistent application of evidence-based practice over time; however, when the classification is based primarily on subjective and qualitative features, then issues of reproducibility, repeatability, and diagnostic drift can limit the quality and consistency of care.[1] In the case of plus disease, there are multiple standard

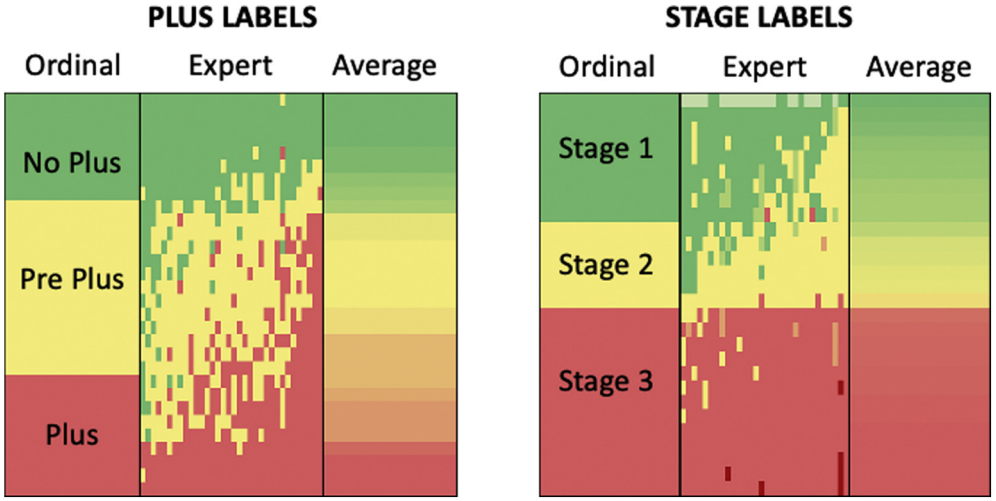

Figure 2. Spectrum of disease severity for plus and stage in retinopathy of prematurity (ROP). In each case, the middle portion of the figure represents the individual expert labels for each image in the dataset for plus (N = 30) and stage (N = 28). Each row represents one image, and the columns in the "Expert" section depict individual expert classifications. Experts were ranked in order of least aggressive diagnosis to most aggressive diagnosis from left to right. Images were ranked from least severe to most severe by average expert classification. Color code represents the underlying class label from green to red in order of increasing severity (no plus, preplus, plus, or stage 0, 1, 2, 3, or 4). The "Ordinal" column represents the mode classification, reflecting the current ICROP classification schema, and the "Average" column represents the average disease classification severity from the individual ICROP experts. Average disease severity better reflects expert diagnosis than an ordinal classification system.

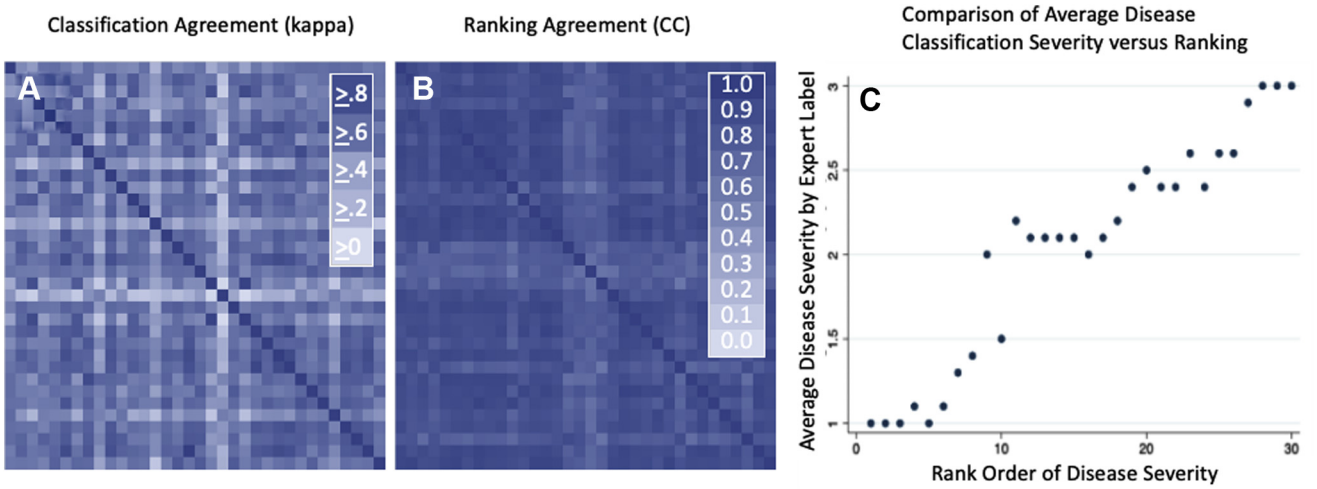

Figure 3. Classification versus comparison agreement. **A,** Interexpert agreement on plus disease label for 34 experts. Inset legend reports weighted kappa color scale for pairwise agreement for each expert relative to each other. Mean weighted kappa for all interobserver pairs is 0.67. **B,** Interexpert agreement for overall disease rankings for relative disease severity for 34 experts, as measured by correlation coefficient (CC). Mean CC for all interobserver pairs is 0.88. **C,** Correlation between average disease severity according to ordinal labels of 34 experts versus rank ordered severity using relative rankings (CC = 0.96).

photographs of plus disease from the original and the 2005 revisited ICROP documents[18,19]; nonetheless, there was widespread disagreement in the number of images representing plus disease among members of the ICROP3 committee (Fig 2). Both this number and the observations that, despite differences in labels (Figs 2 and 3A), experts tend to agree on relative disease severity (Fig 3B) and that these can both produce estimates of a spectrum of disease severity that are highly correlated (Fig 3C) are all consistent with prior research studies of expert consensus.[22] Similar disagreements are shown here for early stages of ROP and elsewhere for image-based classification of all components of the ICROP classification.[23,24] As such, it should not be surprising that we see variation in application of the same evidence in practice.[14,30] Based on these results and the available literature, the updated ICROP document acknowledges that preplus and plus disease run a spectrum and that clinical judgment should be applied when making treatment decisions.[34–36]

Clinical validation study design will be important to ensure the ethical implementation of AI, including evidence of effectiveness in each target population and for each intended use. Study designs should replicate clinical workflow and human interaction. Especially in the case of ROP, where clinical classification is subjective, selection, and validation of the reference standard diagnosis will be critical. Abramoff et al[1] proposed a framework for determining the level of reference standard (I–IV). The highest reference standard (level I) would be an independent reading center whose diagnoses are demonstrably related to improved clinical outcomes. However, this level of evidence does not currently exist for ROP, for the reasons stated above,

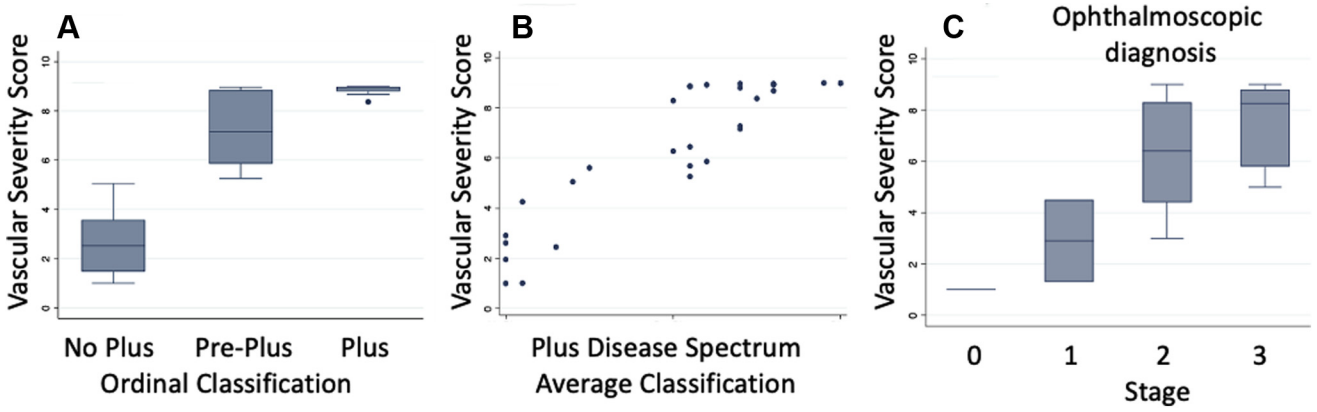

Figure 4. Relationship between deep learning–derived vascular severity score (VSS) and the mode plus classification, average plus classification, and associated ophthalmoscopic diagnosis of stage in plus disease dataset. **A,** Box plot of VSS versus mode plus disease classification ($P < 0.001$). **B,** Scatter plot of VSS versus average disease severity classification (correlation coefficient = 0.90). **C,** Box plot of VSS from plus disease images compared with ophthalmoscopic diagnosis of stage in the same eyes ($P < 0.001$). The VSS corresponds to the current mode classification of plus disease, a continuous spectrum of plus disease as determined by expert classifications, and with the ophthalmoscopic diagnosis of stage in the same eyes (not shown on images).

including evidence of geographic and temporal diagnostic drift for plus disease diagnosis.[29,30] Thus, at the present time, the best available reference standard would be level III, based on the adjudication of multiple graders, with published rates of repeatability and reproducibility. Finally, standards for minimum accepted criteria have yet to be established, such as evidence of improved clinical diagnostic performance with an assistive device, compared with the reference standard diagnosis. Such criteria may need to be higher for an autonomous device given the potential lifelong morbidity of the missed diagnosis of plus disease.

Study Limitations

There are a number of limitations to this analysis. First, the members of the ICROP committee were provided with demographic information or evidence of the zone or stage of disease and were not responsible for clinical care of the patients whose images were included. In that way, this experiment fails to reflect the cognitive processes involved in making the clinical diagnosis of plus disease. However, it does reflect the way plus disease was *supposed to be* diagnosed in most of the existing clinical trials, as an independent disease feature, unrelated to other demographic or clinical features.[20] Second, these datasets represent images taken of babies in North America on a single camera system and may not reflect the full diversity of disease severity or image appearance in other racial or ethnic populations, on other camera systems, or in other parts of the world. Validating the scale on more diverse populations is necessary to determine its utility globally, and early efforts evaluating the i-ROP DL system are promising.[15] Third, 3 of the ICROP members were part of the reference standard diagnosis process for the i-ROP system. As such, to ensure that this did not bias the association, we re-reran the analysis for Figure 3A and B excluding these 3 clinicians and found that both the association between VSS and mode diagnosis and relationship between average score and VSS remained the same ($P < 0.001$ and CC = 0.90, respectively). Fourth, the datasets were small and hand-selected for potentially publishable (higher quality) images. Fifth, the use of softmax probability output for development of a VSS is not the only, and may not be the optimal, method to develop a continuous scale for ROP.

In conclusion, this article demonstrates that both plus and stage in ROP present along a continuum, which can quantified using deep learning. This represents one potential application of deep learning in clinical medicine in general, and ophthalmology in particular, moving from qualitative to quantitative diagnosis. The use of ROP SaMD for treatment decisions based on plus disease will require validation in a clinical trial setting given the low reproducibility within the ICROP committee as to the level of severity constituting plus disease and the likely presence of temporal diagnostic drift since clinical trials established the benefit of identifying plus disease.[13,29,30] The potential global impact may be even higher for ROP than for some other diseases because the burden of blindness from ROP disproportionately affects infants in low- and middle-income countries,[4] and thus, an ROP SaMD that has demonstrated effectiveness in those populations, achieves regulatory authorization,[1] and is implemented in a sustainable, scalable manner could have a significant impact on the epidemiology of blindness from ROP in the future.[15,37]

Acknowledgments

CCOI Executive Committee. The authors thank Michael Abramoff, MD, PhD; Mark Blumenkranz, MD; Malvina Eydelman, MD; David Myung, MD, PhD; Joel S. Schuman, MD; Carol Shields, MD; Aaron Lee, MD; and Michael Repka, MD.

CCOI ROP Workgroup. The authors thank Michael F. Chiang, MD, MA; J. Peter Campbell, MD, MPH; Darius M. Moshfeghi, MD; Eric Nudleman, MD, PhD; Paisan Ruamviboonsuk, MD; D. Hunter Cherwek, MD; Carol Y. Cheung, PhD; R.V. Paul Chan, MD; and Antonio Capone, Jr, MD.

Members of the Third International Classification for ROP Committee. The authors thank Audina Berrocal, MD; Gil Binenbaum, MD, MSCE; Michael Blair, MD; J. Peter Campbell, MD, MPH; Antonio Capone, Jr, MD; R.V. Paul Chan, MD; Yi Chen, MD; Michael F. Chiang, MD, MA; Shuan Dai, MD; Anna Ells, MD; Alistair Fielder, MD; Brian Fleck, MD; William Good, MD; Mary Elizabeth Hartnett, MD; Gerd Holmstrom, MD, PhD; Shunji Kusaka, MD, PhD; Andres Kychenthal, MD; Domenico Lepore, MD; Birgit Lorenz MD, PhD; Maria Ana Martinez-Castellanos, MD; Sengul Ozdek, MD; Dupe Popoola, MBBS; Graham Quinn, MD, MSCHE; James Reynolds, MD; Parag Shah, DNB; Michael Shapiro, MD; Andreas Stahl, MD; Cynthia Toth, MD; Anand Vinekar, MD; Linda Visser, MD; David Wallace, MD, MPH; Wei-Chi Wu, MD, PhD; Peiquan Zhao, MD; and Andrea Zin, MD, MSc.

Footnotes and Disclosures

Originally received: November 8, 2021.
Final revision: January 31, 2022.
Accepted: February 3, 2022.
Available online: February 12, 2022. Manuscript no. D-21-02185.

[1] Casey Eye Institute, Department of Ophthalmology, Oregon Health & Science University, Portland, Oregon.

[2] National Eye Institute, National Institutes of Health, Bethesda, Maryland.

[3] Byers Eye Institute, Horngren Family Vitreoretinal Center, Department of Ophthalmology, Stanford University, Palo Alto, California.

[4] Department of Ophthalmology, University of California, San Diego, California.

[5] Department of Ophthalmology, Rajavithi Hospital, Bangkok, Thailand.

[6] Orbis International, New York, New York.

[7] Department of Ophthalmology and Visual Sciences, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China.

[8] Department of Radiology, MGH/Harvard Medical School, Charlestown, Massachusetts.

[9] Massachusetts General Hospital & Brigham and Women's Hospital Center for Clinical Data Science, Boston, Massachusetts.

[10] Center for Devices and Radiological Health, US Food and Drug Administration, Silver Spring, Maryland.

[11] Department of Ophthalmology and Visual Sciences, University of Illinois at Chicago, Chicago, Illinois.

[12] Associated Retinal Consultants, Oakland University William Beaumont School of Medicine, Royal Oak, Michigan.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have made the following disclosure(s): A.C.: Equity owner — Phoenix Technology Group, BroadSpot Imaging Corporation; Patent co-holder — "Automated system for measurement of zone 1 in assessment of severity of retinopathy of prematurity" (patent number: 9622657).

R.V.P.C.: Scientific Advisory Board — Phoenix Technology Group; Consultant —Alcon; Research support — Regeneron, Genentech; Equity owner — Siloam Vision.

M.F.C.: Consultant — Novartis (previously); Equity owner — Inteleretina; Research support — Genentech.

N.K.-C.: Research support —Genentech.

J.P.C.: Consultant — Boston AI labs, which has licensed the i-ROP DL algorithm from Massachusetts General Hospital, Oregon Health & Science University, University of Illinois Chicago, and Northeastern University; Equity owner — Siloam Vision; Research support —Genentech.

D.M.M.: Equity owner — Visunex Medical Systems Inc., LLC, Pr3vent, Inc.; Consultant for ROP — Akebia, Bayer, Novartis (previously); Steering Committee — Regeneron Butterfleye Trial; Grant support — OHSU via Genentech for ROP.

P.R.: Advisory Board — Roche; Consultant — Novartis, Research support — Roche, Novartis.

This work was supported by grants R01EY19474, R01 EY031331, and P30 EY10572 from the National Institutes of Health (Bethesda, MD) and by unrestricted departmental funding and a Career Development Award (to J.P.C.) from Research to Prevent Blindness (New York, NY).

The Food and Drug Administration participates as a member of the Collaborative Community on Ophthalmic Imaging. This article reflects the views of the authors and should not be construed to represent Food and Drug Administration's views or policies.

HUMAN SUBJECTS: Human subjects were included in this study. The i-ROP study was approved by an institutional review board at Oregon Health & Science University and at each center. Written informed consent was obtained from parents of all infants whose images were included in the study. All research adhered to the tenets of the Declaration of Helsinki.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Campbell, Chiang, Chen, Singh, Kalpathy-Cramer, Ostmo, Chan

Data collection: Campbell, Chiang, Chen, Moshfeghi, Nudleman, Ruambivoonsuk, Cherwek, Cheung, Singh, Kalpathy-Cramer, Ostmo, Eydelman, Chan, Capone

Analysis and interpretation: Campbell, Chiang, Chen, Moshfeghi, Nudleman, Cherwek, Cheung, Singh, Kalpathy-Cramer, Ostmo, Eydelman, Chan, Capone

Obtained funding: N/A; Study was performed as part of the authors' regular employment duties. No additional funding was provided.

Overall responsibility: Campbell, Chiang, Chen, Moshfeghi, Nudleman, Ruambivoonsuk, Cherwek, Cheung, Singh, Kalpathy-Cramer, Ostmo, Eydelman, Chan, Capone

Abbreviations and Acronyms:
AI = artificial intelligence; **CC** = correlation coefficient; **CCOI** = Collaborative Community in Ophthalmic Imaging; **ICROP** = International Classification of Retinopathy of Prematurity; **ICROP3** = International Classification of ROP, Third Edition; **i-ROP** = Imaging and Informatics in ROP; **i-ROP DL** = Imaging and Informatics in ROP deep learning; **ROP** = retinopathy of prematurity; **SaMD** = Software as a Medical Device.

Keywords:
Artificial intelligence, Deep learning, Disease classification, Interobserver agreement, Retinopathy of prematurity, Severity score.

Correspondence:
J. Peter Campbell, MD, MPH, Casey Eye Institute, Department of Ophthalmology, Oregon Health & Science University, 515 SW Campus Drive, Portland, OR 97239. E-mail: campbelp@ohsu.edu.

References

1. Abramoff MD, Cunningham B, Patel B, et al. Foundational considerations for artificial intelligence utilizing ophthalmic images. *Ophthalmology*. 2022;129:e14–e32.
2. U.S. Food & Drug Administration (FDA). Digital Health Center of Excellence, Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. https://www.fda.gov/media/145022/download. Accessed January 11, 2021.
3. Collaborative Community on Ophthalmic Imaging. https://www.cc-oi.org/. Accessed August 31, 2021.
4. Blencowe H, Lawn JE, Vazquez T, et al. Preterm-associated visual impairment and estimates of retinopathy of prematurity at regional and global levels for 2010. *Pediatr Res*. 2013;74:35–49.
5. Blencowe H, Cousens S, Chou D, et al. Born too soon: the global epidemiology of 15 million preterm births. *Reprod Health*. 2013;10:S2.
6. Zhang J, Liu Y, Mitsuhashi T, Matsuo T. Accuracy of deep learning algorithms for the diagnosis of retinopathy of prematurity by fundus images: a systematic review and meta-analysis. *J Ophthalmol*. 2021;2021:8883946.
7. Scruggs BA, Chan RVP, Kalpathy-Cramer J, et al. Artificial intelligence in retinopathy of prematurity diagnosis. *Transl Vis Sci Technol*. 2020;9:5.
8. Li J, Huang K, Ju R, et al. Evaluation of artificial intelligence-based quantitative analysis to identify clinically significant severe retinopathy of prematurity. *Retina*. 2022;42:195–203.
9. Greenwald MF, Danford ID, Shahrawat M, et al. Evaluation of artificial intelligence-based telemedicine screening for retinopathy of prematurity. *J AAPOS*. 2020;24:160–162.
10. 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.
11. 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 Ophthalmol*. 2019;137:1029–1036.

12. Bellsmith KN, Brown J, Kim SJ, et al. Aggressive posterior retinopathy of prematurity: clinical and quantitative imaging features in a large North American cohort. *Ophthalmology*. 2020;127:1105–1112.
13. Campbell JP, Kim SJ, Brown JM, et al. Evaluation of a deep learning-derived quantitative retinopathy of prematurity severity scale. *Ophthalmology*. 2021;128:1070–1076.
14. Choi RY, Brown JM, Kalpathy-Cramer J, et al. Variability in plus disease identified using a deep learning-based retinopathy of prematurity severity scale. *Ophthalmol Retina*. 2020;4:1016–1021.
15. Campbell JP, Singh P, Redd TK, et al. Applications of artificial intelligence for retinopathy of prematurity screening. *Pediatrics*. 2021;147:e2020016618.
16. 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.
17. Blencowe H, Moxon S, Gilbert C. Update on blindness due to retinopathy of prematurity globally and in India. *Indian Pediatr*. 2016;53(Suppl 2):S89–S92.
18. The Committee for the Classification of Retinopathy of Prematurity. An international classification of retinopathy of prematurity. *Arch Ophthal*. 1984;102:1130–1134.
19. International Committee for the Classification of Retinopathy of Prematurity. The International Classification of Retinopathy of Prematurity revisited. *Arch Ophthalmol*. 2005;123:991–999.
20. 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.
21. International Committee for the Classification of Retinopathy of Prematurity. International Classification of Retinopathy of Prematurity, 3rd edition. *Ophthalmology*. 2021;128:e51–e68.
22. 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.
23. Campbell JP, Ryan MC, Lore E, et al. Diagnostic discrepancies in retinopathy of prematurity classification. *Ophthalmology*. 2016;123:1795–1801.
24. Chiang MF, Jiang L, Gelman R, et al. Interexpert agreement of plus disease diagnosis in retinopathy of prematurity. *Arch Ophthalmol*. 2007;125:875–880.
25. Gschließer A, Stifter E, Neumayer T, et al. Inter-expert and intra-expert agreement on the diagnosis and treatment of retinopathy of prematurity. *Am J Ophthalmol*. 2015;160:553–560.e3.
26. Hewing NJ, Kaufman DR, Chan RVP, Chiang MF. Plus disease in retinopathy of prematurity. *JAMA Ophthalmol*. 2013;131:1026–1027.
27. Kim SJ, Campbell JP, Kalpathy-Cramer J, et al. Accuracy and reliability of eye-based vs quadrant-based diagnosis of plus disease in retinopathy of prematurity. *JAMA Ophthalmol*. 2018;136:648–655.
28. Moleta C, Campbell JP, Kalpathy-Cramer J, et al. Plus disease in retinopathy of prematurity: diagnostic trends in 2016 vs. 2007. *Am J Ophthalmol*. 2017;176:70–76.
29. 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*. 2017;123:1–7.
30. 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;136:803–810.
31. 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.
32. 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;123:2345–2351.
33. Chiang MF, Quinn GE, Fielder AR, et al. International Classification of Retinopathy of Prematurity, Third Edition. *Ophthalmology*. 2021;128:e51–e68.
34. Rajan RP, Kohli P, Babu N, et al. Treatment of retinopathy of prematurity (ROP) outside International Classification of ROP (ICROP) guidelines. *Graefes Arch Clin Exp Ophthalmol*. 2020;258:1205–1210.
35. 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.
36. Campbell JP, Mathenge C, Cherwek H, et al. Artificial intelligence to reduce ocular health disparities: moving from concept to implementation. *Transl Vis Sci Technol*. 2021;10:19.
37. Chiang MF, Thyparampil PJ, Madigan D, et al. Interexpert agreement in identification of zone 1 disease in retinopathy of prematurity. Available at: https://iovs.arvojournals.org/article.aspx?articleid=2365536.