Acta Ophthalmologica

Semi-automated quantification of geographic atrophy with blue-light autofluorescence and spectral-domain optical coherence tomography: a comparison between the region finder and the advanced retinal pigment epithelium tool in the clinical setting

Adrian Reumueller,[1] Stefan Sacu,[1] Maria Georgia Karantonis,[1] Irene Steiner,[2] Guenther Weigert[1] and Ursula Schmidt-Erfurth[1]

[1]Department of Ophthalmology and Optometry, Medical University of Vienna, Vienna, Austria
[2]Center for Medical Statistics, Informatics, and Intelligent Systems, Section for Medical Statistics, Medical University of Vienna, Vienna, Austria

ABSTRACT.

Purpose: **To compare inter- and intraobserver reliability and intermodality agreement on quantification of geographic atrophy, using two routinely available quantification tools, based on blue-light fundus autofluorescence (BAF) and spectral-domain optical coherence tomography (SD-OCT).**

Methods: **Quantifications of atrophic lesions within the central 5 mm of 30 eyes from 30 patients (mean age: 76.1 years) were independently performed by two clinicians on BAF images using the REGION FINDER (RF; Heidelberg Engineering) and on SD-OCT using the advanced retinal pigment epithelium tool (ARPET; Carl Zeiss Meditec) at baseline and follow-up (mean interval: 336 days). Inter- and intraobserver reliability was determined by intraclass correlation coefficients (ICC) and Bland–Altmann plots. Additionally, graders rated the experienced difficulty of each measurement.**

Results: **Intraclass correlation coefficients (ICC) showed excellent inter- and intraobserver reliability with values between 0.994 and 0.998 for RF and slightly higher values for ARPET of 0.997 and 0.999. Bland–Altman plots showed smaller variability for ARPET. Mean interobserver differences (95% CI) for size measurements were −0.11 (−0.27; 0.05) (baseline) and −0.05 mm² (−0.18; 0.08) (follow-up) for RF and −0.04 (−0.14; 0.06) and −0.06 mm² (−0.14; 0.02) for ARPET. Measurements of lesions were on average 0.57 mm² (0.35; 0.79) or 7.6% larger in ARPET. Lesion size between graders did not differ significantly. There was no statistically significant difference in relative enlargement rates between methods. There was poor to moderate agreement between graders when rating the experienced difficulty of each measurement.**

Conclusion: **Semi-automated analysis of geographic atrophy with RF and ARPET is equally reliable and reproducible in clinical settings, despite both algorithms require frequent adjustment by users. The ARPET restricts size measurements to the central 5 mm, which limits its ability to fully track GA progression. Results of both tools are not interchangeable as measurements with ARPET result in larger lesion sizes.**

Key words: advanced RPE tool – blue-light autofluorescence – geographic atrophy – interobserver reliability – intraobserver reliability – optical coherence tomography – outer retinal atrophy – region finder – semi-automated quantification

Acta Ophthalmol. 2019: 97: e887–e895

doi: 10.1111/aos.14085

Introduction

Age-related macular degeneration (AMD) is a major burden for health care systems, and demographic trends predict a significant rise in its global prevalence, from an estimated 196 to 288 million patients between 2020 and 2040 (Wong et al. 2014; Colijn et al. 2017). Geographic atrophy (GA) presents as localized atrophy of the retinal pigment epithelium (RPE), the outer retina and the choriocapillaris in advanced AMD, causing central scotomas and permanent loss of visual acuity (Sayegh et al. 2017; Fleckenstein et al. 2018). While no approved treatment for GA is available at present, several studies and trials focus on understanding the progression patterns or altering the enlargement speed of atrophic areas (Bandello et al. 2017; Evans & Lawrenson 2017; Sacconi et al. 2017). Monitoring GA development and progression over time is therefore invaluable in both research and clinical settings. Various imaging modalities are being used for this matter, ranging from conventional fundus colour photography and near-infrared reflectance imaging (IR) over blue-light fundus autofluorescence (BAF) or green-light fundus autofluorescence (GAF) to spectral-domain (SD), swept-source or polarization-sensitive optical coherence tomography (Schütze et al. 2013; Yehoshua et al. 2015; Domalpally et al. 2016). While multimodal imaging with a mixture of several modalities can be used to measure size and enlargement of GA lesions in clinical trials and professional reading centres, limited technical and financial resources can hinder such an approach in clinical routine settings (Holz et al. 2017). In an attempt to support graders and clinicians alike, semi-automated algorithms have been developed for different imaging modalities, aiming to improve inter- and intraobserver reliability (Schmitz-Valckenberg et al. 2011; Chen et al. 2013; Hu et al. 2013; de Sisternes et al. 2017). Computer-assisted GA analysis is a promising approach, but the number of commercially available software is very limited and there is a need for quality and performance assessments in 'real-world' settings, as pointed out in a recently published review (Wintergerst et al. 2017).

The present study aims to provide such information by comparing inter- and intraobserver reliability of two proprietary semi-automated quantification tools in a clinical setting, outside reading centre conditions. The region finder (RF) software in conjunction with BAF using the Heidelberg Spectralis imaging platform (Heidelberg Engineering, Heidelberg, Germany) was compared with the advanced RPE tool (ARPET) in conjunction with SD-optical coherence tomography (SD-OCT) on a Cirrus HD-OCT platform (Carl Zeiss Meditec, Dublin, OH, USA), as these imaging modalities and software tools are widely available and easily accessible (Schmitz-Valckenberg et al. 2011; Yehoshua et al. 2011). Additionally, factors with a possible influence on agreement between graders such as the presence of multifocal lesions were analysed. We further assessed if there was a difference in measured lesions size or lesion progression between the two software tools. Finally, we evaluated if graders were agreeing on the subjective difficulty of every measurement and if the used quantification tool had an influence on the difficulty rating.

Material and Methods

This retrospective longitudinal study comparing two tools for semi-automated quantification of GA lesions was approved by the Medical University of Vienna institutional review board and conducted in compliance with the Declaration of Helsinki.

Study cohort

Subjects under regular clinical surveillance due to GA were selected from the database of the outpatient clinic for macular diseases. Selection procedure was performed as followed: In total, thirty patients with the diagnosis 'advanced nonneovascular AMD' were found who had received same day BAF on a Heidelberg Spectralis imaging platform and SD-OCT imaging on a Cirrus HD-OCT platform with a follow-up of at least 6 months between 01 January 2015 and 31 December 2017. Patients were then pseudonymized with a random number generator Research Randomizer–Version 4.0, Urbaniak, & Plous, (2013), and sequentially, both eyes of each patient (n = 60 eyes) were screened regarding the inclusion and exclusion criteria: Eligible eyes had to be free from a history of neovascularization or other non-AMD related retinal pathologies and refractive errors >6 dioptres. Blue-light fundus (BAF) images had to be acquired with a Spectralis HRA + OCT platform (Heidelberg Engineering), operating with a 488 nm solid state laser and a transverse resolution of 10 μm in high-speed mode. Spectral-domain (SD)-OCT images had to be acquired with a Cirrus HD-OCT 5000 platform (Carl Zeiss Meditec), operating with an 840 nm superluminescent diode, axial and transverse resolutions of 5 and 15 μm and a scan speed of 27 000 A-scans. Image sets from both modalities had to be performed on the same day by trained orthoptists or a professional medical photographer at baseline and at least 6 months later for follow-up. Spectralis BAF images required to be centred at the macular with a 30° × 30° standard field of view (768 × 768 pixels) and at least 20 images in high-speed mode had to be acquired using the built-in automatic real-time tracking function (ART). Images in high-speed mode were selected, as the significantly faster image acquisition (8.8 frames/seconds) compared with high-resolution mode (4.6 frames/seconds) reduces the risk of blurry or unclean averaged images due to patients' eye movement, a frequent problem in patients with advanced AMD. Corresponding IR images were available in high-speed mode with a 30° standard field of view (768 × 768 pixels) and at least 15 images in ART mode. Cirrus SD-OCT scans required to be '200 × 200 Macular Cubes' without motion artefacts and with a signal quality value of at least 6 or higher.

Of the 60 eyes, 12 were not eligible; six due to image quality (insufficient BAF quality due to advanced cataract, minor to major motion artefacts or signal quality of 5 or lower in OCT), three had not been imaged with the required BAF or OCT settings and three eyes had either intermediate or neovascular AMD. Only one eye was included per patient to avoid possible inclusion bias caused by individual factors that could interfere with a measuring method (for example variations in autofluorescence signal strength or size of the central area with physiologically reduced autofluorescence). In patients with two eligible eyes, which was the case for 18 of the 30 patients, the study eye was chosen randomly using a random number generator (www.randomizer.org).

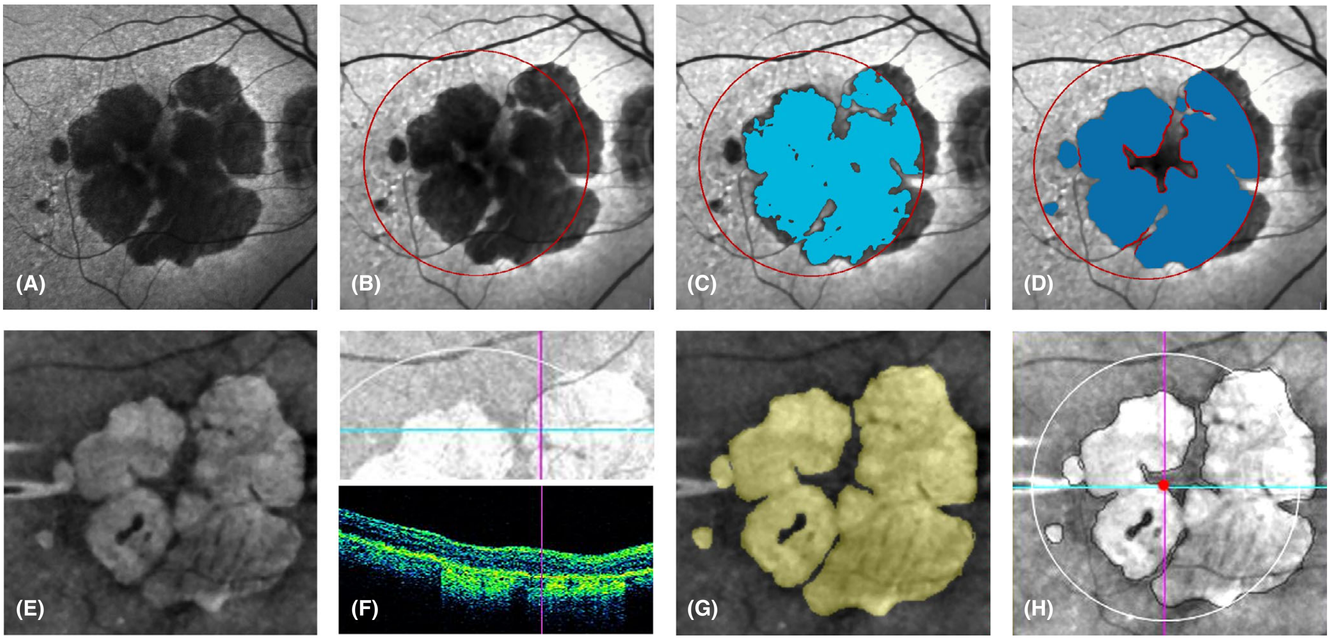

Fig. 1. Steps of measuring atrophic areas with REGION FINDER (first row) and advanced retinal pigment epithelium tool (ARPET) (second row). (A) Blue-light fundus (BAF) image shows non-autofluorescent atrophic areas in dark grey to black. Borders of the spared fovea are hardly visible. (B) A circular constraint (red circle) was manually placed to limit the workspace of graders analogous to the 5 mm border of the ARPET (compare to the white circle in image H). (C) After planting a seed, a grader 'filled' the region by adjusting the threshold for grey values (blue area). Without additional constraints, the spared fovea is falsely included in the atrophic lesion by the algorithm. (D) Final result. Manual constraints were drawn by grader 1 to demarcate foveal sparing. The measured area is 'capped' at the 5 mm circular constraint. (E) Optical coherence tomography (OCT) fundus image. In contrast to BAF, atrophic areas appear bright and foveal sparing is clearly visible. (F) Sub-retinal pigment epithelium (RPE) slab (top) and corresponding OCT image (bottom). B-Scan position is shown by the blue line. Areas with abnormal beam penetration are depicted in bright white. (G) Sub-RPE illumination map allows graders to manually correct the algorithm's suggested area of atrophy. The yellow area is erased or enlarged using the information of the sub-RPE slab and B-Scan (image F). (H) Final Result. Atrophic areas are shown in bright white with black borders. Only areas within the 5 mm circle are measured. There are subtle differences regarding the presence of atrophy between BAF and spectral-domain-OCT (compare the areas in images D and G).

All images had been taken in mydriasis, achieved routinely by instillation of 1% tropicamide (Mydriaticum Agepha, Vienna, Austria) and 2.5% phenylephrine eye drops. Regarding the characteristics of the atrophic lesions, no exclusion criteria were used and eyes with uni- and multifocal lesions as well as well-demarcated and less well-demarcated lesion borders were included.

Procedures

Included images were anonymized and put into random order by a co-author (GW) not involved in the assessment of GA size. Measurements with the RF would have been possible over the whole 30° × 30° BAF image, while the ARPET limited measurements to a 5 mm circle on the SD-OCT image. To make the working area comparable between the two programs, the circular constraint tool of the RF was used by the co-author (GW) to draw a 5 mm circle on the BAF image analogous to the 5 mm circle of the ARPET, using vessels and GA borders as landmarks for placement (Fig. 1A,B,H). The positions of the 5 mm circular constraints on RF images and the 5 mm circles on the ARPET images were not altered for the rest of the study, ensuring that both graders performed measurements in the same area.

Grader 1 (AR) and grader 2 (MGK) performed measurements independently, on separate days and with the same image order to prevent uneven learning curves. Measurements were performed 1 week apart, beginning with ARPET and RF baseline images and followed by ARPET and RF follow-up images. For intraobserver reliability, one grader (AR) did a second measurement cycle. All measurements were performed directly on the same Cirrus HD-OCT and Spectralis HRA + OCT platforms.

For the ARPET, graders were instructed to start with the suggested quantification of the algorithm and to use the integrated tools (brush, eraser, bucket), as well as the integrated OCT B-scans to adjust GA lesion size but could not use any other information or imaging modalities (Fig. 1E–G).

For the RF, graders were allowed to use available constraints (line, circular, freehand and contour) and other image correcting functions as well as the corresponding IR image, but could not use images of other modalities, such as blue reflectance images, infrared autofluorescence, GAF or OCT, and were also not allowed to copy atrophic lesions or set constraints from previous images. After planting a 'seed' on the darkest spots within a GA area, graders would increase the growth power stepwise until it exceeded the borders of the GA lesions and then decrease it by one increment, as suggested in detail elsewhere (Fig. 1C,D) (Schmitz-Valckenberg et al. 2011).

For every image, graders were instructed to note if GA was mono- or multifocal. Additionally, the subjective level of difficulty on a scale ranging from easy to somewhat challenging to very challenging was noted for each measurement performed.

Statistical analysis

Analysis of the main question

The main goal of the study was to analyse the inter- and intraobserver reliability for measuring GA lesion size with RF and ARPET. Interobserver reliability was investigated using Bland–Altman plots. Estimated mean differences with 95% confidence intervals are reported for each method and each visit separately. Estimates and 95% confidence intervals for the interclass correlation coefficient (ICC) were computed with the 'irr' software package for R (cran.r-project.org, M. Gamer), based on a single-rating, absolute agreement, two-way random-effects model. Intraobserver reliability was analysed in a similar manner, whereby ICCs were calculated based on a single-rating, absolute agreement, two-way mixed-effects model. Based on 95% confidence intervals, values between 0.5 and 0.75, between 0.75 and 0.9, and >0.90 were considered as moderate, good and excellent reliability (Koo & Li 2016).

Analysis of secondary questions

To examine if the presence of multifocal lesions, retained RPE fragments within the atrophy or atrophy borders within the foveal region had an influence on the agreement between graders, different symbols in Bland–Altman plots were used and analysed descriptively.

To analyse if the grader or the used quantification tool had an influence on measured GA size, a mixed model (SAS Proc mixed) was calculated with 'GA size' as dependent variable and 'patients' as random factors. To analyse if the used quantification tool had an influence depending whether the lesion extended beyond the central 5 mm or not, an interaction term between 'lesion extending beyond central 5 mm' and 'method' was included in the mixed model.

To analyse if the used quantification tool or the grader had an effect on the measured GA progression between baseline and follow-up, a mixed model was calculated with the proportion 'baseline GA size/follow-up GA size' as dependent variable, 'method' and 'grader' as independent variables and 'patient number' as random factor. To meet the model assumptions, 'baseline GA size/follow-up GA size' was transformed by an arcsine square root transformation. We also investigated if the two graders were performing differently regarding the used quantification tool, which was not significant and hence not included in the mixed model.

To analyse if the graders agreed on the difficulty rating of size measurements, interobserver and intraobserver reliability was analysed by weighted kappa coefficients with Cicchetti Allison weights. To analyse if the visit, method and the presence of atrophy borders within the foveal region had an influence on the difficulty rating, an ordinal logistic regression model with patient number as repeated factor (SAS Proc genmod) was calculated. p-values based on the score test as well as odds ratios with 95% Wald confidence intervals are reported.

Statistical analysis was conducted with R 3.3.2 (r-project.org, RCore Team), SAS 9.4 (SAS Institute, Cary, NC, USA) and SPSS 24 (IBM Corp, Armonk, NY, USA). p-values of <0.05 were considered statistically significant.

Results

Patients, eyes and GA lesions

Characteristics of the included patients, eyes and GA lesions are shown in Table 1. Regarding demarcation of lesions, GA areas were generally well demarcated in ARPET as the sub-RPE slab allowed for a high contrast between atrophic and nonatrophic regions. In RF, region borders were either well demarcated, which was the case for nine eyes with multifocal and seven eyes with monofocal lesions at baseline (eight of each at follow-up), or less well demarcated, which was the case for seven multifocal and seven monofocal lesions at baseline and follow-up. Mean confocal scanning laser detector sensitivity for BAF images was 78.4 (SD: 7.7) at baseline and 86.4 (SD: 7.8) at the follow-up visit. Mean signal strength for OCT images was 7.9 (SD: 1.1) at baseline and 7.9 (SD: 1.0) at the follow-up visit.

Interobserver reliability and GA lesion size

Geographic atrophy (GA) lesion sizes at baseline and follow-up for both graders as well as mean differences in measurements between graders are shown in Table 2. Plotting the differences between measurements against their respective mean value (Bland–Altman plot) revealed higher variability in differences for RF in comparison with ARPET (Fig. 2). Lesion size did not correlate visibly or statistically with the measurement-difference between graders (Spearman's ρ = 0.164 and −0.023 for RF and 0.012 and 0.225 for ARPET). Intraclass correlation coefficients (ICCs; 95% CI) showed excellent interobserver reliability for RF with values of 0.994 (0.987; 0.997) at baseline and 0.996 (0.993; 0.998) at follow-up. Intraclass correlation coefficients (ICCs) for ARPET were slightly higher with 0.997 (0.995; 0.999) at baseline and 0.998 (0.996; 0.999) at follow-up.

Table 1. Characteristics of included patients and eyes.

Patients	*n* = 30
Sex	
Male	13
Female	17
Ethnicity	Caucasian
Mean age	76.1 years (SD: 7.6, range 60–88)
Eyes	*n* = 30
Right eyes	17
Left eyes	13
Pseudophakic	
At baseline	13
At follow-up	13
Lesions type	
Multifocal	
At baseline	16
At follow-up*	15
Monofocal	
At baseline	14
At follow-up*	15
Foveal sparing	
At baseline	9
At follow-up	7

* Changes in lesion type at follow-up are caused by geographic atrophy progression.

Intraobserver reliability and GA lesion size for repeated measurements

Geographic atrophy (GA) lesions sizes and mean differences in repeated measurements of Grader 1 are shown in Table 2. Similar to interobserver agreement, variability in differences was slightly larger for RF than for ARPET as shown in the Bland–Altman Plots (Fig. 3).

Intraclass correlation coefficients (ICCs; 95% CI) showed excellent intraobserver reliability for RF with values of 0.997 (0.994; 0.999) at baseline and 0.998 (0.995; 0.999) at follow-up. Intraclass correlation coefficients

Table 2. Size of atrophic area (mm²) within the central 5 mm, assessed with the REGION FINDER and advanced retinal pigment epithelium (RPE) tool, n = 30 eyes.

	Baseline visit					Follow-Up visit				
	Mean	SD	Median	Min|max	Mean difference (95% CI)	Mean	SD	Median	Min|max	Mean difference (95% CI)
REGION FINDER										
Grader 1	8.62	3.96	9.67	2.95|15.78	Grader 1–Grader 2	10.06	4.08	11.40	3.40|17.44	Grader 1–Grader 2
Grader 2	8.73	3.93	9.82	3.07|15.76	−0.110 (−0.269; 0.050)	10.11	4.11	11.66	3.65|17.32	−0.049 (−0.180; 0.081)
Grader 1R*	8.52	3.85	9.83	2.90|15.54	Grader 1–Grader 1R 0.103 (0.003; 0.202)	10.08	4.10	11.67	3.43|17.69	Grader 1–Grader 1R 0.017 (−0.122; 0.088)
Advanced RPE Tool										
Grader 1	8.90	3.66	9.70	3.20|14.20	Grader 1–Grader 2	10.35	3.76	11.50	3.50|15.60	Grader 1–Grader 2
Grader 2	8.93	3.67	9.90	3.00|14.10	−0.037 (−0.135; 0.062)	10.41	3.71	11.60	3.80|15.50	−0.057 (−0.138; 0.025)
Grader 1R*	8.86	3.66	9.70	3.10|14.40	Grader 1–Grader 1R 0.040 (−0.019; 0.099)	10.33	3.74	11.70	3.50|15.60	Grader 1–Grader 1R 0.023 (−0.029; 0.076)

* Repeated measurement of grader 1 for intraobserver reliability.

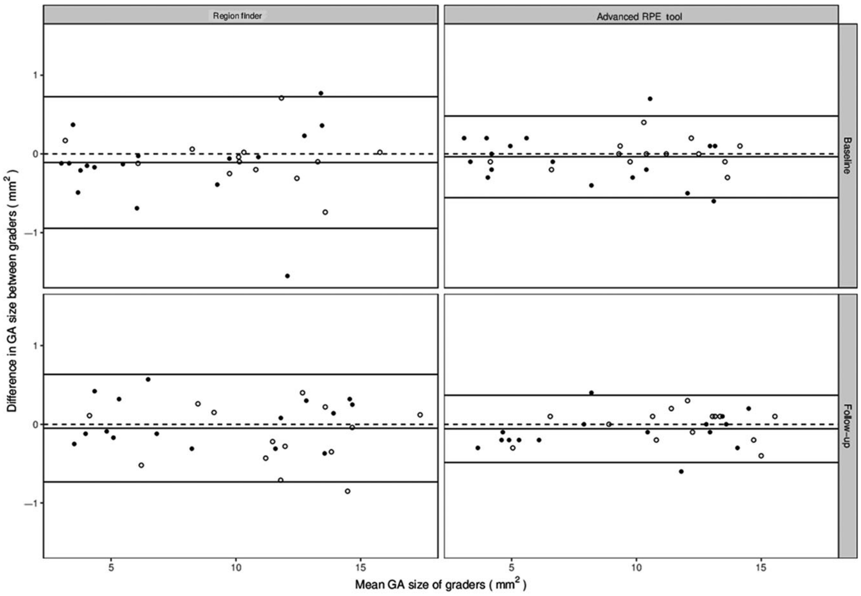

Fig. 2. Interobserver agreement between grader 1 and grader 2 for the REGION FINDER and Advanced retinal pigment epithelium Tool at baseline and follow-up, shown with Bland–Altman plots. Measurements closer to a difference of zero (dashed line) represent higher agreement between graders. Central solid lines represent mean differences. Limits of agreement are shown by the upper and lower solid lines in every plot. Black circles mark eyes where manual constraints had to be placed within the central area of hypoautofluorescence in blue-light fundus (n = 17 at baseline, n = 16 at follow-up). No difference in the distribution between white and black dots is visible.

(ICCs) for ARPET were slightly higher with 0.999 (0.998; 1) at baseline and 0.999 (0.999; 1) at follow-up.

Interobserver performance regarding number of lesions and foveal involvement

There was total agreement between graders and methods when assessing if a patient had mono- or multifocal GA. Lesion numbers per image ranged between 1 and 23 (median = 2) at baseline and between 1 and 17 (median = 1.5) at follow-up and were similar for graders and methods.

To prevent false-positive region growth in RF, 17 images at baseline and 16 at follow-up required placement of constraints in the foveal region. Overall, constraints had to be placed in 55 from 60 BAF images by graders, with an average of 2.67 constraints (SD: 1.73) at baseline and 3.28 (SD: 2.52) at follow-up. Analysis of the Bland–Altman plots showed that neither multifocal lesions (Fig. 4), nor retained RPE fragments (not shown) or GA involvement of the foveal region (Fig. 2) influenced the variability in differences between graders.

Effects of method and grader on measured GA size

Mixed model analysis showed that GA lesion size was not affected by the grader [mean difference (95% CI): 0.063 (−0.11; 0.23), p = 0.470] but by the used quantification tool [ARPET versus REGION FINDER: mean difference (95% CI): 0.19 (0.017; 0.37), p = 0.032] and the interaction between the used tool and the factor ‘lesion extending beyond central 5 mm’ (p < 0.0001). For the latter, subset analyses were performed to investigate the effect of the used tool separately for patients with lesions extending beyond the automatically (ARPET) or manually (RF) placed 5 mm margin (n = 12) and for patients with lesions that did not expand beyond the central 5 mm circle (n = 18), respectively. For patients with lesions that did not expand over the central 5 mm circle, GA measurements were on average significantly larger when performed with ARPET [mean difference (95% CI): 0.57 (0.35; 0.79), p < 0.0001] compared with the REGION FINDER. Mean size for these lesions at baseline was 6.22 mm² (SD: 2.92) for RF and 6.72 mm² (SD: 2.91) for ARPET and 7.67 mm² (SD: 3.32) and 8.25 mm² (SD: 3.25) at follow-up.

For patients with lesions that were capped by the 5 mm margin (‘lesions extending beyond central 5 mm = yes’), no statistically significant difference between methods was observed [ARPET versus RF: mean difference (95% CI): −0.19 (−0.46; 0.084), p = 0.18]. Regarding the enlargement rate, there was no statistically significant difference between the two methods (mixed model, p = 0.746).

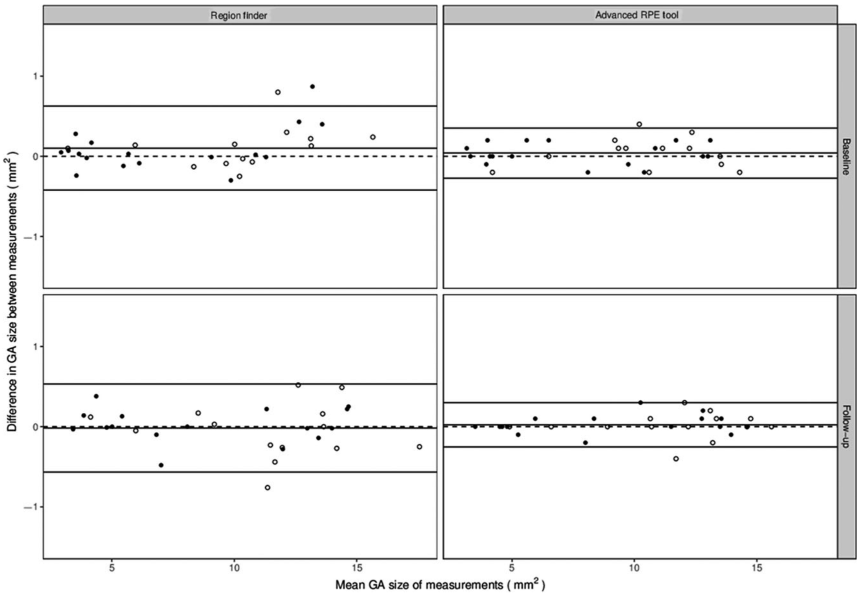

Fig. 3. Intraobserver agreement for grader 1 for the REGION FINDER and Advanced retinal pigment epithelium Tool at baseline and follow-up, shown with Bland–Altman plots. Central solid lines represent mean differences. Limits of agreement are shown by the upper and lower solid lines in every plot. Black circles mark eyes where manual constraints had to be placed within the central area of hypoautofluorescence in blue-light fundus (*n* = 17 at baseline, *n* = 16 at follow-up). White circles mark eyes where no constraints had to be placed within the central area. Results are comparable to interobserver agreements (compare Fig. 1) but show generally less variability and narrower limits of agreement.

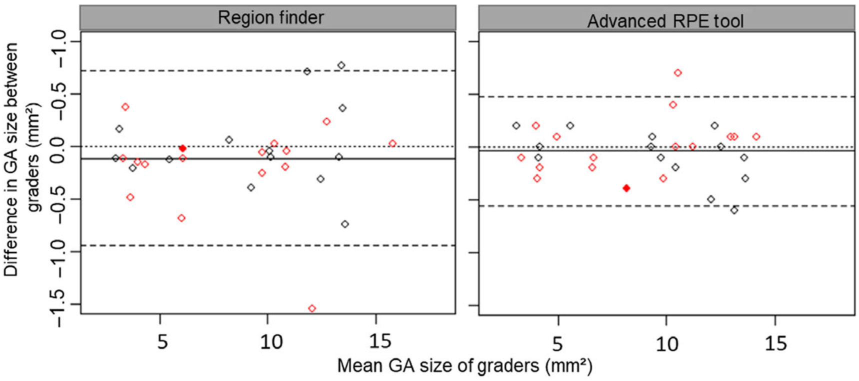

Fig. 4. Influence of lesion type on interobserver agreement at baseline. Central solid lines represent mean differences. White circles mark eyes with monofocal lesions. Red circles mark eyes with multifocal lesions. The solid red circle marks the eye with the highest number of lesions (*n* = 23). No influence of multifocal lesions on interobserver agreement can be found.

Subjective difficulty for graders

Ordinal difficulty ratings were similarly distributed between graders with RF at baseline and follow-up. Both graders rated ARPET measurements slightly easier at baseline and at follow-up, but no statistically significant effect of grader, visit, method or the interaction between foveal constrains and method on the difficulty rating could be observed (ordinal logistic regression with repeated measurements: method: p = 0.34, visit: p = 0.24, grader: p = 0.76, interaction foveal constraints* method: p = 0.95). Blue-light fundus (BAF) measurements were rated as more difficult, when the graders had to place constraints in the central foveal area [foveal constraints yes versus no: odds ratio (95% CI) = 2.93 (1.22; 7.01), p = 0.0249]. For the individual measurements, weighted kappa coefficients (95% CI) showed only low to moderate interobserver agreement [at baseline: RF: 0.153 (−0.132; 0.437), ARPET: 0.463 (0.211; 0.715); at follow-up: RF: 0.250 (−0.028; 0.528), ARPET: 0.391 (0.160; 0.622)] and moderate to substantial intraobserver agreement [at baseline: RF: 0.344 (0.069; 0.620), ARPET: 0.756 (0.546; 0.966); at follow-up: RF. 0.407 (0.156; 0.658), ARPET. 0.554 (0.323; 0.785)].

Discussion

The purpose of this study was to compare two commercially available options for semi-automated quantification of GA in a clinical setting: the RF in conjunction with BAF and the ARPET with SD-OCT.

The two-dimensional BAF images used with the RF are not true anatomical images, but distribution maps of fluorescent compounds contained in lipofuscin, a not degradable, highly oxidized aggregation of cross-linked proteins that can be found in several human tissues as a result of cellular ageing and oxidative stress and other fluorophores (Delori et al. 1995; von Rückmann et al. 1995; Jung et al. 2007). The contrast difference of autofluorescent nonatrophic areas and nonautofluorescent atrophic lesions improves analysis over fundus colour photos (von Rückmann et al. 1997). The RF utilizes these contrast differences to form regions of similar BAF intensity, a principle known as binary enlarged objects (‘BLOB’) – detection, and calculates their size with a scaling factor registered during image acquisition, which has been reported to be more accurate than manual outlining of GA lesions (Schmitz-Valckenberg et al. 2002, 2011). Drawbacks emerge from the technique of BAF and the RF algorithm: First, the distinction between healthy and atrophic regions around the central fovea is known to be challenging due to physiological autofluorescence signal reduction in this area (Delori et al. 1995; Sayegh et al. 2015; Domalpally et al. 2016). This signal reduction results in false-positive region growth of the RF’s algorithm when grading foveal atrophic lesions, a problem we encountered frequently as nearly half of all BAF images required placement of central constraints (Fig. 1C), which had a negative impact on the difficulty ratings of the graders but no significant impact on agreement between graders. Second, wrong region growth of the algorithm into vessels and beyond less well-demarcated lesions borders has been described (Schmitz-

Valckenberg et al. 2002, 2011), which required constraint placement in 92% of BAF images in the present study, underlining that BLOB-based region-detection is highly prone to false-positive region growth. While the combination of BAF analysis with imaging techniques that show less to none foveal signal reduction, such as GAF or IR, can help in distinguishing healthy from pathologic regions, it does not solve the false-positive region growth, which is inherent to BLOB-based lesion detection.

In contrast to two-dimensional imaging techniques, the ARPET uses cross-sectional information of a three-dimensional volume, acquired by a high-definition SD OCT (Cirrus HD-OCT). Summing and merging the signal of each A-scan allows for creation of OCT fundus images (see Fig. 1E) which can be used to outline atrophic areas (Bearelly et al. 2009; Yehoshua et al. 2011; Pilotto et al. 2015). In contrast to a conventional OCT fundus image, the ARPET utilizes the different light scattering properties of retinal tissue such as RPE, choriocapillaris or choroid to create a sub-RPE en-face image based on beam penetration, known as the sub-RPE slab (Fig. 1H) (Yehoshua et al. 2011; Augustin 2012). Although the ARPET algorithm detects GA areas in the sub-RPE slab automatically as opposed to manual seed planting with the RF, manual correction was necessary in all images, showing that the software regularly demarcates lesion borders incorrectly. Additionally, while en-face OCT images have advantages over BAF images such as a constant signal around the fovea or availability of depth information (Fig. 1A,E,F), they are prone to motion artefacts due to longer scanning duration and interpolated, meaning that resolution depends on the number of performed scans (Sayegh et al. 2011; Yehoshua et al. 2011; Pilotto et al. 2015).

Although both methods rely on different technical backgrounds and software algorithms, they showed excellent inter- and intraobserver reliability and consistent measurements were performed without a measurable effect of graders. For RF, the mean interobserver differences between GA size measurements were 0.11 mm² (baseline) and 0.05 mm² (follow-up), which is very similar to the results of professional graders, who reported interobserver differences from 0.00 to 0.44 mm² in 30 eyes from 30 patients but were additionally allowed to use blue reflectance images to obtain additional information on the lesion borders and could copy constraints and regions from baseline to follow-up visits (Schmitz-Valckenberg et al. 2011). A second study on 29 eyes from 20 GA patients assessed intra- and interobserver agreements with RF between nonprofessional graders and found mean interobserver differences from 0.02 to 0.27 mm² (Panthier et al. 2014). Mean intraobserver differences were smaller than interobserver differences in the mentioned studies (−0.02 to 0.32 mm² for professional graders (Schmitz-Valckenberg et al. 2011), −0.01 to −0.17 mm² for nonprofessionals (Panthier et al. 2014)) as well as in the presented one (−0.02 and 0.10 mm²).

For ARPET, mean inter- and intraobserver differences in GA size were at the level of RF, with values around 0.05 mm². While no detailed reports for ARPET were found, two studies by Yehoshua et al. (2011, 2015) stated excellent inter- and intraobserver reliability when using a 6 × 6 mm square SD-OCT fundus image (not to be confused with the sub-RPE slab of ARPET) and measuring GA size with external proprietary software (adobe photoshop cs2, Adobe Systems, San José, California, USA) and a stylus-driven digitizing tablet. A big advantage of RF and ARPET over such approaches is, that both programs are 'ready to use' on either the Heidelberg Spectralis imaging platform or the Cirrus SD-OCT and measurements can be done easily on the spot, which is more important for clinical use than for specialised reading centres.

Bland–Altman plots revealed that variability between graders' measurements was even smaller with ARPET. This finding is unexpected, as ARPET relies more on direct mouse-driven region border correction and seems therefore to be more prone to human error than utilization of the region growth algorithm of RF. However, this finding was consistent for all measurements. Bland–Altman plots for RF in the presented study showed even slightly less variability in comparison with other reports, indicating that graders of this study achieved the expected performance (Schmitz-Valckenberg et al. 2011; Panthier et al. 2014). Region size outputs between the tools differ, as RF rounds region size to three decimal places whereas ARPET to just one decimal place, but the impact in decimal places is too subtle to explain for the difference in variabilities. From the current point of view, a possible explanation could be the threshold-driven approach of the region growing algorithm, as local disagreement between graders still affects the whole perimeter of the lesion. Panthier et al. (2014) reported a strong rise in observer disagreement for GA lesions of a size above 15.75 mm². They addressed this issue to the region growing algorithm, as size changes by adjusting the grey-scale threshold is ultimately proportional to the current lesion size, thus making larger regions more prone to substantial disagreement. A small area with a low contrast between atrophic and nonatrophic regions on a BAF image can thus affect the measured size of the whole lesion.

Intraclass correlation coefficients (ICC) were between 0.994 and 0.998 for RF and between 0.997 and 0.999 for ARPET, indicating excellent inter- and intraobserver agreement. A recent paper comparing BAF and GAF for measuring GA (both performed with RF) reported ICCs of 0.995 (GAF) and 0.991 (BAF), which were marginally smaller than the ones found in this study (Pfau et al. 2017).

Regarding the influence of methods on GA size, measurements were on average about 7.6% or 0.5 mm² larger when performed with ARPET. This noticeable difference is in agreement with an abstract on 15 eyes with GA, reporting mean lesion size of 4.8 mm² with ARPET and 4.1 mm² with manual outlining on colour fundus photos (Sharma et al. 2011). Post hoc comparison of the measurements in the current study showed that demarcations of GA areas on sub-RPE slabs were smoother and slightly larger than on BAF images (Fig. 5). Regarding lesion borders, BAF images appeared more detailed. Occasionally, RPE presence between GA lesions was indicated by autofluorescence in BAF, whereas the same region on SD-OCT indicated atrophy and absence of RPE (Fig. 1B, D,F,H). The less detailed images found with ARPET are most likely a result of the lower transversal resolution of the

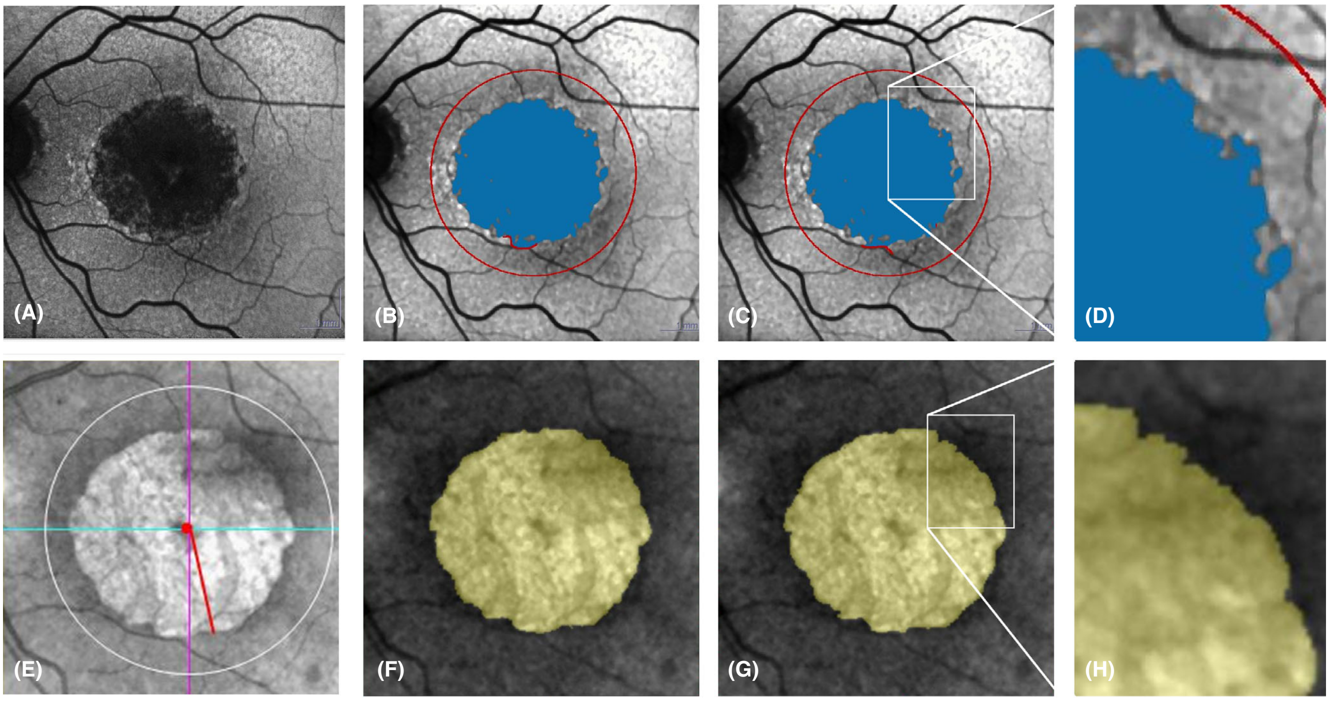

Fig. 5. Comparison of geographic atrophy (GA) measurements between graders with REGION FINDER (RF) and advanced retinal pigment epithelium tool (ARPET). (A,E) Blue-light fundus (BAF) and sub-retinal pigment epithelium (RPE) slab of monofocal GA. (B,C,F,G) Final results of grader 1 (B and F) are very similar to the results of grader 2 (C and G). Lesion size was 9.360 (grader 1) and 9.367 (grader 2) with RF and 10.7 and 10.7 with ARPET. (D,H) Magnification of corresponding areas (white rectangles in C and G) shows that region borders are more detailed in BAF while the sub-RPE slab shows smoother but larger areas.

Cirrus OCT compared with the Heidelberg Spectralis confocal scanning laser ophthalmoscope. Additionally, the OCT en-face and sub-RPE slap images are interpolated, which results in a loss of information. Besides differences in area demarcation, the scaling factors for both imaging platforms are not identical and a difference in region size can also be caused by the underlying technique of image scaling. We did not find a difference between the methods regarding GA size for lesions that did expand over the 5 mm border, which can be explained as followed: The 5 mm margin acts as a smooth artificial lesion border in RF as well as ARPET, thus masking the differences in depiction of the real lesion borders between the two methods and concealing possible differences in lesion size.

Ordinal difficulty ratings for each measurement were not useful, as agreement between graders was moderate at best. Overall difficulty was slightly lower with ARPET in comparison with RF, but no statistical significant difference was found. The weak agreement on difficulty ratings had no effect on the excellent inter- and intraobserver reliability of size measurements, indicating that graders agree with a measurement even if they disagree with the difficulty of performing it.

To the best of our knowledge, this is the first study comparing GA analysis with RF against ARPET. However, several limitations arise: Due to its retrospective nature, typical design problems such as selection bias can occur. Carefully defined inclusion and exclusion criteria were used to minimize this risk, and all included images were acquired by the same team of qualified orthoptists and medical photographers. Despite the excellent levels of agreement, the clinicians in this study were nonprofessional graders and results of this study have to be regarded as such. We did not assess resource factors such as grading time duration, as these attributes are of more relevance for professional reading centres and should therefore be performed in these settings. Regarding image resolution, choosing BAF images in high-resolution mode (1536 × 1536 pixels) could have provided even more details than in high-speed mode (768 × 768 pixels), which is beneficial when analysis of highly magnified images (15° × 15°) is performed. However, this was not the case in our study and high-speed mode has successfully been used when quantifying GA with RF (Schmitz-Valckenberg et al. 2011).

Finally, the workspace of RF had to be manually adjusted with the circular constraint to overlap with ARPET, as the latter limits the workspace automatically to a 5 mm circle. However, not doing so would have led to selection bias, as the RF algorithm causes errors in inter- and intraobserver agreement for large lesions as stated above (Panthier et al. 2014).

Concluding, RF and ARPET are equally suitable for measuring GA size in clinical settings as they both reach excellent inter- and intraobserver reliability but require significant adjustments by graders. REGION FINDER (RF) allows for analysis of larger areas, compared with ARPET, which restricts the user to a 5 mm circle. This limits the use of the ARPET to lesions that are not extensively progressed, ultimately limiting its potential to fully track GA enlargement in every patient in clinical practice, which was the case in 40% of our subjects. However, it also has to be noted that size measurements of large lesions with the RF are less reliable, as mentioned above. The characteristic hypoautofluorescent area of the central fovea in BAF images had no significant impact on reliability, but measurements were more challenging when GA borders were present at the foveal region. Intraclass correlation coefficients (ICCs) for size measurements are excellent, even if graders agree generally poorly on the subjective difficulty of the performed measurements.

Contrary to what one would expect, inter- and intraobserver agreement is even better with ARPET, although the clinical relevance of the difference is debatable. The choice of method has an impact on the measured GA size, as lesions were on average 7% larger when using ARPET in comparison with RF. As a result, size measurements are not interchangeable between methods and clinicians outside of clinical trials should stick to a single imaging modality for longitudinal observations or aim for a multimodal imaging approach. Future studies will be necessary to find the best combination of imaging modalities as recently suggested by the Classification of Atrophy Meetings group (Holz et al. 2017).

References

Augustin AJ (2012): Advanced retinal pigment epithelium analysis by SD-OCT to monitor dry AMD progression. Imaging Med **4**: 251–259.

Bandello F, Sacconi R, Querques L, Corbelli E, Cicinelli MV & Querques G (2017): Recent advances in the management of dry age-related macular degeneration: a review. F1000Research **6**: 245.

Bearelly S, Chau FY, Koreishi A, Stinnett SS, Izatt JA & Toth CA (2009): Spectral domain optical coherence tomography imaging of geographic atrophy margins. Ophthalmology **116**: 1762–1769.

Chen Q, de Sisternes L, Leng T, Zheng L, Kutzscher L & Rubin DL (2013): Semi-automatic geographic atrophy segmentation for SD-OCT images. Biomed Opt Express **4**: 2729–2750.

Colijn JM, Buitendijk GHS, Prokofyeva E et al. (2017): Prevalence of age-related macular degeneration in Europe: the past and the future. Ophthalmology **124**: 1753–1763.

Delori FC, Dorey CK, Staurenghi G, Arend O, Goger DG & Weiter JJ (1995): *In vivo* fluorescence of the ocular fundus exhibits retinal pigment epithelium lipofuscin characteristics. Invest Ophthalmol Vis Sci **36**: 718–729.

Domalpally A, Danis R, Agrón E, Blodi B, Clemons T, Chew E; Age-Related Eye Disease Study 2 Research Group (2016): Evaluation of geographic atrophy from color photographs and fundus autofluorescence images: age-related eye disease study 2 report number 11. Ophthalmology **123**: 2401–2407.

Evans JR & Lawrenson JG (2017): Antioxidant vitamin and mineral supplements for slowing the progression of age-related macular degeneration. Cochrane Database Syst Rev **7**: CD000254.

Fleckenstein M, Mitchell P, Freund KB, Sadda S, Holz FG, Brittain C, Henry EC & Ferrara D (2018): The progression of geographic atrophy secondary to age-related macular degeneration. Ophthalmology **125**: 369–390.

Holz FG, Sadda SR, Staurenghi G et al. (2017): Imaging protocols in clinical studies in advanced age-related macular degeneration: recommendations from classification of atrophy consensus meetings. Ophthalmology **124**: 464–478.

Hu Z, Medioni GG, Hernandez M, Hariri A, Wu X & Sadda SR (2013): Segmentation of the geographic atrophy in spectral-domain optical coherence tomography and fundus autofluorescence images. Invest Ophthalmol Vis Sci **54**: 8375–8383.

Jung T, Bader N & Grune T (2007): Lipofuscin. Ann N Y Acad Sci **1119**: 97–111.

Koo TK & Li MY (2016): A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med **15**: 155–163.

Panthier C, Querques G, Puche N, Le Tien V, Garavito RB, Béchet S, Massamba N & Souied EH (2014): Evaluation of semiautomated measurement of geographic atrophy in age-related macular degeneration by fundus autofluorescence in clinical setting. Retina **34**: 576–582.

Pfau M, Goerdt L, Schmitz-Valckenberg S, Mauschitz MM, Mishra DK, Holz FG, Lindner M & Fleckenstein M (2017): Green-light autofluorescence versus combined blue-light autofluorescence and near-infrared reflectance imaging in geographic atrophy secondary to age-related macular degeneration. Invest Ophthalmol Vis Sci **58**: BIO121–BIO130.

Pilotto E, Guidolin F, Convento E, Antonini R, Stefanon FG, Parrozzani R & Midena E (2015): En face optical coherence tomography to detect and measure geographic atrophy. Invest Ophthalmol Vis Sci **56**: 8120–8124.

Research Randomizer–Version 4.0, Urbaniak, G. C., & Plous, S. (2013). Retrieved from http://www.randomizer.org/

von Rückmann A, Fitzke FW & Bird AC (1995): Distribution of fundus autofluorescence with a scanning laser ophthalmoscope. Br J Ophthalmol **79**: 407–412.

von Rückmann A, Fitzke FW & Bird AC (1997): Fundus autofluorescence in age-related macular disease imaged with a laser scanning ophthalmoscope. Invest Ophthalmol Vis Sci **38**: 478–486.

Sacconi R, Corbelli E, Querques L, Bandello F & Querques G (2017): A review of current and future management of geographic atrophy. Ophthalmol Ther **6**: 69–77.

Sayegh RG, Simader C, Scheschy U et al. (2011): A systematic comparison of spectral-domain optical coherence tomography and fundus autofluorescence in patients with geographic atrophy. Ophthalmology **118**: 1844–1851.

Sayegh RG, Zotter S, Roberts PK et al. (2015): Polarization-sensitive optical coherence tomography and conventional retinal imaging strategies in assessing foveal integrity in geographic atrophy. Invest Ophthalmol Vis Sci **56**: 5246–5255.

Sayegh RG, Sacu S, Dunavölgyi R et al. (2017): Geographic atrophy and foveal-sparing changes related to visual acuity in patients with dry age-related macular degeneration over time. Am J Ophthalmol **179**: 118–128.

Schmitz-Valckenberg S, Jorzik J, Unnebrink K & Holz FG (2002): Analysis of digital scanning laser ophthalmoscopy fundus autofluorescence images of geographic atrophy in advanced age-related macular degeneration. Graefes Arch Clin Exp Ophthalmol **240**: 73–78.

Schmitz-Valckenberg S, Brinkmann CK, Alten F et al. (2011): Semiautomated image processing method for identification and quantification of geographic atrophy in age-related macular degeneration. Investig Opthalmol Vis Sci **52**: 7640–7646.

Schütze C, Bolz M, Sayegh R, Baumann B, Pircher M, Götzinger E, Hitzenberger CK & Schmidt-Erfurth U (2013): Lesion size detection in geographic atrophy by polarization-sensitive optical coherence tomography and correlation to conventional imaging techniques. Invest Ophthalmol Vis Sci **54**: 739–745.

Sharma S, Huo S & Kaiser P (2011): Comparison of manual versus automated analysis of spectral domain optical coherence tomography (SDOCT) scans in non-neovascular age related macular degeneration. Invest Ophthalmol Vis Sci **52**: 3688.

de Sisternes L, Jonna G, Moss J, Marmor MF, Leng T & Rubin DL (2017): Automated intraretinal segmentation of SD-OCT images in normal and age-related macular degeneration eyes. Biomed Opt Express **8**: 1926–1949.

Wintergerst MWM, Schultz T, Birtel J, Schuster AK, Pfeiffer N, Schmitz-Valckenberg S, Holz FG & Finger RP (2017): Algorithms for the automated analysis of age-related macular degeneration biomarkers on optical coherence tomography: a systematic review. Transl Vis Sci Technol **6**: 10.

Wong WL, Su X, Li X, Cheung CMG, Klein R, Cheng C-Y & Wong TY (2014): Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. Lancet Glob Health **2**: e106–e116.

Yehoshua Z, Rosenfeld PJ, Gregori G, Feuer WJ, Falcão M, Lujan BJ & Puliafito C (2011): Progression of geographic atrophy in age-related macular degeneration imaged with spectral domain optical coherence tomography. Ophthalmology **118**: 679–686.

Yehoshua Z, de Amorim Filho Garcia CA, Nunes RP et al. (2015): Comparison of geographic atrophy growth rates using different imaging modalities in the COMPLETE study. Ophthalmic Surg Lasers Imaging Retina **46**: 413–422.

Received on October 17th, 2018.
Accepted on February 16th, 2019.

Correspondence:
Stefan Sacu, MD
Department of Ophthalmology
Medical University of Vienna
Waehringer Guertel 18-20, 1090 Vienna
Austria
Tel: +43 1 40400-79400
Fax: +43 1 40 400 79320
Email: stefan.sacu@meduniwien.ac.at