Published in final edited form as:

Ophthalmology. 2017 May ; 124(5): e46–e47. doi:10.1016/j.ophtha.2016.11.041.

REPLY

J. Peter Campbell, MD, MPH[1], **Michael C. Ryan, MD**[1], **R.V. Paul Chan, MD**[2], **Michael F. Chiang, MD**[1,3]

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

[2]Department of Ophthalmology and Visual Sciences, Illinois Eye and Ear, Infirmary, University of Illinois at Chicago, Chicago, Illinois

[3]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon

We thank Yeo et al for their comments. To clarify, although our paper mentioned that the clinical (ophthalmoscopic) examination has “long been considered the gold standard,” we intentionally did not promote the ophthalmoscopic examination as a gold standard in this paper, and used the term “agreement” rather than “accuracy.” This word choice was to emphasize the absence of a perfect gold standard. Instead of focusing on “which was more accurate,” our study addressed “was there a difference” between image-based and ophthalmoscopic classifications. For our ongoing cohort study exploring imaging and genetic analysis of retinopathy of prematurity (ROP), we have developed a reference standard diagnosis for each examination that integrates the results of the clinical ophthalmoscopic classification performed by experts at 7 participating centers with image-based classification by 3 experienced readers. The dataset for this paper’s analysis was derived from the classifications during that process.

In response to the 3 points raised by Yeo et al: (1) Could some of the observed difference between image classifications be due to intra-expert variability? As the authors note, there are several previous papers (including several by our group) that have addressed intragrader variability and that have generally found that it is very high.[1,2] For this analysis, because the data were obtained as part of the larger data acquisition for the cohort study, we did not repeat any classifications to assess intragrader reliability.

(2) Could inexperience with or limitations of the imaging technique explain some of the discrepancy? We agree that experience with fundus photography can affect the quality, reliability, and readability of images and have similarly noted anecdotally that dynamic evaluation can be helpful as part of the bedside examination by a physician. In terms of generalizing the findings to other telemedicine clinical and research programs, dynamic

Correspondence: Michael F. Chiang, MD, Casey Eye Institute, Oregon Health & Science University, 3375 SW Terwilliger Boulevard, Portland, OR 97239. chiangm@ohsu.edu.

Financial Disclosures: The authors made the following disclosures:
M.F.C.: Unpaid member, Scientific Advisory Board — Clarity Medical Systems; Consultant — Novartis
R.V.P.C.: Consultant — Visunex Medical Systems

(video) documentation of findings are less standardized. Therefore, we felt that classification using multiple standardized still images would have greater generalizability and clinical relevance. In addition, all physicians and staff in the i-ROP study group are familiar with the use and limitations of wide-angle imaging, and images determined to be unreadable were not included in this dataset. Thus, we do not believe that technical limitations explain our findings, especially because there were similar rates of disagreement between image-based classifications and the clinical examination, for both graders.

(3) Is disagreement even for milder levels of ROP still relevant for follow-up and resource utilization? We also agree that even though agreement on types 1 and 2 disease was high, there may be relevant implications for imperfect agreement on lower levels of pathology. Our main purpose in writing this paper was to determine the baseline level of agreement that might be expected under controlled circumstances (i.e., research study with expert physician graders) using available technology, recognizing that for a number of reasons noted in the paper, agreement in the real world may be lower.

One of the most striking findings of this paper was that the 2 expert physician graders (M.F.C. and R.V.P.C.), who have collaborated on ROP work for nearly 10 years, would have such limited agreement in image-based classification of stage. We note that the cryotherapy for ROP study showed 12% disagreement among experts performing clinical ophthalmoscopic examinations regarding the presence versus absence of threshold ROP on the same infants, which was higher than the levels of discrepancy in our current study regarding presence of treatment-requiring ROP.[3] Taken together, we feel that these findings suggest that interexpert discrepancy in ROP diagnosis may result from a combination of limitations in fundus photography with inherent subjectivity in the nature of ophthalmic diagnosis and the qualitative definition of parameters such as zone, stage, and plus disease. For diagnosis of plus disease, we have recently published another explanation for some of the interexpert variability related to the parsing of a continuous phenotype into ordinal categories.[4,5] For zone and stage, we believe that these findings support the need for research to improve diagnostic precision and accuracy using other diagnostic modalities such as fluorescein angiography, optical coherence tomography, and perhaps eventually optical coherence tomography angiography, as well as methods such as computer-based image analysis.

Acknowledgments

Supported by NIH grants R01 EY19474, R21 EY22387, and P30 EY10572 from the National Institutes of Health (Bethesda, Maryland), by unrestricted departmental funding from Research to Prevent Blindness (New York, New York), by the iNsight Foundation (New York, New York), and by grant 1622679 from the National Science Foundation (Arlington, Virginia).

References

1. Chiang MF, Wang L, Busuioc M, et al. Telemedical retinopathy of prematurity diagnosis: accuracy, reliability, and image quality. Arch Ophthalmol. 2007;125:1531–1538. [PubMed: 17998515]
2. Daniel E, Quinn GE, Hildebrand PL, et al. Validated system for centralized grading of retinopathy of prematurity: telemedicine approaches to evaluating acute-phase retinopathy of prematurity (e-ROP) study. JAMA Ophthalmol. 2015;133:675–682. [PubMed: 25811772]

3. 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. [PubMed: 12427059]

4. 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. [PubMed: 27591053]

5. 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. [PubMed: 27566853]