


# A Comparison of Artificial Intelligence and Human Diabetic Retinal Image Interpretation in an Urban Health System

Journal of Diabetes Science and Technology  
2022, Vol. 16(4) 1003–1007  
© 2021 Diabetes Technology Society  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/1932296821999370  
journals.sagepub.com/home/dst  


Nikita Mokhashi, BA<sup>1</sup> , Julia Grachevskaya, OD<sup>1</sup>,  
Lorrie Cheng, OD<sup>1</sup>, Daohai Yu, PhD<sup>1</sup>, Xiaoning Lu, MS<sup>1</sup>,  
Yi Zhang, MD, PhD<sup>1</sup>, and Jeffrey D. Henderer, MD<sup>1</sup>

## Abstract

**Introduction:** Artificial intelligence (AI) diabetic retinopathy (DR) software has the potential to decrease time spent by clinicians on image interpretation and expand the scope of DR screening. We performed a retrospective review to compare EyeNuk's EyeArt software (Woodland Hills, CA) to Temple Ophthalmology optometry grading using the International Classification of Diabetic Retinopathy scale.

**Methods:** Two hundred and sixty consecutive diabetic patients from the Temple Faculty Practice Internal Medicine clinic underwent 2-field retinal imaging. Classifications of the images by the software and optometrist were analyzed using sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and McNemar's test. Ungradable images were analyzed to identify relationships with HbA1c, age, and ethnicity. Disagreements and a sample of 20% of agreements were adjudicated by a retina specialist.

**Results:** On patient level comparison, sensitivity for the software was 100%, while specificity was 77.78%. PPV was 19.15%, and NPV was 100%. The 38 disagreements between software and optometrist occurred when the optometrist classified a patient's images as non-referable while the software classified them as referable. Of these disagreements, a retina specialist agreed with the optometrist 57.9% the time (22/38). Of the agreements, the retina specialist agreed with both the program and the optometrist 96.7% of the time (28/29). There was a significant difference in numbers of ungradable photos in older patients ( $\geq 60$ ) vs younger patients ( $< 60$ ) ( $p=0.003$ ).

**Conclusions:** The AI program showed high sensitivity with acceptable specificity for a screening algorithm. The high NPV indicates that the software is unlikely to miss DR but may refer patients unnecessarily.

## Keywords

artificial intelligence, diabetic retinopathy, ophthalmology, screening, telemedicine

## Introduction

Since 1980, the prevalence of diabetes mellitus has increased globally by 110% in men and 58% in women and reached 9% and 7.9% by 2014. In 2016, global prevalence was estimated at 422 million and is projected to reach 629 million by 2045.<sup>1</sup> Over one-third of diabetics have diabetic retinopathy (DR), which is caused by damage to capillaries due to high glucose levels.<sup>1,2</sup> One-third of patients with DR have vision-threatening DR which is defined as either severe non-proliferative DR or proliferative DR. In 2010, DR was cited as the fifth most common cause of preventable blindness.<sup>2</sup> Early and widespread screening can aid in identifying patients with DR to either treat them or increase hypoglycemic therapy to prevent the progression of severe disease to blindness.

Due to these staggering statistics, widely available screening for DR is required. Traditionally, image interpretation is done by a human, which delays care and adds cost. Artificial intelligence (AI) DR software has the potential to increase the availability and decrease the cost of screening. AI can decrease time spent by clinicians on image interpretation, provide point-of-care results to the patient, and expand the

<sup>1</sup>Department of Ophthalmology, Lewis Katz School of Medicine, Philadelphia, PA, USA

### Corresponding Author:

Nikita Mokhashi, BA, Department of Ophthalmology, Temple Ophthalmology, Lewis Katz School of Medicine, 3401 North Broad Street, Philadelphia, PA 19140, USA.  
Email: tuh38058@temple.edu

scope of diabetic retinopathy screening. In order to create such software, thousands of images are run through a software algorithm to “learn” to identify disease and then the AI is taught to become more accurate and efficient.<sup>3</sup> AI software designed to detect age-related macular degeneration, glaucoma, and retinopathy of prematurity has been shown to have promising results.<sup>4</sup> In the realm of diabetic retinopathy, AI programs utilizing standard multiple-field fundus photography, ultra-wide field photography, and smartphone-based photography have been used to successfully diagnose DR.<sup>5</sup> Gulshan et al. developed a deep learning software using 1 28 175 images from EyePACS to train an algorithm, after which they found an 87%-90% sensitivity and 98% specificity.<sup>6</sup>

Another emerging AI program that has demonstrated success is Eyenuk’s EyeArt software (Woodland Hills, CA). It has been used to rapidly screen color fundus images for referable retinopathy. Bhaskaranand et al. conducted a retrospective study of 1 01 710 consecutive patient visits using it and demonstrated 91.3% sensitivity and 91.1% specificity using ophthalmologists and optometrists as the gold standard.<sup>7</sup>

Based on its observed success, we are beginning to employ the system in our urban academic medical center at Temple University. To evaluate how it compared to the human grader, we determined the sensitivity and specificity of the software when compared to an optometrist in grading diabetic retinopathy using the International Classification of Diabetic Retinopathy grading scheme (ICDR).

## Methods

We obtained IRB approval (protocol #26008) to perform a retrospective chart review of 260 consecutive diabetic patients from the Temple Faculty Practice Internal Medicine clinic that underwent 2-field retinal imaging between April 1, 2019 and August 1, 2019. All patients were assigned a subject identification number on a separate spreadsheet to maintain patient anonymity. Criteria for inclusion were age of 18 years or older at the time of participation and a diagnosis of diabetes mellitus per ICD-10 code. Patients with a history of ocular injections, laser treatment of the retina or other intraocular surgery other than cataract surgery, and history of retinal vascular disease such as arterial or venous occlusions were excluded. Images were taken with a Canon CR-2 AF Non-Mydriatic Digital Retinal camera (24 Megapixel), and photographers were trained Temple University medical assistants.

At least 1 optic nerve centered image and 1 macula centered image from each eye were analyzed by the software and resulted in a reading of non-referable diabetic retinopathy, referable diabetic retinopathy, or ungradable image for each eye. No diabetic retinopathy (ICDR 0) and mild diabetic retinopathy (ICDR 1) were classified as non-referable diabetic retinopathy. Moderate (ICDR 2), severe (ICDR 3) and proliferative (ICDR 4) were classified as referable diabetic retinopathy (Table 1). The artificial intelligence program is designed

**Table 1.** Classification of Diabetic Retinopathy Severity by AI Software.

Severity	ICDR	AI software
No DR	ICDR 0	Non-referable
Mild DR	ICDR 1	Non-referable
Moderate DR	ICDR 2	Referable
Severe DR	ICDR 3	Referable
Proliferative DR	ICDR 4	Referable

to pick up signs of macular edema such as hard exudates when grading an image. The presence of these signs would generate a referable diagnosis. If there are hard exudates, optometrists will similarly generate a referable diagnosis. The Temple Ophthalmology optometrist reading was obtained for each photograph and classified using the identical system. Once the image from each eye was classified as referable, non-referable, or ungradable, the outputs from both eyes were combined into a single patient level output. Patient level outputs were used to compare software and the optometrist readings (Table 2).

Ungradable images typically did not have a view of all 4 quadrants of the posterior pole or were mostly black or blurry. Ungradable images were analyzed using McNemar’s test for any differences between ethnicities and age and binary logistic regression for relationships with HbA1c. Sensitivity, specificity, positive predictive value, negative predictive value, agreement rate, and a McNemar’s test were used for analysis of patient-level and eye-level comparison. All disagreements and a subset of 20% ( $n = 29$ ) of agreements between the optometrist and software were adjudicated by a retina specialist and analyzed to determine which grader was more accurate and the reason for discrepancy. We randomly selected 1 in every 10 patients with agreement until reaching a sample size of 29.  $P$ -values less than .05 were considered statistically significant. SAS version 9.4 (SAS Institute Inc., Cary, NC) was used for all the data analyses.

## Results

Of the 260 patients evaluated, the average age (SD) was 60.7 (11.7) years with a range of 21 to 90 years. 39.2% of the patients were male, and 60.8% were female. 78.8% of the patients self-identified as African American/Black, 14.2% as Hispanic/White, 4.6% as Caucasian, 1.2% as Asian/Pacific Islander, and 1.2% as other. The median (IQR) of the HbA1c was 7.0 (6.3, 8.6) for the 243 patients with HbA1c data.

Thirty-eight patients had images from one or both eyes that were ungradable by both the software and optometrist, 15 patients were ungradable only by the human, and 27 patients were ungradable only by the software (Table 3). Patients with ungradable images had a median age of 66 years, vs 61.4 years for gradable images. Due to the program’s algorithm, any patient that had one eye with an

**Table 2.** Classification System of Image Outputs from Optometrist and AI Software.

Optometrist OS	Optometrist OD	Optometrist patient level
1	1	1
1	2	2
2	1	2
2	2	2
1	3	3
3	1	3
2	3	3
3	2	3

AI Software OS	AI Software OD	AI Software patient level
1	1	1
1	2	2
2	1	2
3	3	3

Optometrist patient level	AI Software patient level	Outcome
1	1	Agree
2	2	Agree
1	2	Disagree
2	1	Disagree
3	3	Ungradable (excluded)
3	1 or 2	Ungradable (excluded)
1 or 2	3	Ungradable (excluded)

1, non-referable; 2, referable; 3, ungradable.

**Table 3.** Breakdown of Ungradable Images by Optometrist and AI Software.

Ungradable results	Number of patients	Percentage of total patients
Ungradable by both	38	14.6
Ungradable only by optometrist	15	5.8
Ungradable only by AI software	27	10.4
Total ungradable	80	30.8

ungradable image automatically resulted in both eyes being reported as ungradable. In contrast, the optometrist classified eyes independently, and so one eye could have been ungradable with the other eye classified as referable or non-referable. We decided to exclude those patients from the analysis. Using McNemar’s test, there was no significant difference in patients classified as ungradable vs gradable when stratified by ethnicity ( $p=0.197$ ). Due to sample sizes less than 5 of certain ethnicities, we grouped ethnicity into African American/Black and other for analysis. Using binary logistic regression, we found that HbA1c was not an accurate predictor of a patient having a gradable vs ungradable photo ( $P = .263$ ). However, we determined that there

**Table 4.** Optometrist and AI Software Patient-Level Comparison of Referable vs Non-Referable Readings.

	Optometrist referable	Optometrist non-referable	
AI software referable	9	38	PPV = 19.15%
AI software non-referable	0	133	NPV = 100%
	Sensitivity = 100%	Specificity = 77.78%	

**Table 5.** Optometrist and AI Software Patient-Level Comparison of Referable vs Non-Referable Readings with Additional 3 Patients Included.

	Optometrist referable	Optometrist non-referable	
AI software referable	12	38	PPV = 23.5%
AI software non-referable	0	133	NPV = 100%
	Sensitivity = 100%	Specificity = 77.78%	

was a difference in numbers of ungradable photos in patients with age  $\geq 60$  vs age  $< 60$  ( $P = .003$ ) using McNemar’s test. When calculating specificity and sensitivity, the total of 80 patients with ungradable images were excluded.

On patient level comparison, sensitivity for the AI software was 100%, while specificity was 77.78% (Table 4). Positive predictive value (PPV) was 19.15%, and negative predictive (NPV) value was 100%. Analysis demonstrated an agreement rate of 78.89% (SE = 3.04, 95% CI: 72.19%-84.61%) and a simple kappa coefficient of 0.259 (SE = 0.072, 95% CI: 0.119-0.340). A chi-square test comparing EyeNuk and the optometrist determination of referable and non-referable demonstrated a  $P < .0001$  with a test statistic  $S = 38.0$ .

There were 3 patients that were excluded who had been classified as referable in one eye and ungradable in the other eye by the optometrist. In clinical practice these 3 patients would have been characterized as referable on a patient level but were classified as ungradable and excluded for the purpose of consistency in this study. These 3 patients, if included, would not have significantly affected the PPV or NPV (Table 5).

Of the 142 patients where the program and optometrist agreed on patient level analysis, 139 of those patients had both individual eyes agree, and 3 patients had only one eye agreeing. Of the 38 patients with disagreement, 24 were due to disagreement between both eyes and 14 were due to disagreement with only one eye.

These 38 disagreements all occurred when the optometrist classified a patient’s images as non-referable while the

**Table 6.** Breakdown of Retina Specialist Adjudication where Optometrist and AI Software Disagreed on Classification.

Adjudication of disagreements by retina specialist	Number of patients	Percentage of disagreements
Agreement with Optometrist	22	57.9
Agreement with AI software	8	21.1
Agreement with neither (ungradable classification)	8	21.1

**Table 7.** Breakdown of Retina Specialist Adjudication where Optometrist and AI Software Agreed on Classification.

Adjudication of 20% of AI software/optometrist agreements by retina specialist	Number of patients	Percentage of agreements
Agreement with Retina Specialist (non-referable classification)	26	89.7
Agreement with Retina Specialist (referable classification)	2	6.9
Disagreement with Retina Specialist	1	3.4

software classified them as referable. Adjudication by a retina specialist demonstrated agreement with the optometrist 22 of the 38 times that the images were non-referable, agreement with the AI program 8 times that the images were referable, and agreement with neither 8 times due to classification as ungradable (Table 6). To better understand if the software and optometrist were using the same criteria, 20% ( $n = 29$ ) of images with agreement between the two were also adjudicated by the retina specialist (Table 7). In 96.5% of cases, the retina specialist agreed with both the optometrist and software. There was 1 case (3.4%) in which the software and optometrist classified a patient as referable while the retina specialist classified that patient as non-referable.

## Discussion

On a patient level, the AI program demonstrated a sensitivity of 100% and specificity of 77.78% in this study. The NPV of 100% and PPV of 19.15% demonstrate it is very unlikely to miss disease, though it demonstrates a high false positive rate when using the optometrist as the gold standard. The high NPV from both patient and eye level comparison shows that a negative result can be used to accurately rule out diabetic retinopathy, but the low PPV shows that a positive result is not sufficient to rule in disease. These numbers are promising for using artificial intelligence as a screening tool, where the goal is to identify disease. A screening tool that does not identify disease is ineffective as a screening tool. Unfortunately, increasing the sensitivity typically comes at the cost of a decreased specificity. In our case, a negative screening result can be presumed negative with a high degree of confidence, but the positive results will need to be verified

by an in-person exam. However, the ability to exclude a large number of eyes when using this software as an initial screening tool allows for significant reduction in the quantity of image analyses for eye-care providers.

Of the 38 patients that were classified as false positives when compared to the optometrist, 8 were found to be legitimate positives by a retina specialist. This suggests that the program can identify referable images that an optometrist may miss. While there were still 22 images in which the retina specialist agreed with the optometrist that an image was a false positive, the use of AI would significantly decrease the number of images for ophthalmologist review without missing disease. As the retina specialist agreed with the optometrist and software in 96.5% of the cases with agreement between the two, we can comfortably conclude that the majority of results with agreement are reliable. In the one case with disagreement, the retina specialist noted that there was drusen bilaterally, which could have contributed to a referable result by the optometrist and AI program vs a non-referable result by the retina specialist.

The low kappa coefficient of 0.259 as well as  $P < .0001$  on the McNemar's test of patient- and eye-level data showed that optometrists and the AI program came to different conclusions about referable DR in a substantial number of cases. Another study of 301 patients by Rajalakshmi et al. determined a 95.8% sensitivity and 80.2% specificity for detecting DR with EyeArt when compared to 2 retina specialists. They suggested that the higher number of false positives was due to detection of non-DR retinal lesions like drusen, RPE atrophic patch, retinal telangiectatic vessels at the macula, RPE hypertrophy, tessellated fundus, and retinal vein occlusion.<sup>8</sup> In a study of 69 patients using a smartphone-based camera (RetinaScope), Kim et al. found a lower sensitivity of 77.8% and specificity of 71.5%. It is possible that the sensitivity and specificity were lower in Kim et al.'s study due to the smaller sample size or due to inferior image quality of a phone photograph vs a table camera photograph.<sup>9</sup> On our analysis by a retina specialist, many of the false positive calls occurred when there were laser scars, drusen, or artifacts in the image.

There are several limitations to this study. Each image was graded by only 1 optometrist, which could allow for grading bias between the 2 optometrists involved. However, it was more practical in the context of a real-life screening program where timely feedback to the primary care provider is necessary. Both optometrists were equivalently trained and frequently consulted each other when grading images. High overall agreement (96.5%) with the retina specialist (Table 7) also indicates that the optometrists had similar grading proficiency.

Additionally, there was a high ungradable photo rate, and issues affecting photo quality will need to be explored. The AI program does not independently evaluate images when one eye has an ungradable image, as an ungradable image from one eye results in an ungradable result from both eyes. This system property complicated the comparison with the optometrist thus necessitating exclusion of those patients.



We also noted that older patients ( $\geq 60$  years old) were more likely to have an ungradable photo. When grading nonmydriatic photographs in individuals with diabetes, Scanlon et al. similarly found a 5.8% increased likelihood of having an ungradable image for every 1 year of increased age.<sup>10</sup> Higgs et al. also reported that 54% of nonmydriatic images were ungradable in patients greater than 70 years old vs 13% in patients less than 50 years old.<sup>11</sup> It is possible that older patients were more likely to have conditions such as cataract, leading to lower quality photos. This difference in image quality between age groups suggests that special care must be taken to procure high quality photographs in older individuals to accurately assess progression of diabetic retinopathy. In the future, reducing the ungradable photo rate will help to improve the accuracy of the screening program. Further photographer training or dilation of patients may also be required.

## Conclusions

The EyeArt software showed a very high sensitivity with an acceptable, relatively lower level of specificity for a screening algorithm. The high negative predictive value indicates that it is unlikely to miss DR. However, in this patient population and in this study, the high number of ungradable images and the low positive predictive value mean that some patients will be referred unnecessarily. We believe this an encouraging start to our screening program at Temple University.

In the future, we plan to improve image quality in real time using AI software to provide immediate photo quality feedback for the photo-taker. Other avenues may include additional technician training or dilating patients.

## Abbreviations

AI, artificial intelligence; DR, diabetic retinopathy; ICDR, International Classification of Diabetic Retinopathy grading scheme; NPV, negative predictive value; PPV, positive predictive value.

## ORCID iD

Nikita Mokhashi  <https://orcid.org/0000-0001-9510-3747>

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: We were given free access to the AI software by EyeNuk, Inc. (Woodland Hills, CA) in order to perform the study. EyeNuk, Inc was not involved in the structure or outcomes of the study in any way.

## References

1. Cheloni R, Gandolfi SA, Signorelli C, et al. Global prevalence of diabetic retinopathy: protocol for a systematic review and meta-analysis. *BMJ Open*. 2019;9(3).
2. Lee R, Wong TY, Sabanayagam C. Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. *Eye and Vis*. 2015;2(1):17.
3. Padhy S, Takkar B, Chawla R, et al. Artificial intelligence in diabetic retinopathy: a natural step to the future. *Indian J Ophthalmol*. 2019;67(7):1004-1009.
4. Bellemo V, Lim G, Rim TH, et al. Artificial intelligence screening for diabetic retinopathy: the real-world emerging application. *Curr Diab Rep*. 2019;19(9):1-12.
5. Lim G, Bellemo V, Xie Y, et al. Different fundus imaging modalities and technical factors in AI screening for diabetic retinopathy: a review. *Eye and Vis*. 2020;7(1):1-13.
6. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402-2410.
7. Bhaskaranand M, Ramachandra C, Bhat S, et al. The value of automated diabetic retinopathy screening with the eyart system: a study of more than 100,000 consecutive encounters from people with diabetes. *Diabetes Technol Ther*. 2019;21(11):635-643.
8. Rajalakshmi R, Subashini R, Anjana RM, et al. Automated diabetic retinopathy detection in smartphone-based fundus photography using artificial intelligence. *Eye*. 2018;32(6):1138-1144.
9. Kim TN, Aaberg MT, Li P, et al. Comparison of automated and expert human grading of diabetic retinopathy using smartphone-based retinal photography. *Eye*. 2021;35(1):334-342.
10. Scanlon PH, Foy C, Malhotra R, Aldington SJ. The influence of age, duration of diabetes, cataract, and pupil size on image quality in digital photographic retinal screening. *Diabetes Care*. 2005;28(10):2448-2453.
11. Higgs ER, Harney BA, Kelleher A, Reckless JP. Detection of diabetic retinopathy in the community using a non-mydriatic camera. *Diabet Med*. 1991;8(6):551-555.