RESPONSE TO COMMENT ON LEE ET AL.

Multicenter, Head-to-Head, Real-World Validation Study of Seven Automated Artificial Intelligence Diabetic Retinopathy Screening Systems. Diabetes Care 2021;44:1168–1175

Aaron Y. Lee,[1,2,3] Cecilia S. Lee,[1,2] Matthew S. Hunt,[1] Ryan T. Yanagihara,[1] Marian Blazes,[1] and Edward J. Boyko[4,5]

Diabetes Care 2021;44:e108–e109 | https://doi.org/10.2337/dci21-0007

We thank Dr. Soliz for his expedient comments (1) regarding our work evaluating seven artificial intelligence (AI) algorithms in a head-to-head fashion (2). We welcome the opportunity to clarify these points of concern.

We agree that the two U.S. Food and Drug Administration–approved diabetic retinopathy (DR) systems, one of which was tested in our study, use a threshold of moderate nonproliferative DR (NPDR) or higher for referral. We set the threshold to any DR versus no DR because that represents the clinical practice at the Veterans Affairs (VA) teleretinal system, as stated in our methods. All participating companies were aware of the study threshold prior to submitting their algorithms and were allowed to adjust their algorithms accordingly. As the VA manages one of the largest teleretinal screening systems in the U.S., we believe this threshold constitutes a reasonable real-world cut point for DR detection. Algorithm performance at this threshold is important to investigate because this threshold would most likely be used for referral decisions if these AI systems were to be used in the VA.

Diabetic macular edema (DME) represents an important trigger for referrals in teleretinal systems and is always accompanied by retinopathy of any degree including less than moderate NPDR. Had we used the recommended referral threshold of moderate NPDR or higher, omitting DME as a referral criterion would have constituted a serious error. However, it is not possible to have DME without having some level of retinopathy, and thus referral of all cases of DR (including mild DR) would encompass all patients with DME in the reference standard grading.

We also agree that most studies report metrics at the case level. In our study, we report all metrics at the case level for each encounter level. As we stated in our methods, we asked the companies to use all images available per patient encounter to reach a referral decision, including ungradable images, which resulted in one referral decision for each time a patient underwent screening. If one patient had a few ungradable images but multiple acceptable-quality images sufficient to determine whether referral was needed, then the overall encounter was still graded appropriately as per clinical standards. Therefore, the performance we reported in our article was based on each case and not each image. This was true both for the human VA teleretinal grader and for the AI algorithm output.

As stated in our discussion, the primary difference between the two sites was the use of routine dilation in Atlanta versus nonmydriatic imaging in Seattle. We also note that we performed stratified analyses by location in Fig. 1, and this provides a valuable opportunity to understand the impact of mydriasis on AI algorithms. We believe that this difference and other discrepancies in real-world imaging protocols between screening locations represent important variations for understanding the performance of AI algorithms.

We again thank Dr. Soliz for his comments.

Funding. A.Y.L. is supported by the U.S. Food and Drug Administration. This material is the result of work supported with resources and the use of facilities at VA Puget Sound and the VHA Innovation Ecosystem. This study was supported by National Institutes of Health grants K23EY029246 and R01AG060942 and an unrestricted grant from Research to Prevent Blindness.

The sponsors/funding organizations had no role in the design or conduct of this research. The contents do not represent the views of the U.S. Department of Veterans Affairs, the U.S. Food and Drug Administration, or the U.S. Government.

[1]*Department of Ophthalmology, University of Washington School of Medicine, Seattle, WA*
[2]*Department of Ophthalmology, Puget Sound Veteran Affairs, Seattle, WA*
[3]*eScience Institute, University of Washington, Seattle, WA*
[4]*Seattle Epidemiologic Research and Information Center, Department of Veteran Affairs Medical Center, Seattle, WA*
[5]*Department of Medicine, University of Washington, Seattle, WA*

Corresponding author: Aaron Y. Lee, leeay@uw.edu

Duality of Interest. A.Y.L. reports grants from Santen, Carl Zeiss Meditec, and Novartis and personal fees from Genentech, Topcon, and Verana Health outside of the submitted work. M.S.H. reports a commercial relationship with Boston Scientific, patent ID US20180228896A1. E.J.B. reports personal fees from Bayer AG. No other potential conflicts of interest relevant to this article were reported.

References

1. Soliz P. Comment on Lee et al. Multicenter, head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. Diabetes Care 2021;44:1168–1175 (Letter). Diabetes Care 2021;44:e107. DOI: 10.2337/dc21-0151

2. Lee AY, Yanagihara RT, Lee CS, et al. Multicenter, head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. Diabetes Care 2021; 44:1168–1175