Validity and reproducibility of ophthalmologist photo grading of diabetic retinopathy and glaucoma

Screening for diabetic retinopathy (DR) is an important component of diabetes management to prevent vision loss, yet many diabetic patients do not receive screening eye examinations.[1] Tele-ophthalmology programs administering fundus photographs offer a potential solution, and technological advances have made automated photo-interpretation a real possibility.[2] Automated algorithms are typically trained on images whose disease status has been classified by a human, so the validity of human grading is important. And yet, although the variability of ophthalmologist grading for glaucoma has been described previously, limited information exists for DR.[2–4] This study attempts to fill that gap in order to better inform optimal methods of human grading for use in automated algorithms.

In this cross-sectional diagnostic accuracy study, a convenience sample of diabetic patients were recruited from the ophthalmology clinic at Chiang Mai University, Thailand. After pupillary dilation, the fundus of each eye was examined by an ophthalmologist, who noted the presence of DR according to the International Clinical Diabetic Retinopathy Severity Scale as well as the vertical cup-to-disk ratio (VCDR).[5] Each eye was then photographed (TRC-NW6S, Topcon, Tokyo, Japan; manufacturer's software used to create an 85° montage image from photographs captured at 9 fixed locations using a repositionable internal fixation light). The mosaic image was graded for the presence of DR according to the International Clinical Diabetic Retinopathy Severity Scale and for the VCDR by 5 ophthalmologists who were masked to each other's grades and to all participant identifiers. Although mosaic images are not considered the gold standard for DR or glaucoma screening, they have been shown to have sensitivities equivalent to in-person ophthalmology examination.[6] Glaucoma suspect was defined as a VCDR ≥0.6. Cohen's kappa statistic was used to assess agreement between photo grades and clinical examination and also agreement between the 5 photo graders, using bootstrapped 95% confidence intervals (CIs) resampled at the participant level to account for correlation of eyes from the same person (N = 999 replications). Ethics approval was obtained from University of California, San Francisco, and Chiang Mai University.

In total, 235 eyes from 119 participants completed an ophthalmologist examination and fundus photography. Of these, 222 photo montages were judged to have adequate image clarity and coverage of the optic disk and macula by a majority of photo graders and were included in the analysis. On the reference standard ophthalmologist examination, 19 of 222 eyes were assessed as having a VCDR ≥0.6, and 97 eyes were diagnosed with DR—43 (19%) with nonproliferative DR and 54 (24%) with proliferative DR. The validity of photo grading is summarized in Table 1, which shows that agreement between each photo grader and the reference standard was greater for DR than VCDR, and that creating a majority consensus grade improved the agreement much more for VCDR than for DR assessment. The precision of photo grading is summarized in Table 2, which shows that inter-rater reproducibility between the 5 photo graders was significantly higher for DR (κ = 0.75, 95% CI 0.70–0.82) than glaucoma (κ = 0.40, 95% CI 0.28–0.52) (difference 0.35 higher for DR, 95% CI 0.21–0.50).

In this study, inter-rater reproducibility was higher for DR assessment compared with VCDR assessment, and photo grades for DR were a more valid indicator of disease (i.e., better agreement with reference standard). The study was conducted using a single digital camera, and so it is unclear whether these results are generalizable to lower-resolution

Table 1—Assessment of vertical cup-to-disk ratio and diabetic retinopathy by photo grading and agreement with in-person eye examination by an ophthalmologist

	Finding on photography			
Photo grader	Present	Absent	Cannot determine	Agreement with eye exam, Cohen's κ (95% CI)†
Any DR				
1	92	124	6	0.91 (0.86–0.98)
2	62	159	1	0.70 (0.58–0.81)
3	100	122	0	0.94 (0.89–0.98)
4	102	120	0	0.86 (0.80–0.93)
5	97	123	2	0.87 (0.81–0.94)
Consensus*	93	129	0	0.89 (0.81–0.95)
VCDR ≥0.6				
1	23	198	1	0.56 (0.34–0.74)
2	11	208	3	0.38 (0.15–0.58)
3	19	198	5	0.47 (0.24–0.66)
4	29	193	0	0.58 (0.34–0.77)
5	14	189	19	0.35 (0.15–0.53)
Consensus*	18	204	0	0.73 (0.52–0.90)

CI, confidence interval; DR, diabetic retinopathy; VCDR, vertical cup-to-disk ratio.
*Results are also shown for the consensus grade, classified as that upon which at least 3 of the 5 graders agreed.
†From 3 × 3 contingency table (present, absent, cannot determine); 95% bootstrap CIs were resampled at the person level to account for correlation of eyes from the same person.

Table 2—Inter-rater agreement between 5 photo graders for retinopathy findings and cup-to-disk ratio thresholds

Classification	Number*	Cohen's κ (95% CI)†
Any retinopathy feature	93	0.75 (0.69–0.81)
Microaneurysms	80	0.65 (0.59–0.71)
Cotton-wool Spots	13	0.46 (0.29–0.59)
Intraretinal hemorrhage	83	0.44 (0.33–0.56)
Hard exudates	41	0.57 (0.46–0.68)
Neovascularization of the disk	7	0.43 (0.24–0.60)
Neovascularization elsewhere	8	0.42 (0.29–0.52)
Fibrous proliferation of the disk	15	0.73 (0.57–0.85)
Fibrous proliferation elsewhere	18	0.63 (0.46–0.75)
Preretinal hemorrhage	6	0.59 (0.26–0.76)
Vitreous hemorrhage	3	0.24 (0.10–0.35)
Glaucoma		
VCDR ≥0.6	18	0.40 (0.27–0.52)
VCDR ≥0.7	4	0.21 (0.07–0.32)

CI, confidence interval; VCDR, vertical cup-to-disk ratio.
*Number of photographs for which a majority of graders judged the finding to be present.
†From 3 × 3 contingency table (present, absent, cannot determine); 95% bootstrap CIs were resampled at the person level to account for correlation of eyes from the same person.

https://doi.org/10.1016/j.jcjo.2019.11.006
ISSN 0008-4182

imaging systems such as the ultra-widefield scanning ophthalmoscope. Nonetheless, these results suggest that automated algorithms for DR could be trained on fundus photographs graded by even a single person, whereas algorithms for VCDR would benefit from grading based on a consensus of multiple graders.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.jcjo.2019.11.006.

Blake M. Snyder,[*,†,1] Irene Shyu,[*,‡,1] Arezu Haghighi,[*] Voraporn Chaikitmongkol,[§] Janejit Choovuthayakorn,[§] Linda Hansapinyo,[§] Thidarat Leeungurasatien,[§] Nawat Watanachai,[§] Sakarin Ausayakhun,[§] Somsanguan Ausayakhun,[§] Jeremy D. Keenan[*]

[*]University of California, San Francisco, CA; [†]University of Colorado School of Medicine, Aurora, CO; [‡]Vanderbilt University School of Medicine, Nashville, TN; [§]Chiang Mai University, Chiang Mai, Thailand

Originally received Sep. 10, 2019. Final revision Oct. 22, 2019. Accepted Nov. 3, 2019.

Correspondence to:
Jeremy D. Keenan, MD, MPH; jeremy.keenan@ucsf.edu.

References

1. Piyasena M, Murthy GVS, Yip JLY, et al. Systematic review on barriers and enablers for access to diabetic retinopathy screening services in different income settings. PLoS One 2019;14:e0198979.
2. Gupta V, Bansal R, Gupta A, Bhansali A. Sensitivity and specificity of nonmydriatic digital imaging in screening diabetic retinopathy in Indian eyes. Indian J Ophthalmol 2014;62:851–6.
3. Lichter PR. Variability of expert observers in evaluating the optic disc. Trans Am Ophthalmol Soc 1976;74:532–72.
4. Olson JA, Strachan FM, Hipwell JH, et al. A comparative evaluation of digital imaging, retinal photography and optometrist examination in screening for diabetic retinopathy. Diabet Med 2003;20:528–34.
5. Wilkinson CP, Ferris 3rd FL, Klein RE, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology 2003;110:1677–82.
6. Shiba T, Yamamoto T, Seki U, et al. Screening and follow-up of diabetic retinopathy using a new mosaic 9-field fundus photography system. Diabetes Res Clin Pract 2002;55:49–59.

Footnotes and Disclosure

The authors have no proprietary or commercial interest in any materials discussed in this article.

This study was supported by the JaMel Perkins Family Foundation, the Fortisure Foundation, That Man May See, Research to Prevent Blindness, the Littlefield Trust, the Peierls Foundation, the Doris Duke Charitable Foundation, and the University of California, Berkeley, Blum Center for Developing Economies. Shyu was supported by an award from the Office of Medical Student Research at Vanderbilt University School of Medicine. Yen and Snyder were Doris Duke International Clinical Research Fellows.

[1] Blake M. Snyder and Irene Shyu contributed equally to this work.