



Inter-rater reliability for diagnosis of geographic atrophy using spectral domain OCT in age-related macular degeneration

Shruti Chandra^{1,2} · Rajna Rasheed¹ · Piyali Sen^{1,2} · Deepthy Menon¹ · Sobha Sivaprasad^{1,2}

Received: 4 January 2021 / Revised: 5 February 2021 / Accepted: 22 February 2021 / Published online: 8 March 2021
© The Author(s), under exclusive licence to The Royal College of Ophthalmologists 2021

Abstract

Purpose To evaluate the inter-rater reliability for identification of complete retinal pigment epithelium and outer retinal atrophy (cRORA) on SD-OCT images as defined by the Classification of Atrophy Meetings (CAM) group.

Methods Fifty images of anonymized SD-OCT line scans of eyes with cRORA due to AMD were selected. Each .tiff image was saved in both black-on-white (BW) and white-on-black (WB) format. Five retina-trained clinicians graded both sets of images twice for the diagnosis of cRORA based on the CAM group definition. Fleiss kappa statistic was calculated for inter-rater reliability and Cohen's kappa statistic for intra-grader and inter-grader reliability between any two graders.

Results The inter-grader reliability varied from as low as 0.28 to 0.92 for WB images and 0.34 to 0.86 for BW images. However, the inter-grader and intra-grader agreement was κ WB 0.92; κ BW 0.86 and κ 0.92 respectively, for graders accustomed to the CAM criteria. Fleiss kappa was κ 0.49 (p value < 0.0001) for WB images and κ 0.34 (p value < 0.0001) for BW images. Overall, the agreement was better using WB images for all parameters except RPE attenuation/loss.

Conclusion There is significant variability in diagnosis of cRORA on SD-OCT by retina-trained ophthalmologists in the real world. The study highlights the need for training to recognise the different features of cRORA prior to its implementation in clinical practice.

Introduction

One of the phenotypes of late stages of age-related macular degeneration (AMD) is geographic atrophy (GA). Since its first description in 1970 by Gass, various terminologies and classifications have been used to define this entity on colour fundus photographs [1–4]. Typically, GA is defined as any sharply delineated roughly round or oval area of hypopigmentation or depigmentation with increased visibility of the underlying choroidal vessels of at least 175 μ m in diameter on 30° or 35° colour fundus photographs (CFP) [5]. However, discriminating the edges of atrophy is a difficult task on monoscopic images and cannot be easily delineated

using image analysis. Additionally, precursor changes at the cellular level that precede the development of GA are not identifiable on CFP.

The advent of fundus autofluorescence (FAF) has advanced our understanding significantly. Not only did it better delineate lesion boundaries, but several phenotypes of GA also became evident with associated prognostic significance [6, 7]. In addition, hyperautofluorescence at the margin of GA may precede cell death and growth of GA. For the first time, regulators have accepted FAF as a structural surrogate of disease progression [8]. However, the FAF lesion phenotypes are not always reproducible [9]. FAF is also dependent on lipofuscin loss from retinal pigment epithelium (RPE). Therefore, the focus is diverted to other imaging modalities that can identify changes before RPE cell loss and can complement CFP and FAF.

Spectral domain optical coherence tomography (OCT) has become the mainstay imaging modality for macular diseases including AMD. The high axial resolution of OCT allows layer-by-layer evaluation of retinal and choroidal tissue, enabling cross-sectional phenotyping of GA. More importantly, longitudinal OCT scans can also reveal temporal changes and identify precursor lesions of GA.

✉ Sobha Sivaprasad
sobha.sivaprasad@nhs.net

¹ National Institute of Health Research Moorfields Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust, London, UK

² University College London, Institute of Ophthalmology, London, UK

In the light of these advantages, the Classification of Atrophy Meetings (CAM) group proposed a new classification system based on OCT, unique to atrophy associated with AMD [10]. The new classification system considered microstructural changes in the outer retina and RPE to define four types of atrophy based on the correlated histopathological changes in the retina. These include iORA (incomplete outer retinal atrophy), cORA (complete outer retinal atrophy), iRORA (incomplete RPE and outer retinal atrophy) and cRORA (complete RPE and outer retinal atrophy) [8].

The entity, cRORA, the equivalent of GA, has been defined by the CAM group as having the following features on OCT: (i) Zone of hypertransmission of $\geq 250 \mu\text{m}$, (ii) Zone of attenuation or disruption of RPE band of $\geq 250 \mu\text{m}$, (iii) Evidence of overlying photoreceptor degeneration characterised by features that include outer nuclear layer (ONL) thinning, external limiting membrane (ELM) loss, and ellipsoid zone (EZ) or interdigitation zone (IZ) loss. Although these features are well-described, multiple novel parameters had to be evaluated by a team of retinal experts around the world and several meetings and grading exercises had to be completed to reach a consensus on the components of cRORA. In the absence of reliable quantification methods, the interpretations of these parameters are based on subjective recognition of descriptive characteristics. Due to the diversity of features seen in individual OCT images, image artefacts and variations in image quality, it can be challenging to accurately ascertain presence of these features. The agreement between clinicians provides a measure of this challenge.

The purpose of this study was to evaluate the inter-rater reliability for identification of cRORA on Spectralis Heidelberg SD-OCT line scans (Heidelberg Engineering, Heidelberg, Germany). We then investigated the technical issues faced by the graders in evaluation of each individual parameter to recommend approaches to mitigate them.

Methods

The study adhered to the tenets of the Declaration of Helsinki. Only anonymised images were analysed so institutional review board approval was not required.

Image acquisition and processing

One of the authors (SC) selected and extracted Spectralis Heidelberg SD-OCT line scans (Heidelberg Engineering, Heidelberg, Germany) that were routinely done on patients with GA. Scans were either on iRORA and cRORA. The line scan had to demonstrate an atrophic lesion size $\geq 250 \mu\text{m}$ measured on the Heyex overlay software. To

assure consistent image quality and similar preconditions for each evaluated scan, we used well-resolved SD-OCT scans. These were pre-defined as line scans with a minimum of 20 dB signal-to-noise ratio; minimum of sixteen frames per B-scan using the average real time mode; clear media and qualitatively confirmed visibility and distinction of each outer retinal layer.

Each of the 50-line scans of the 50 patients were anonymised and exported from the Heidelberg software in .tiff format in the standard setting. Each line scan was saved in two different colour settings (i.e. inverted grey-scale showing images as either positives or negatives-referencing terminology used in black and white photography) identified as white-on-black (WB) and black-on-white (BW) in the study. Each folder contained a WB and BW scan of an anonymised patient.

Grading characteristics

The two sets of the images (WB and BW) of each of the 50 patients were analysed for three parameters mentioned in the CAM grading for cRORA. They include: (i) Zone of hypertransmission of $\geq 250 \mu\text{m}$, (ii) Zone of attenuation or disruption of RPE band of $\geq 250 \mu\text{m}$, (iii) Evidence of overlying photoreceptor degeneration whose features include ONL thinning, ELM loss and EZ or IZ loss.

Graders

The 50 pairs of images were interpreted by five clinicians (four medical retina fellows and one consultant) who are skilled OCT readers from the same retina centre (Moorfields Eye Hospital, London, UK). Two of the five graders were more accustomed to the CAM grading. All graders were familiarised with the CAM grading and trained to identify the parameters on test sets before grading the study images independently.

Grading procedure

The five graders analysed the 50 pairs of SD-OCT horizontal cross-sectional scans without application of any image modifications. The images were presented in two separate data sets of 50 images each, and the readers were masked to the grading outcome. The readers were required to go through both the images and data sets in the same order. The graders were asked to document their response as yes or no for presence or absence cRORA and each parameter that defines cRORA. The folders were then randomly re-numbered, and graders re-graded the images to evaluate intra-grader agreement. Additionally, they were asked to make a notation of whether WB or BW scans was most helpful in identifying each particular feature for a set of

Table 1 Interpretation of Cohen's and Fleiss kappa.

Cohen's kappa value	Interpretation	Fleiss kappa value	Interpretation
0–0.20	None	<0.00	Poor agreement
0.21–39	Minimal	0.00–0.20	Slight agreement
0.40–0.59	Weak	0.21–0.40	Fair agreement
0.60–0.79	Moderate	0.41–0.60	Moderate agreement
0.80–0.90	Strong	0.61–0.80	Substantial agreement
>0.90	Almost perfect	0.81–1.00	Almost perfect

Table 2 Inter-grader and Intra-grader agreement (Cohen's kappa) for cRORA in white on black images.

Grader	#1	#2	#3	#4	#5
#1	0.88	0.64	0.44	0.28	0.30
#2	0.64	0.94	0.39	0.33	0.34
#3	0.44	0.39	0.94	0.65	0.68
#4	0.28	0.33	0.65	0.95	0.92
#5	0.30	0.34	0.68	0.92	0.92

images. They could also choose that there was no difference in terms of setting, in their ability to evaluate respective OCT features.

Statistical analysis

Statistical analysis was performed using SPSS Statistics version 24 (IBM), Microsoft Excel for Mac version 15.33 (Microsoft), and the web-based Kappa Programme [11]. Inter-grader agreement was evaluated as a measure of reliability. The higher the inter-grader correlation coefficient, the more reliable was the identification and detectability of respective morphologic alteration. The responses being categorical variables, Fleiss' kappa (κ) was used for this purpose. To measure the intra-grader agreement and inter-grader agreement among two graders Cohen's kappa was used (Table 1). The significance was set at $p \leq 0.05$.

Results

Fifty pairs of single SD-OCT horizontal scans were graded twice by each grader. Of the 50 images, 36 images demonstrated cRORA and the rest had iRORA as confirmed by two graders with previous experience in CAM grading.

The inter-grader and intra-grader Cohen's kappa values for cRORA diagnosis are shown for white-on-black (Table 2) and black-on-white images (Table 3). The intra-grader reliability by Cohen's kappa was in the range of

Table 3 Inter-grader and Intra-grader agreement (Cohen's kappa) for cRORA in black on white images.

Grader	#1	#2	#3	#4	#5
#1	0.93	0.49	0.48	0.37	0.50
#2	0.49	0.90	0.55	0.39	0.34
#3	0.48	0.55	0.95	0.57	0.58
#4	0.37	0.39	0.57	0.93	0.86
#5	0.38	0.34	0.58	0.86	0.45

0.88–0.92 for white-on-black images which is strong to almost perfect agreement. The Cohen's kappa for black-on-white images ranged from 0.45 to 0.95, being >0.90 for four out of five graders. The inter-grader reliability varied from as low as 0.28 to almost perfect value of 0.92 for white-on-black images. Similarly, it ranged from 0.34 to 0.86 for black-on-white images. The inter-grader agreement was almost perfect for two graders (kappa WB 0.92, p value < 0.0001; kappa BW 0.86, p value < 0.0001) who were accustomed to the CAM criteria.

The Fleiss kappa values (κ) are shown in Table 4. There was moderate agreement in identifying cRORA using white-on-black images (κ 0.49, p value < 0.0001) and fair agreement using BW images (κ 0.34, p value < 0.0001). The RPE attenuation/loss was parameter detected most reliably in both sets of images whereas hypertransmission was the most poorly detected parameter. Overall, the agreement was better using WB images for all parameters except RPE attenuation/loss.

The graders noted that RPE attenuation/loss was the relatively easier parameter to identify in the images, better detected on BW images. Hypertransmission was the least reliable parameter according to the graders and was particularly difficult to distinguish on BW images. Inner layer changes also were more clearly identified using WB versus BW images. However, the graders observed that to reliably detect the presence of cRORA it is better to analyse both images together. Examples of challenges in grading cRORA are shown in Fig. 1.

Discussion

The study assessed the reliability of detection of cRORA on SD-OCT images. There are four key findings. First, inter-grader reliability for any two graders was better for WB images than BW images and was almost perfect for graders more accustomed with CAM classification. Second, intra-grader agreement was high across all images suggesting parameters assessed inaccurately across one set of images were assessed incorrectly across the second set, reinforcing the importance of repeated training. Third, inter-grader

Table 4 Fleiss kappa showing Inter-grader agreement across five graders for all parameters.

Parameter	Fleiss kappa (95% CI)	Standard error	Level of agreement
White on black images			
1. cRORA overall	0.49 (0.22–0.45)	0.06	Moderate
2. Hypertransmission	0.36 (0.20–0.52)	0.08	Fair
3. Inner retinal changes	0.46 (0.30–0.62)	0.08	Moderate
4. RPE attenuation/loss	0.68 (0.52–0.84)	0.08	Substantial
Black on White Images			
1. cRORA overall	0.34 (0.33–0.65)	0.08	Fair
2. Hypertransmission	0.34 (0.22–0.48)	0.06	Fair
3. Inner layer changes	0.38 (0.30–0.56)	0.40	Fair
4. RPE attenuation/loss	0.72 (0.65–0.79)	0.60	Substantial

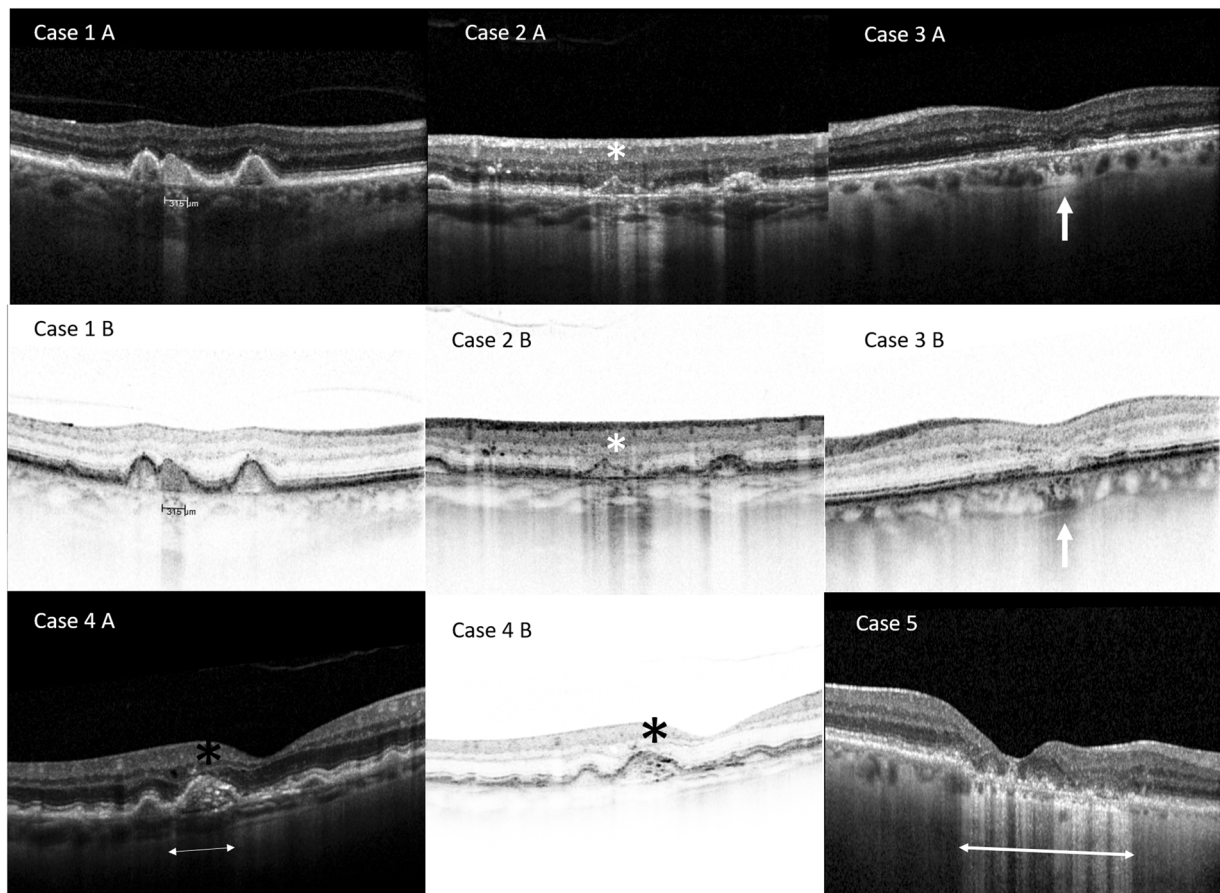


Fig. 1 Examples of optical coherence tomography (OCT) images analysed in the study. Case 1A: White on black (WB) image showing hypertransmission of 315 μm width and associated inner retinal changes overlying the PED. In the corresponding Black on White (BW) image (Case 1B), the RPE loss is better appreciated. Case 2A and 2B: This image shows the presence of a persistent hyper-reflective line in the bed of cRORA, which could be confused as attenuated RPE. This has been termed persistent basal laminal deposit (*white asterisk*) by the CAM classification. Case 3A: All signs of cRORA are noticeable (EZ/ELM layer changes and hypertransmission of 250 μm),

however the RPE is intact (*white vertical arrow*). This is again more evident in the BW image (*white vertical arrow*; Case 3B). Case 4A: There is loss of RPE and EZ and ELM layer changes overlying the PED (*black asterisk*), but the hypertransmission is absent (*white horizontal arrow*). Case 4B: The BW image confirms the definite absence of RPE (*black asterisk*). Case 5: This case shows an example of discontinuous transmission overlying a region of cRORA caused due to back shadowing secondary to dispersed pigmented cells (*white horizontal arrow*).

reliability using Fleiss kappa was fair to moderate indicating the subjectivity of defined parameters. Finally, all parameters were detected better on WB images except RPE

attenuation. RPE attenuation/loss was the parameter detected with highest agreement whereas hypertransmission had the lowest agreement. Taken together these findings suggest

that diagnosis of cRORA on OCT images may be quite subjective and have an impact on clinical diagnosis when treatment become available for this condition. Adequate training combined with use of both WB and BW images may enhance our capability to improve our diagnostic capability of this entity.

The first finding was that inter-grader reliability was better for WB images. There are a couple of possible explanations for this finding. First, all graders were accustomed to viewing WB images routinely in clinic and this may have introduced a bias to better identification of structures in this set of images. Second, recognising minute aspects of outer retinal changes require advanced skills possibly acquired by repeated evaluation of these images. The RPE, EZ and ELM on OCT are seen as multiple hyper-reflective lines with almost similar reflectivity. The presence of drusen, subretinal drusenoid deposits, patchy intraretinal pigment migration or presence of outer retinal tubulations are some of the features that distort the continuity of these layers and increase the difficulty in differentiating them. Repetitive exposure to scans with these characteristics may enhance the expertise in accurately identifying pathologies.

Secondly, intra-grader agreement was high across all images. A layer or parameter identified wrongly on repeated occasions indicate the challenge in identifying these parameters due to the heterogeneity of these parameters and highlight the importance of experience required to assess these entities.

The third finding was the fair to moderate agreement based on Fleiss kappa for cRORA diagnosis. Kappa was designed to consider the possibility of guessing but it has its limitations. The assumptions kappa makes about rater independence and other factors are not well supported, leading to an excessively low estimate of agreement [12]. As it cannot be directly interpreted, it has become conventional for researchers to accept low kappa levels in inter-rater reliability studies [10]. Table 1 shows us the interpretation of Fleiss kappa coefficient. However, this interpretation is applicable to social science. Low levels of kappa are unacceptable in medicine or clinical research where results may change clinical practice and may lead to poorer clinical outcomes. An agreement of over 0.80 is considered acceptable in medicine related research [12]. Thus, even though our results show moderate agreement, they are not adequate. This points towards use of more descriptive definitions of each parameter or employing reliable quantification methods for more subjective parameters such as hypertransmission. These steps may help improve inter-grader agreement.

Finally, two parameters (hypertransmission and inner retinal layer changes) were detected with higher reliability on WB images and RPE attenuation was more reliably graded on BW images. Use of a combination of WB and BW images is required to better detect the parameters to reliably diagnose

cRORA. Even though it was not evaluated in this particular study, the authors suggest use of other modalities including near-infrared reflectance (NIR) and FAF alongside OCT B-scan to improve cRORA diagnosis. We did not employ use of NIR and FAF in this study to avoid bias in detecting cRORA as this grading was based completely on OCT features. Moreover, the authors evaluated .tiff files of these images to maintain uniformity across the grading. In the real world, dynamic manipulation of images with ability to adjust colour/contrast settings and use multimodal imaging, may lead to precise identification of morphologic features on OCT. However, most reliability studies are done on static images due to ease of implementation [13].

The study scrutinized the difficulties faced by the graders during the grading process with respect to each parameter. The most challenging parameter was continuous hypertransmission. The CAM group used the term hypertransmission as it best conveyed the cause for the observed phenomenon. It is recognised though that hypertransmission may not always penetrate to the underlying choroid especially in eyes with tall pigment epithelial detachments (PED). However, as it one of the key features for cRORA diagnosis absence of definite hypertransmission may lead graders to misdiagnose cRORA especially those lesions lying on top of a PED [10]. Continuity of hypertransmission was another feature that was difficult to detect. In some eyes with RPE loss, the upward intraretinal migration of pigment tends to cause back shadowing in the OCT scan thereby interrupting the continuous hypertransmission resulting it as graded to be discontinuous. So even though the RPE cells are lost as per cRORA definition, the hypertransmission can be discontinuous and thus appears as if all criteria are not met. RPE attenuation was the most reliably detected parameter, however the agreement was still less than 0.80 (κ 0.72). The CAM group acknowledged presence of persistent laminar deposit might interfere with accurate assessment of RPE attenuation [10].

Ascertaining RPE continuity was increasingly complex in eyes with multiple drusen. In the case of BW images, the markedly pigmented RPE nuclei characteristically stand out as dark black line against the less dark EZ and ELM lines. This probably enhanced the ease with which one can detect its loss. Inner retinal layer changes were better documented on WB images.

There are several limitations to the study. First, only five graders were included. This may limit generalisability of our findings. However, we note that this is a fairly new classification and, to the best of our knowledge, a study on inter-grader reliability of clinicians not involved in the CAM classification has not been done previously. Second, there are technical limitations on image grading. We used pre-saved images where colour and image contrast settings may have affected the assessment of parameters. Third, we did not

re-train our graders and perform a post-test grading to test if repeated training will improve reliability. However, as two of the graders who have repeatedly used the CAM criteria had higher agreement, it reinforces that training will likely improve the reliability of diagnosis cRORA based on these criteria.

In conclusion, the CAM classification provides a well-thought through classification and criteria for OCT-defined atrophy in the setting of AMD. The ability to identify these OCT changes reproducibly is essential to understand the natural history of the disease, to identify high-risk signs of progression, and to study the effects of early interventions. This study adds insight into the reproducibility of these parameters in the real world and the need for training for clinicians to accurately identify them, prior to implementing their use in clinical practice.

Summary

What was known before

- CAM classification defines OCT based parameters for accurate identification of atrophy.
- However, inter-grader reliability for diagnosis of cRORA is poor for individuals not accustomed with the CAM criteria.

What this study adds

- This study emphasizes the need for training clinicians to identify cRORA accurately, prior to implementing it in clinical practice.

Acknowledgements SC and RR are supported by the ORNATE India grant (GCRF UKRI MR/P207881/1). The research was supported by the NIHR Biomedical Research Centre and the Clinical Research Facility at Moorfields Eye Hospital NHS Foundation trust and UCL Institute of Technology.

Funding The study is funded by Fight For Sight (Grant code 1905).

Compliance with ethical standards

Conflict of interest The authors declare no competing interests.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Schmitz-Valckenberg S, Sadda S, Staurengi G, Chew EY, Fleckenstein M, Holz FG. Geographic atrophy: semantic considerations and literature review. *Retina*. 2016;36:2250–64.
2. Klein R, Davis MD, Magli YL, Segal P, Klein BE, Hubbard L. The Wisconsin age-related maculopathy grading system. *Ophthalmology*. 1991;98:1128–34.
3. Age-Related Eye Disease Study Research Group. The age-related eye disease study system for classifying age-related macular degeneration from stereoscopic colour fundus photographs: the age-related eye disease study report number 6. *Am J Ophthalmol*. 2001;132:668–81.
4. Gass JDM. *Stereoscopic atlas of macular diseases*. 1st ed. St. Louis: The C. V. Mosby Company; 1970.
5. Bird AC, Bressler NM, Bressler SB, Chisholm IH, Coscas G, Davis MD, et al. An international classification and grading system for age-related maculopathy and age-related macular degeneration. The International ARM Epidemiological Study Group. *Surv Ophthalmol*. 1995;39:367–74.
6. Fleckenstein M, Charbel Issa P, Helb HM, Schmitz-Valckenberg S, Finger RP, Scholl HP, et al. High-resolution spectral domain-OCT imaging in geographic atrophy associated with age-related macular degeneration. *Investig Ophthalmol Vis Sci*. 2008;49:4137–44.
7. Schmitz-Valckenberg S, Fleckenstein M, Göbel AP, Sehmi K, Fitzke FW, Holz FG, et al. Evaluation of autofluorescence imaging with the scanning laser ophthalmoscope and the fundus camera in age-related geographic atrophy. *Am J Ophthalmol*. 2008;146:183–92.
8. Shen LL, Liu F, Nardini HG, Del Priore LV. Reclassification of fundus autofluorescence patterns surrounding geographic atrophy based on progression rate: a systematic review and meta-analysis. *Retina*. 2019;39:1829–39.
9. Biarnés M, Forero CG, Arias L, Alonso J, Monés J. Reappraisal of geographic atrophy patterns seen on fundus autofluorescence using a latent class analysis approach. *Investig Ophthalmol Vis Sci*. 2014;55:8302–8.
10. Sadda SR, Guymier R, Holz FG, Schmitz-Valckenberg S, Curcio CA, Bird AC, et al. Consensus definition for atrophy associated with age-related macular degeneration on OCT: classification of atrophy report 3. *Ophthalmology*. 2018;125:537–48.
11. StatsToDo. Kappa (Cohen & Fleiss) for ordinal data program. 2020. <https://www.statstodo.com/CohenFleissKappa.Pgm.php>.
12. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med*. 2012;22:276–82.
13. Mendonça LSM, Perrott-Reynolds R, Schwartz R, Madi HA, Cronbach N, Gendelman I, et al. Deliberations of an International Panel of Experts on OCTA nomenclature of nAMD. *Ophthalmology*. 2020;S0161-6420:31198–2.