# Eyenuk Commentary on

## "Multicenter, Head-to-Head, Real-World Validation Study of Seven Automated Artificial Intelligence Diabetic Retinopathy Screening Systems." *Diabetes Care*, January 4, 2021

**Background**: A multicenter, non-interventional device retrospective validation study evaluating a total of 311,604 retinal images from 23,724 veterans who presented for teleretinal diabetic retinopathy (DR) screening at the Veterans Affairs (VA) Puget Sound Health Care System (HCS) or Atlanta VA HCS from 2006 to 2018. Five companies provided seven algorithms, including the FDA-cleared EyeArt technology from Eyenuk, Inc (www.eyenuk.com).

We would like to congratulate the study team for completing such a major and complex undertaking, involving invitation of 23 artificial intelligence (AI) companies, analyzing 7 algorithms from 5 companies, and extracting data from 2 VA health systems. This study has trailblazed a model for large scale head-to-head AI algorithm comparisons in the future. Eyenuk is also appreciative of the fact that our EyeArt algorithm has now been validated on the large VA dataset.

**Conclusions that can be drawn from the study**: To our knowledge, this study is the largest head-to-head AI algorithm comparison analysis published to-date. The study analyzed seven AI algorithms under the same analytical protocol by using the same retinal images. As a result, the study appropriately identified the best performing algorithm (*Algorithm G*, which was the EyeArt technology from Eyenuk).

- *Algorithm G* was the only algorithm that was statistically indistinguishable from the standard of care.
- Most notably, *Algorithm G* did not miss a single case of moderate or severe non-proliferative or proliferative DR in the arbitration set, achieving 100% sensitivity for each (Figure 2 from the publication[1]).
- *Algorithm G* also enabled the highest amount of cost savings to the VA teleretinal screening program.

**Limitations from the study that prevent other conclusions**: This University of Washington VA study included analysis with some limitations that do not allow a direct interpretation of sensitivity and specificity. Below we list the limitations and provide analysis results that address these limitations to the extent possible.

- Imperfect "reference standard": In the analysis of complete set, algorithms are compared to VA graders who themselves are shown to have 82% sensitivity and 84% specificity in the expert arbitration set.
    - Any comparison with human graders can only provide agreement between AI algorithms and the human graders but cannot characterize the true clinical performance (sensitivity and specificity) of the AI systems because the chosen "reference standard" in the complete set may not be accurate.
- To characterize the true performance of the AI systems, the reading center Early Treatment Diabetic Retinopathy Study (ETDRS) reference standard would be needed, which was not available for this study.
    - In a pivotal prospective multi-center clinical trial, EyeArt was compared against the rigorous ETDRS reference standard, providing exceptional performance (summarized below).

| ETDRS validation | EyeArt's More than mild DR result | EyeArt's Vision-threatening DR result |
|---|---|---|
| Sensitivity | 96% | 92% |
| Specificity | 88% | 94% |

- Arbitration set: With an ETDRS clinical reference standard lacking, the best available clinical reference in this study comes from the arbitration set where multiple retina experts have graded a subset of encounters.
    - In the arbitration set (7,379 images from 735 encounters), *Algorithm G*'s performance is perfect (Figure 2 in the paper[1]): The *Algorithm G* achieves 100% sensitivity for moderate NPDR or worse, 100% sensitivity for severe NPDR or worse, and 100% sensitivity for proliferative DR.
- Lumping of "positive" results with "ungradable" results for analyses: In the analyses conducted, such lumping skews the definition of sensitivity (or positive agreement) and specificity (or negative agreement). Hence, the reported sensitivity and specificity numbers cannot be understood well, nor can they be compared with those reported in other studies.
    - When analysis is repeated by not lumping *positive* results with *ungradable* result, ***Algorithm G*'s sensitivity and specificity for detecting mild DR or worse is found to be 99.4% and 85.2% respectively, when using arbitrators' results as the gold standard.** This is an excellent performance in line with results reported in other publications that study EyeArt performance.

| | EyeArt detecting mild DR or worse (N=542) |
|---|---|
| Sensitivity | **99.4%** (95% CI: 96.9%-100%) |
| Specificity | **85.2%** (95% CI: 81.2%-88.6%) |

- It is a retrospective analysis. One implication of this is that it does not allow full alignment with the AI system's imaging protocols.
    - In real world clinical use, FDA-cleared systems such as Eyenuk's are integrated with the camera and the photographers are trained on the imaging protocol to be used.

**Conclusions:** Authors attempt to make the following conclusions from the study in the publication.

- "*The DR screening algorithms showed significant performance differences… Although some algorithms in our study performed well from a screening perspective, others would pose safety concerns.*": We agree that all AI is not created equal. FDA cleared systems have gone through rigorous prospective validation against gold standard (ETDRS) clinical reference standard in intended use settings and are expected to perform well.
- "*These results argue for rigorous testing of all such algorithms on real-world data before clinical implementation.*": Authors are making this general conclusion based on poor performance of some algorithms, but our view is that for systems that have FDA clearance after already going through more rigorous prospective clinical trial validation (than the UW study), additional testing is not necessary.

**References**:

1. Aaron Y. Lee et al., "Multicenter, Head-to-Head, Real-World Validation Study of Seven Automated Artificial Intelligence Diabetic Retinopathy Screening Systems," Diabetes Care, January 4, 2021, https://doi.org/10.2337/dc20-1877.

**Regulatory Note:** The US FDA-cleared version of EyeArt is indicated for use by healthcare providers to automatically detect more than mild diabetic retinopathy and vision-threatening diabetic retinopathy (severe nonproliferative diabetic retinopathy or proliferative diabetic retinopathy and/or diabetic macular edema) in eyes of adults diagnosed with diabetes who have not been previously diagnosed with more than mild diabetic retinopathy. It is indicated for use with Canon CR-2 AF and Canon CR-2 Plus AF cameras in both primary care and eye care settings.

The validation study conducted on the VA dataset evaluated the ability of EyeArt to detect any diabetic retinopathy (mild DR or higher) on images captured using a Topcon TRC-NW8 fundus camera, which is currently only available in EyeArt versions outside the US.