tvst

Special Issue

Artificial Intelligence in Retinopathy of Prematurity Diagnosis

Brittni A. Scruggs[1], R. V. Paul Chan[2], Jayashree Kalpathy-Cramer[3], Michael F. Chiang[1,4], and J. Peter Campbell[1,4]

[1] Casey Eye Institute, Department of Ophthalmology, Oregon Health & Science University, Portland, OR, USA
[2] Department of Ophthalmology, University of Illinois, Chicago, IL, USA
[3] Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital/Harvard Medical School, Boston, MA, USA
[4] Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA

Correspondence: J. Peter Campbell, Oregon Health & Science University, Department of Ophthalmology, 3375 SW Terwilliger Blvd, Portland, OR 97239, USA. e-mail: campbelp@ohsu.edu

Received: November 18, 2019
Accepted: November 21, 2019
Published: February 10, 2020

Keywords: retinopathy of prematurity; artificial intelligence; machine learning; pediatric retina

Citation: Scruggs BA, Chan RVP, Kalpathy-Cramer J, Chiang MF, Campbell JP. Artificial intelligence in retinopathy of prematurity diagnosis. Trans Vis Sci Tech. 2020;9(2):5, https://doi.org/10.1167/tvst.9.2.5

Retinopathy of prematurity (ROP) is a leading cause of childhood blindness worldwide. The diagnosis of ROP is subclassified by zone, stage, and plus disease, with each area demonstrating significant intra- and interexpert subjectivity and disagreement. In addition to improved efficiencies for ROP screening, artificial intelligence may lead to automated, quantifiable, and objective diagnosis in ROP. This review focuses on the development of artificial intelligence for automated diagnosis of plus disease in ROP and highlights the clinical and technical challenges of both the development and implementation of artificial intelligence in the real world.

Introduction

Retinopathy of prematurity (ROP) is a leading cause of childhood blindness worldwide. This vasoproliferative retinal disease affects extremely preterm infants, and low gestational age and low birth weight are the two strongest risk factors for the development of ROP.[1,2] The Multicenter Study of Early Treatment for Retinopathy of Prematurity (ET-ROP) study found that 68% of infants born <1251 g developed at least mild ROP.[3] A large population study of 184,700 preterm babies with ROP found that >10% of children became blind or severely visually impaired.[4] An estimated 20,000 babies go blind around the world each year,[4] and the Global Burden of Disease study estimated that 257,000 years lived with disability worldwide in 2010 were associated with visual impairment secondary to ROP.[5] This is despite the fact that blindness from ROP is often preventable with appropriate primary, secondary, and tertiary prevention.[6]

There are a number of challenges for ROP screening and diagnosis using current technology. ROP screening requires either bedside ophthalmoscopic screening or telemedicine using remote interpretation of digital fundus imaging; however, in a typical population, only 5% to 10% of babies within a screening population will develop sight-threatening ROP. Unfortunately, there

are often barriers to ensuring consistent screening of at risk babies especially in low- and middle-income countries, including inadequate equipment and training, personnel shortages, and inconsistent examinations between clinicians and/or institutions.[7] Moreover, owing to differences in the level of oxygen regulation, the population at risk in these regions is significantly higher.

Automated image analysis and deep learning (DL) systems for ROP have the potential to improve ROP care by (1) improving the efficiency, accuracy, and objectivity of diagnosis and (2) facilitating quantitative disease monitoring and risk prediction.[8] This review focuses on recent major advances, controversial topics, and knowledge gaps in artificial intelligence (AI), machine learning, and DL research as it relates to ROP diagnosis and management.

ROP Diagnosis and Current Limitations

ROP is classified based on the location, extent, and severity of disease according to the guidelines described by the International Classification of Retinopathy of Prematurity (ICROP) in 1984 and 2005.[9–11] The cryotherapy for ROP (CRYO-ROP) study defined threshold ROP (i.e., ROP requiring treatment) as five or more contiguous or eight total clock-hours of extraretinal fibrovascular proliferation (i.e., stage 3 ROP) in zones I or II in the presence of plus disease,[12] defined as venous dilatation and arteriolar tortuosity in two or more quadrants within the posterior pole that is greater than or equal to that of standard published photographs selected by expert consensus.[13] The subsequent ET-ROP trial further classified ROP into type 1 and type 2 prethreshold treatment to guide the treatment of infants before the development of threshold ROP.[3] Treatment of type 1 ROP, defined as any eye with 1) zone 1 disease (any stage) with plus disease; 2) zone 1 stage 3 disease without plus disease; or 3) zone 2 stage 2 or 3 disease with plus disease, remains the currently accepted treatment threshold for ROP. In 2005, an intermediate level (pre-plus) was introduced to the ICROP classification reflecting the fact that vascular changes present on a continuum. Vascular dilation and tortuosity increase with more posterior disease and with higher stage and extent of peripheral disease. The development of neovascularization of the retina (i.e., stage 3 ROP) can result in tractional membranes and retinal detachments (e.g., stages 4 and 5 ROP).[9] Thus, timely and accurate diagnosis is essential to prevent blindness from ROP.

There are several potential challenges to ensuring every at-risk baby is diagnosed accurately and in a timely manner. Besides wide disparities worldwide in the distribution of ophthalmologists between rural and urban settings and between countries, the diagnosis of ROP is based on subjective assessment of disease severity (zone, stage, and plus), and it is well-established that there is wide interobserver variability for all three components.[14] In a recent review, Ghergherehchi et al.[13] offered several potential explanations for variability in plus disease diagnosis: (1) attention to undefined vascular features (i.e., venous tortuosity which can be quite striking but which is not formally a diagnostic criterion); (2) differences in field of view (the standard photograph is quite narrow, but when we examine an eye we see a wider field of view and it is possible different examiners focus on different parts of the retina); (3) unfamiliarity with digital images; (4) magnification of the standard photograph; and (5) differences in plus disease thresholds along a continuum. As with any medical specialty, some experts are aggressive in their treatment plans, whereas others are more conservative. Significant interexpert variability across different regions has also been reported.[13,15,16] Further, most examiners do not routinely perform photography at the time of examination, which hinders their ability to objectively make comparisons across serial examinations or with other examiners. The lack of objective diagnosis of ROP and the high rates of interobserver variability have been a key limiting step in the development of AI technology in ROP.

Telemedicine in ROP

The ability to easily image a neonatal retina paved the way for telemedicine to provide a more efficient method to screen at-risk babies. There are now multiple examples of successful telemedicine programs in the United States and around the world.[17–21] The Stanford University Network for Diagnosis of Retinopathy of Prematurity (SUNDROP) trial found that telemedicine had 100% sensitivity, 99.8% specificity, 93.8% positive predictive value, and 100% negative predictive value for detection of treatment-warranted ROP.[22] Especially in regions where there are too few trained or willing ophthalmologists to manage ROP screening and treatment, telemedicine can allow a single provider to screen babies over a large geographic area. Digital fundus imaging also enabled for the first time large databases of retinal images of babies with ROP, which is an essential step for the development of AI for automated image-based diagnosis.[23]

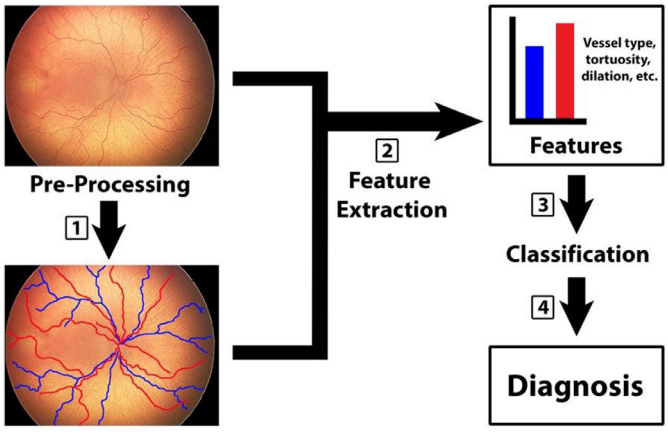

Figure 1. Machine learning in ROP. Early efforts to quantify the vascular changes in ROP used user-defined features of dilation and tortuosity (steps 1 and 2) without a computer-based classification (step 3). For example, the ROPTool used a semiautomated process to sum these features into a score[64] that correlated with the expert disease labels. Machine learning uses a classifier (step 3), such as a support vector machine, that learns the best relationship between the features (step 2) and the diagnosis (step 4).[26] Deep CNNs differ from traditional feature extraction and machine learning systems by allowing the CNN to learn features that best correlate the input image (step 1) with the diagnosis (4) with or without preprocessing but without explicit human defined features (step 2).[27,28,42]

The Development of AI Systems for ROP Diagnosis

Computer-based systems for ROP diagnosis have been around more than a decade.[24] Some of the first systems used hand-crafted measures of dilation and tortuosity to attempt to produce an objective metric of severity. There are more than a dozen ways to algorithmically measure tortuosity and dilation, and the earliest systems varied in the equations used and in the methods used to identify the blood vessels. In 2012, Wittenberg et al.[24] reviewed the four main computer-based ROP diagnostic systems at the time: ROPTool, Retinal Image multiScale Analysis, Vessel Map, and Computer Assisted Image Analysis of the Retina. All of these systems were feature extraction–based systems; that is, they had a system, manual or semiautomated, to quantify dilation and/or tortuosity into a value that had some variable diagnostic agreement with clinical diagnosis of ROP, as shown in Figure 1. As opposed to subsequent machine learning and DL techniques, there was no learning performed by the computer. Feature combinations and diagnostic cut-points were determined manually by the human users. Some of systems were semiautomatic with requirements of clinicians to label or select findings within the images, and, in general, the outputs did not correlate well enough with clinical diagnose to be widely used. However, these systems laid the groundwork for future machine learning.[25]

In 2015, Ataer-Cansizoglu et al.[26] reported a machine learning model for automated diagnosis of plus disease that performed as well as experts. Unlike prior systems, although this model used traditional features, a trained support vector machine (SVM) was used to determine the combination(s) of features and field of view that best correlated with expert diagnosis. A SVM is an type of supervised machine learning classifier that learns the best relationship between features and diagnosis.[26] The accuracy of the system was highest (95%) when it incorporated vascular tortuosity from both arteries and veins with the widest field of view (i.e., 6-disc diameter radius). Moreover, when the images were cropped to the field of view of the standard photograph, the accuracy was <85%, suggesting that clinicians factor in vascular information from a larger area of retina than the standard photograph depicted. Despite expert-level performance, this system was limited in its clinical usefulness because it required manual tracing and segmentation of the vessels as an input.[26]

To our knowledge, Worrall et al.[27] were the first to demonstrate fully automated plus disease diagnosis using a convolutional neural networks (CNN). Using a real-world dataset with multiple conflicting labels from experts, they found that their best classifier performed as well as some of the human graders, but they concluded that "a classifier can only ever be as good as its training data, as such we need to look to less human-dependent training data if we are to surpass human performance."[27] In 2018, Brown et al.[28] reported the results of a fully automated DL-based system for automated three-level diagnosis of plus disease. This deep CNN, called the i-ROP DL system, was trained on >5000 images with a single reference standard diagnosis (RSD) based on multiple expert diagnosis (consensus diagnosis of three independent image gradings and the ophthalmoscopic diagnosis). Using five-fold cross-validation, the area under the curve for plus disease diagnosis was 0.98. On an independent dataset of 100 images, the i-ROP DL system had higher diagnostic agreement with the RSD than seven of eight ROP experts. For the diagnosis of plus disease, the sensitivity and specificity of the algorithm were 93% and 94%, respectively. These sensitivity and specificity values increased to 100% and 94%, respectively, when including pre-plus disease or worse.[28]

Most studies to date have focused on computer-based systems to diagnose plus disease; however, there are a number of reports of using DL to grade ROP

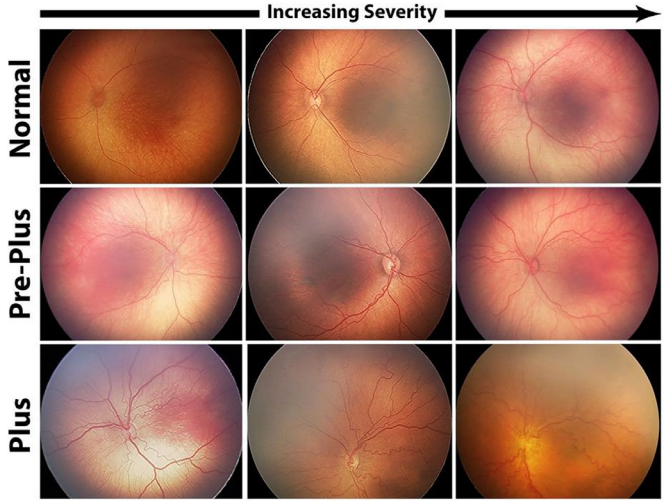

Figure 2. The continuum of vascular change. Each row depicts images with a label of normal vessels, pre-plus, or plus disease based on multiple (>3) expert consensus from ophthalmoscopy and image grading. For any row, there is increasing tortuosity and dilation of the vessels from left to right demonstrating a continuous range of vascular change within current ordinal categories of disease. Quantification of plus disease using AI has shown promise in the diagnosis and monitoring of disease change over time.

severity category or classify zone or stage specifically.[29,30] For example, a DL system called DeepROP achieved a sensitivity of 96.62% (95% confidence interval, 92.29%–98.89%) and a specificity of 99.32% (95% confidence interval, 96.29%–9.98%) for the detection of ROP (vs no ROP).[31,32] Zhao et al.[30] reported the development of a DL system that can automatically draw the border of zone 1 on a fundus image as a diagnostic aid. Mulay et al.[29] were the first to report the identification of a peripheral ROP ridge (stage) directly in a fundus image. Thus, DL seems to hold promise for automated and objective diagnosis of ROP in digital fundus images; however, none of these systems are yet available for clinical use.

A Continuous Severity Score for ROP

Vascular disease in ROP presents on a continuum, as shown in Figure 2, and experts have been shown to have poor absolute agreement on classification (normal, plus, or pre-plus), but good relative agreement on disease severity.[33] This finding motivated the development of a continuous vascular severity score using the i-ROP DL system. Redd et al. reported that a scale from 1 to 9 could accurately detect type 1 ROP with an area under the curve of 0.95 and, in theory, could decrease the number of ophthalmoscopic examinations by 80% in a real-world telemedicine program because most babies will have no or mild disease.[34] More recently, Taylor et al.[35] implemented the i-ROP DL algorithm to assign a continuous ROP vascular severity score (1–9) and to classify images based on severity: no ROP, mild ROP, type 2 ROP, and pre-plus disease, or type 1 ROP. Using reference standard diagnoses, this retrospective cohort study concluded that the continuous ROP vascular score was associated both with the ICROP category of disease at a single point in time, and clinical progression of ROP over time.[35] Gupta et al.[36] showed that these continuous scores reflected posttreatment regression in eyes with treatment requiring-ROP. Using i-ROP data, this group also found that eyes requiring multiple treatment sessions (laser or intravitreal injection of bevacizumab) had higher pretreatment ROP vascular severity scores compared with eyes requiring only a single treatment, suggesting that treatment failure may be related to more aggressive disease or disease treated at a later stage.[36] Using a similar automated quantitative severity scale for ROP diagnosis may help to optimize treatment regimens by better predicting the preterm infants that are at risk for treatment failure and disease recurrence.[36] Future clinical trials may use a quantitative scale to help evaluate treatment thresholds.

Challenges to AI Implementation

Ting et al.[37] recently published on the clinical and technical challenges of DL applications in ophthalmology. Although AI holds great promise for improving care for ROP, the gap between scientific discovery and clinically useful implementation of technology remains wide. The main potential challenges hindering the deployment of DL systems include ensuring generalizability, explainability, and overcoming regulatory and medicolegal issues.[37] We discuss these topics as they apply to AI for ROP diagnosis.

Generalizability

It is well-established that CNNs often do not generalize well to unseen data. This can be due to differences in the populations being studied, differences in the ways the images were acquired, technical differences between camera systems, and other unknown factors. ROP phenotypes that are seen in low- and middle-income countries are often qualitatively different from phenotypes seen in North America; therefore, it is critical to validate the performance of AI systems on the target population before clinical use. At a minimum, AI systems need to be further validated using datasets widely tested in different populations (i.e., patients

with different levels of pigmentation, image quality) and on different devices (e.g., various cameras, fields of view). Open access datasets and software could alleviate such issues and encourage timely clinical application for ROP diagnosis.[38]

Explainability

One common criticism for AI algorithms is that acceptance by physicians and patients may be reduced owing to the inability to explain how the algorithm arrived at a conclusion. We agree with efforts to improve explainability; however, there are problems with all of the currently available techniques to explain the workings of CNNs.[39] Moreover, it is interesting to consider whether and how the same standard of explainability is applied to the cognitive reasoning of clinicians. When human experts are asked to explain what features they used to arrive at a diagnosis they often (1) disagree[40] and (2) cite different features based on clinical judgment that led to the diagnosis.[41] It may be that the eventual clinical adoption of technology is predicated on developing methodology for understanding the high-level features CNNs use for discrimination. But, it may also be that clinical adoption occurs after rigorous clinical validation demonstrating improvement in outcomes despite a lack of complete transparency. The primary goal of AI research in ROP is to improve outcomes. As a field, we may need to determine whether explainability or improved outcomes is our primary goal given the black box nature of clinical diagnosis in general.[38] This is analogous to the fear of self-driving cars causing a fatal accident. Technology never will be perfect, but it may be better than the status quo.

One approach to have more explainable AI is to combine DL methods with traditional feature extraction, and several groups have attempted this for plus disease.[42,43] Mao et al.[42] trained a DL network to segment retinal vessels and the optic disc and to diagnosis plus disease based on automated quantitative characterization of pathologic features, such as vessel tortuosity, width, fractal dimension, and density. Graziani et al.[43] compared the black box CNN features with known hand-crafted features using regression concept vectors. Although these approaches do not explain how the CNNs make their decisions, both methods demonstrated that disease-specific features (e.g., dilation and tortuosity) correlate with the CNN's diagnostic outcome and may provide enough face validity to satisfy clinicians and regulators. In addition, these quantitative metrics may be useful for disease monitoring over time.

Regulatory and Medicolegal Issues

ROP care is the highest medicolegal risk within ophthalmology; thus, any discussion of AI assistance for ROP care must consider this issue. As AI enters clinical medicine, there is increasing awareness of the need to adjudicate liability from care decisions informed by AI.[44] To this end, there is a distinction between autonomous and assistive AI systems. In the former, such as IDx-DR,[45] decisions are based solely on the output of the AI system. In the latter, the output is used as an aid to clinical diagnosis by a physician. The US Food and Drug Administration is rapidly innovating their methods of evaluation to ensure safe implementation of these technologies into clinical care given a precise indication for use and evidence of effectiveness in a real world population with consideration of these challenges regarding generalization, explainability, and bias.[46] As these technologies become more commonplace, the regulatory requirements will likely continue to evolve, as will the medicolegal implications.

AI for ROP Education

Multiple computer-based analyses have also been used to explain interobserver variability in diagnosis, which may help to standardize ROP education. Chiang et al. and Gelman et al. measured agreement and accuracy of plus disease diagnosis among ROP experts compared with the Retinal Image multi-Scale Analysis system.[47–50] They found that experts often focused on qualitative features outside of the published definitions of ROP.[50] Ataer-Cansizoglu et al.[40,51] analyzed 66 image features and quantified inter-expert variability in evaluating these ROP features and found significant differences in the features used between different examiners. A quantitative analysis of vascular features was also performed by Woo et al.[52] in 2015 looking specifically at aggressive posterior ROP, a rapidly progressive form of ROP predominantly in the posterior region with marked plus disease. Woo et al.[52] found that aggressive posterior ROP was often misdiagnosed, and interexpert agreement was poor when comparing retinal images of normal, plus, or aggressive posterior ROP eyes. Some clinicians refer to a two-quadrant rule for the level of disease in ROP, that is, an eye has plus disease if two or more quadrants analyzed individually have findings greater than the standard photography. Kim et al.[53] compared quadrant diagnosis versus eye-level diagnosis and found lower accuracy when clinicians diagnosed plus disease one quadrant at a time, suggesting that clinicians subconsciously

evaluate the whole eye, even when they intend to carefully evaluate plus by quadrant.

If ROP experts often do not agree on how to diagnose ROP, or on the diagnosis of individual babies, it is not surprising that ROP trainees find the task of ROP diagnosis challenging as well. It is well-established that ophthalmology graduates complete residency, as well as retina and pediatric ophthalmology fellowship programs, without confidence in their ability to diagnose ROP.[54–56] A recent survey of 95 international ophthalmology trainees found that <33% of learners performed ROP screenings under direct supervision.[57] Chan et al.[58] demonstrated that there was significant variability in diagnostic accuracy among retinal fellows when analyzing ROP images compared with reference standard diagnoses. Both Chan et al.[58] and Myung et al.[59] demonstrated the inconsistent accuracy of detecting type 2 ROP and treatment-requiring ROP by fellows. These studies raise serious concerns for ROP screening performed by inexperienced examiners, and there are no accepted criteria for minimum necessary supervision, examinations, treatments, and so on for clinical competency for ROP cases. Improved global education for ROP training is necessary to ensure treatments are performed adequately. The development of AI systems for automated diagnosis in ROP may facilitate the incorporation of these algorithms within medical training to standardize ROP education through tele-education, and perhaps, ROP certification.[56]

Recommendations for Future Research in AI Applied to ROP

The 2018 US Food and Drug Administration approval of IDX for diabetic retinopathy screening in the primary care setting set the stage for clinical use of AI[45,60]; this technology, which assesses image quality and detects diabetic retinopathy based on biomarkers (e.g., hemorrhages, exudates), was the first automated screening in all medical specialties with no need for physician input.[45] As discussed, similar technology exists for the detection of plus disease and ROP staging, yet there are several hurdles that ROP researchers must tackle to advance this AI technology into the clinical arena. First, the automated DL-enhanced algorithms must first be integrated into commonly used cameras (e.g., RetCams) or into cloud-based systems. This effort will require collaborations with engineers, system developers, and, in some cases, the development of infrastructure for telemedicine. AI-assisted ROP screening may be most useful in regions that have a dearth of trained ROP providers. Low-cost fundus cameras, including smartphone-based devices, have also been proposed to lower the capital costs required for telemedicine infrastructure; however, this strategy has not been validated on a large scale.[8] Moreover, validation studies in the intended population need to be performed, because the image quality, fundus pigmentation, prevalence of disease, and ROP phenotype can vary dramatically between geographic regions.

Experts have poor absolute agreement on ROP vascular disease classification (normal, plus, or pre-plus) but good relative agreement on ROP disease severity.[33] As previously described, this finding motivated the development of a continuous vascular severity score using the i-ROP DL system. This continuous ROP vascular score provided a more quantitative diagnosis for plus disease and was associated with both the ICROP category of disease at a single point in time and clinical progression of ROP over time.[35] The quantitative ROP vascular severity group (1–9), which was described and evaluated by Gupta et al.[36] and Taylor et al.[35] consistently reflects plus disease presence (i.e. more advanced ROP stages are associated with increased plus score)[2], ROP disease progression, and posttreatment regression. However, it is not uncommon for preterm infants to have comorbidities, such as pulmonary hypertension, which cause increased retinal vessel dilation and tortuosity. Thus, the ROP vascular severity score and the CNNs that detect these subtle changes should be tested for the ability to differentiate between ROP- and non–ROP-related vascular changes. More precise definitions of AP-ROP also need to be developed for the purpose of training CNNs to distinguish AP-ROP from either ROP or plus disease.

Demonstrating abnormal regions of the retina or affected vessels as a detailed figure or heat map would be helpful for clinicians to identify areas at risk for progression. Such technology would likely have clinical implications for older children with an ROP history to determine which patients with persistent avascular retina are at risk of further sequelae and need prophylactic laser treatment. Further, there are structural changes on optical coherence tomography (OCT) that correlate with severe ROP.[61] OCT and OCT angiography could identify the earliest structural and angiographic signs preceding disease progression to advanced stages of ROP, such as early vitreoretinal traction preceding retinal detachment[61] and AI-assisted OCT image acquisition and interpretation is an active area of research.[62] Maldonado et al.[63] demonstrated three-dimensional structural changes in patients with severe ROP. DL-based

OCT for improving the efficiency in image acquisition, auto-segmentation, and extraction of quantifiable biomarkers in ROP represents a potential area of future research.

Conclusions

There are multiple reasons that AI may improve diagnosis of ROP worldwide. ROP diagnosis has been shown to be both subjective and qualitative, and AI may add objectivity and improve accuracy. The screening burden in low- and middle-income countries is often too high to be met with the available workforce, and AI-assisted disease screening may be a paradigm-shifting strategy to improve efficiency. AI could provide objectivity to ROP education, and AI-based augmented reality may represent a future pedagogical tool to improve trainee performance on ROP diagnosis and treatment. AI has enabled the development of an ROP severity score that correlates with ICROP disease classification and shows promise for quantitative disease monitoring, improved risk prediction, and posttreatment identification of treatment failure and recurrence. Deployed into a telemedicine system, AI could significantly benefit ROP clinical care by improving the efficiency, accuracy, and objectivity of ROP diagnosis. AI may also improve early identification of severe ROP before the development of retinal detachment. Finally, at the population level, quantitative diagnosis may enable epidemiologic monitoring of disease severity between neonatal care units within a geographic region and over time (Redd, T et al., *IOVS*, 2019, ARVO E-Abstract, A0207).

Acknowledgments

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 and a Career Development Award (JPC) from Research to Prevent Blindness (New York, NY). Also supported by the Heed Foundation.

Disclosure: **B.A. Scruggs,** None; **R.V.P. Chan,** i-ROP DL system (P); **J. Kalpathy-Cramer,** i-ROP DL system (P); **M.F. Chiang,** i-ROP DL system (P); **J.P. Campbell,** i-ROP DL system (P)

References

1. Flynn JT, Bancalari E, Bachynski BN, et al. Retinopathy of prematurity. Diagnosis, severity, and natural history. *Ophthalmology*. 1987;94:620–629.
2. Fierson WM, American Academy of Pediatrics Section on Ophthalmology, American Academy of Ophthalmology, American Association for Pediatric Ophthalmology and Strabismus, American Association of Certified Orthoptists. Screening examination of premature infants for retinopathy of prematurity. *Pediatrics*. 2018;142.
3. Good WV, Early Treatment for Retinopathy of Prematurity Cooperative Group. Final results of the Early Treatment for Retinopathy of Prematurity (ETROP) randomized trial. *Trans Am Ophthalmol Soc*. 2004;102:233–248; discussion 248–250.
4. Blencowe H, Lawn JE, Vazquez T, Fielder A, Gilbert C. Preterm-associated visual impairment and estimates of retinopathy of prematurity at regional and global levels for 2010. *Pediatr Res.* 2013;74(suppl 1):35–49.
5. Vos T, Flaxman AD, Naghavi M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*. 2012;380:2163–2196.
6. Norman M, Hellström A, Hallberg B, et al. Prevalence of severe visual disability among preterm children with retinopathy of prematurity and association with adherence to best practice guidelines. *JAMA Netw Open*. 2019;2:e186801.
7. Gilbert C. Retinopathy of prematurity: a global perspective of the epidemics, population of babies at risk and implications for control. *Early Hum Dev*. 2008;84:77–82.
8. Valikodath N, Cole E, Chiang MF, Campbell JP, Chan RVP. Imaging in retinopathy of prematurity. *Asia Pac J Ophthalmol (Phila)*. 2019;8:178–186.
9. International Committee for the Classification of Retinopathy of Prematurity. The international classification of retinopathy of prematurity revisited. *Arch Ophthalmol*. 2005;123:991–999.
10. An International Classification of Retinopathy of Prematurity. II. The classification of retinal detachment. The international committee for the classification of the late stages of retinopathy of prematurity. *Arch Ophthalmol.* 1987;105:906–912.
11. An International Classification of Retinopathy of Prematurity. The committee for the classification

of retinopathy of prematurity. *Arch Ophthalmol*. 1984;102:1130–1134.
12. Multicenter Trial of Cryotherapy for Retinopathy of Prematurity. One-year outcome–structure and function. Cryotherapy for retinopathy of prematurity cooperative group. *Arch Ophthalmol.* 1990;108:1408–1416.
13. Ghergherehchi L, Kim SJ, Campbell JP, Ostmo S, Chan RVP, Chiang MF. Plus disease in retinopathy of prematurity: more than meets the ICROP? *Asia Pac J Ophthalmol (Phila)*. 2018;7:152–155.
14. 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;134:651–657.
15. 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.
16. 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;32:74–80.
17. Fierson WM, Capone A, American Academy of Pediatrics Section on Ophthalmology, American Academy of Ophthalmology, American Association of Certified Orthoptists. Telemedicine for evaluation of retinopathy of prematurity. *Pediatrics*. 2015;135:e238–e254.
18. Quinn GE, Ying GS, Daniel E, et al. Validity of a telemedicine system for the evaluation of acute-phase retinopathy of prematurity. *JAMA Ophthalmol*. 2014;132:1178–1184.
19. Weaver DT, Murdock TJ. Telemedicine detection of type 1 ROP in a distant neonatal intensive care unit. *J AAPOS*. 2012;16:229–233.
20. Chiang MF, Melia M, Buffenn AN, et al. Detection of clinically significant retinopathy of prematurity using wide-angle digital retinal photography: a report by the American Academy of Ophthalmology. *Ophthalmology*. 2012;119:1272–1280.
21. Ells AL, Holmes JM, Astle WF, et al. Telemedicine approach to screening for severe retinopathy of prematurity: a pilot study. *Ophthalmology*. 2003;110:2113–2117.
22. Fijalkowski N, Zheng LL, Henderson MT, et al. Stanford University Network for Diagnosis of Retinopathy of Prematurity (SUNDROP): five years of screening with telemedicine. *Ophthalmic Surg Lasers Imaging Retina*. 2014;45:106–113.
23. Chee RI, Darwish D, Fernandez-Vega A, et al. Retinal telemedicine. *Curr Ophthalmol Rep*. 2018;6:36–45.
24. Wittenberg LA, Jonsson NJ, Chan RV, Chiang MF. Computer-based image analysis for plus disease diagnosis in retinopathy of prematurity. *J Pediatr Ophthalmol Strabismus*. 2012;49:11–19; quiz 10, 20.
25. Wilson CM, Wong K, Ng J, Cocker KD, Ells AL, Fielder AR. Digital image analysis in retinopathy of prematurity: a comparison of vessel selection methods. *J AAPOS*. 2012;16:223–228.
26. Ataer-Cansizoglu E, Bolon-Canedo V, Campbell JP, et al. Computer-based image analysis for plus disease diagnosis in retinopathy of prematurity: performance of the "i-ROP" system and image features associated with expert diagnosis. *Transl Vis Sci Technol*. 2015;4:5.
27. Worrall DE, Wilson CM, Brostow GJ. Automated retinopathy of prematurity case detection with convolutional neural networks. *Deep Learning and Data Labeling for Medical Applications*; 2016; Athens, Greece.
28. 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.
29. Mulay S, Ram K, Sivaprakasam M, Vinekar A. Early detection of retinopathy of prematurity stage using deep learning approach. *Paper presented at: SPIE Medical Imaging 2019*, 2019; San Diego, California, United States.
30. Zhao J, Lei B, Wu Z, et al. *A Deep Learning Framework for Identifying Zone I in RetCam Images*. Vol. 7. IEEE Access 2019:103530–103537.
31. Wang J, Ju R, Chen Y, et al. Automated retinopathy of prematurity screening using deep neural networks. *EBioMedicine*. 2018;35:361–368.
32. Hu J, Chen Y, Zhong J, Ju R, Yi Z. Automated analysis for retinopathy of prematurity by deep neural networks. *IEEE Trans Med Imaging*. 2019;38:269–279.
33. 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.
34. 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.

35. 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.
36. 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.
37. Ting DSW, Peng L, Varadarajan AV, et al. Deep learning in ophthalmology: The technical and clinical considerations. *Prog Retin Eye Res*. 2019.
38. Reid JE, Eaton E. Artificial intelligence for pediatric ophthalmology. *Curr Opin Ophthalmol*. 2019;30:337–346.
39. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. In. *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*. Montreál, Canada. 2018.
40. Ataer-Cansizoglu E, Kalpathy-Cramer J, You S, Keck K, Erdogmus D, Chiang MF. Analysis of underlying causes of inter-expert disagreement in retinopathy of prematurity diagnosis. Application of machine learning principles. *Methods Inf Med*. 2015;54:93–102.
41. Hewing NJ, Kaufman DR, Chan RV, Chiang MF. Plus disease in retinopathy of prematurity: qualitative analysis of diagnostic process by experts. *JAMA Ophthalmol*. 2013;131:1026–1032.
42. Mao J, Luo Y, Liu L, et al. Automated diagnosis and quantitative analysis of plus disease in retinopathy of prematurity based on deep convolutional neural networks. *Acta Ophthalmol*. 2019.
43. Graziani M, Brown JM, Andrearczyk V, et al. Improved interpretability for computer-aided severity assessment of retinopathy of prematurity. *SPIE Medical Imaging*; 2019; San Diego, California.
44. Shah NH, Milstein A, Bagley PhD SC. Making machine learning models clinically useful. *JAMA*. 2019.
45. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med*. 2018;1:39.
46. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25:44–56.
47. Gelman R, Jiang L, Du YE, Martinez-Perez ME, Flynn JT, Chiang MF. Plus disease in retinopathy of prematurity: pilot study of computer-based and expert diagnosis. *J AAPOS*. 2007;11:532–540.
48. Chiang MF, Gelman R, Jiang L, Martinez-Perez ME, Du YE, Flynn JT. Plus disease in retinopathy of prematurity: an analysis of diagnostic performance. *Trans Am Ophthalmol Soc*. 2007;105:73–84; discussion 84–75.
49. Chiang MF, Gelman R, Martinez-Perez ME, et al. Image analysis for retinopathy of prematurity diagnosis. *J AAPOS*. 2009;13:438–445.
50. Chiang MF, Gelman R, Williams SL, et al. Plus disease in retinopathy of prematurity: development of composite images by quantification of expert opinion. *Invest Ophthalmol Vis Sci*. 2008;49:4064–4070.
51. Bolón-Canedo V, Ataer-Cansizoglu E, Erdogmus D, et al. Dealing with inter-expert variability in retinopathy of prematurity: A machine learning approach. *Comput Methods Programs Biomed*. 2015;122:1–15.
52. Woo R, Chan RV, Vinekar A, Chiang MF. Aggressive posterior retinopathy of prematurity: a pilot study of quantitative analysis of vascular features. *Graefes Arch Clin Exp Ophthalmol*. 2015;253:181–187.
53. 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.
54. Patel SN, Martinez-Castellanos MA, Berrones-Medina D, et al. Assessment of a tele-education system to enhance retinopathy of prematurity training by international ophthalmologists-in-training in Mexico. *Ophthalmology*. 2017;124:953–961.
55. Campbell JP, Swan R, Jonas K, et al. Implementation and evaluation of a tele-education system for the diagnosis of ophthalmic disease by international trainees. *AMIA Annu Symp Proc*. 2015;2015:366–375.
56. Chan RV, Patel SN, Ryan MC, et al. The Global Education Network for Retinopathy of Prematurity (Gen-Rop): development, implementation, and evaluation of a novel tele-education system (an American Ophthalmological Society Thesis). *Trans Am Ophthalmol Soc*. 2015;113:T2.
57. Al-Khaled T, Mikhail M, Jonas KE, et al. Training of residents and fellows in retinopathy of prematurity around the world: an international web-based survey. *J Pediatr Ophthalmol Strabismus*. 2019;56:282–287.
58. Paul Chan RV, Williams SL, Yonekawa Y, Weissgold DJ, Lee TC, Chiang MF. Accuracy of retinopathy of prematurity diagnosis by retinal fellows. *Retina*. 2010;30:958–965.

59. Myung JS, Paul Chan RV, Espiritu MJ, et al. Accuracy of retinopathy of prematurity image-based diagnosis by pediatric ophthalmology fellows: implications for training. *J AAPOS*. 2011;15:573–578.
60. Abràmoff MD, Folk JC, Han DP, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmol*. 2013;131:351–357.
61. Campbell JP. Why do we still rely on ophthalmoscopy to diagnose retinopathy of prematurity? *JAMA Ophthalmol*. 2018;136:759–760.
62. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24:1342–1350.
63. Maldonado RS, Toth CA. Optical coherence tomography in retinopathy of prematurity: looking beyond the vessels. *Clin Perinatol*. 2013;40:271–296.
64. Abbey AM, Besirli CG, Musch DC, et al. Evaluation of screening for retinopathy of prematurity by ROPtool or a lay reader. *Ophthalmology*. 2016;123:385–390.