AMERICAN ACADEMY OF OPHTHALMOLOGY®

Correspondence

Re: Angela S Li et al Gradeability and Reproducibility of Geographic Atrophy Measurement in GATHER-1, a Phase II/III Randomized Interventional Trial

TO THE EDITOR: We were interested to read the paper by Li et al. [1] The authors aimed to validate GATHER-1 inclusion criteria and the study's primary anatomic end point by assessing the reproducibility of geographic atrophy (GA) measurements and factors that affect reproducibility. For 286 participants, blue-light fundus autofluorescence, color fundus photographs, fluorescein angiograms, and OCT scans were obtained on the study eye and fellow eye. The main outcome measures were gradeability and reproducibility of fundus autofluorescence imaging data. Imaging data included lesion area, confluence of GA with peripapillary atrophy (PPA), whether GA involved the foveal centerpoint, and type of hyperautofluorescence pattern. They reported that gradeability (90.5%) and interreader gradeability concordance (90.2%) were high across all visits. Moreover, there was no difference in gradeability, gradeability concordance, or lesion-area concordance for images with PPA-confluent GA compared with those with nonconfluent PPA.

The article provides insight into the decision that there is high gradeability and excellent reproducibility measures across all images. However, their approach is limited in 2 important methodological and statistical ways.[2–5]

First, it is possible to have the prevalence of concordance equal to 90% and disconcordance to 10% while getting different kappa values (0.11 as poor vs. 0.81 as very good, respectively).[2,3] However, it does mean that focusing on concordance is not an appropriate approach to assessing either gradeability or reproducibility (Table 1). Prevalence-adjusted kappa is suggested for correctly assessing gradeability or reproducibility for qualitative variables. Prevalence-adjusted kappa is calculated on a frequency table with a standardized marginal and will make similarly standardized kappas comparable. In general, the low prevalence level results in a substantial reduction in kappa values, which can be misleading.

Table 1. Limitation of Concordance Cells and Simple Kappa Value to Assess Gradeability and Reproducibility

	Reader 1		
Reader 2	+	−	Total (%)
Situation (a)			
+	**89**	5	94
−	5	**1**	6
Kappa = 0.11 (poor)			
Total	94	6	100
Situation (b)			
+	**45**	5	50
−	5	**45**	50
Kappa = 0.81 (very good)			
Total	50	50	100

Bold indicates frequency of concordance cells.

Second, our approach to assess gradeability and reproducibility should be individually based rather than a global average approach. The authors correctly applied the Bland Altman plot for all images (A) and screening-visit images (B) and showed high interreader reproducibility across all image sizes. However, for all gradeability analyses, interreader gradeability concordance, and interreader lesion-area concordance, they applied either a chi-square test or a McNemar test (for paired data) to compare the 2 proportions. The authors utilized the nonparametric Wilcoxon rank sum test or Wilcoxon signed-rank test (for paired data) to compare the 2 distributions in terms of quantitative data, such as lesion area or interreader absolute difference in mm^2. The global average used for these test statistics can easily result in a misleading message. We suggest applying a prevalence-adjusted kappa value and intraclass correlation coefficient for qualitative and quantitative variable, respectively.[4,5]

SIAMAK SABOUR, MD, PHD
FARIBA GHASSEMI, MD

Safety Promotion and Injury Prevention Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran; Department of Clinical Epidemiology, School of Public Health and Safety, Shahid Beheshti University of Medical Sciences, Tehran, Iran; Eye Research Center, Farabi Eye Hospital, Tehran University of Medical Sciences, Tehran, Iran; Retina and Vitreous Service, Farabi Eye Hospital, Tehran University of Medical Sciences, Tehran, Iran

Disclosure(s):
All authors have completed and submitted the ICMJE disclosures form. The authors have no proprietary or commercial interest in any materials discussed in this article.
Author Contributions:
Conception and design: Sabour, Ghassemi
Data collection: N/A
Analysis and interpretation: Sabour, Ghassemi
Obtained funding: N/A
Overall responsibility: Sabour, Ghassemi

Available online: .

Correspondence:
Siamak Sabour, MD, PhD, Chamran Highway, Department of Clinical Epidemiology, School of Public Health and Safety, Shahid Beheshti University of Medical Sciences, Velenjak, Daneshjoo Blvd, Tehran, Iran. E-mail: s.sabour@sbmu.ac.ir.

References

1. Li AS, Myers J, Stinnett SS, et al. Gradeability and reproducibility of geographic atrophy measurement in GATHER-1, a phase II/III randomized interventional trial. *Ophthalmol Sci*. 2023;4:100383.
2. Sabour S. Reliability of a new modified tear breakup time method: methodological and statistical issues. *Graefes Arch Clin Exp Ophthalmol*. 2016;254:595–596.
3. Sabour S, Ghassemi F. Accuracy, validity, and reliability of the infrared optical head tracker (IOHT). *Invest Ophthalmol Vis Sci*. 2012;53:4776.
4. Sabour S, Ghassemi F. Accuracy and reproducibility of the ETDRS visual acuity chart: methodological issues. *Graefes Arch Clin Exp Ophthalmol*. 2016;254:2073–2074.
5. Sabour S. Reproducibility of dynamic Scheimpflug-based pneumotonometer and its correlation with a dynamic bidirectional pneumotonometry device: methodological issues. *Cornea*. 2015;34:e14–e15.