tvst

https://doi.org/10.1167/tvst.8.3.37

Article

Reliability of Spectral-Domain OCT Ellipsoid Zone Area and Shape Measurements in Retinitis Pigmentosa

Travis B. Smith[1], Maria A. Parker[1], Peter N. Steinkamp[1], Albert Romo[1], Laura R. Erker[1], Brandon J. Lujan[1], and Ning Smith[2]

[1] Casey Eye Institute, Oregon Health & Science University, Portland, OR, USA
[2] Center for Health Research, Kaiser Permanente, Portland, OR, USA

Correspondence: Travis B. Smith, Casey Eye Institute, 3375 SW Terwilliger Boulevard, Portland, OR 97239, USA. e-mail: smittrav@ohsu.edu

Received: 19 November 2018
Accepted: 22 April 2019
Published: 11 June 2019

Keywords: ellipsoid zone; retina; optical coherence tomography; reliability; shape

Citation: Smith TB, Parker MA, Steinkamp PN, Romo A, Erker LR, Lujan BJ, Smith N. Reliability of spectral-domain OCT ellipsoid zone area and shape measurements in retinitis pigmentosa. Trans Vis Sci Tech. 2019;8(3):37, https://doi.org/10.1167/tvst.8.3.37

Purpose: We investigate the ellipsoid zone (EZ) area and EZ boundary shape measurement reliability and the operability characteristics of two methods of EZ boundary delineation in spectral-domain optical coherence tomography (SD-OCT).

Methods: EZ boundaries in SD-OCT scans of 122 eyes from 64 subjects with autosomal dominant retinitis pigmentosa were delineated by three raters using two methods, termed the profile and en face methods. For each method, we determined the measurement reliabilities for boundary area (EZ area) and boundary shape, percentage of eyes with measurable EZ (measurability), time required, and effect of rater experience.

Results: With expert raters, inter- and intrarater area intraclass correlation coefficients (ICCs) were 0.986 and 0.980 (profile) and 0.959 and 0.976 (en face), respectively. In comparison, the corresponding shape ICCs were 0.906 and 0.891 (profile) and 0.845 and 0.885 (en face), indicating lower reliability for the raw measurements ($P \leq 0.01$). Only profile method interrater reliability depended on experience. Average measurement times per eye were 8.2 (profile) and 4.1 (en face) minutes. Measurability percentages were 99.2% (profile) and 73.0% (en face).

Conclusions: The slower profile method had better measurability, and with expert raters yielded the best area and shape reliabilities. The faster, but less sensitive, en face method still showed excellent reliability, and was less dependent on experience. Shape analysis reveals the boundary measurements underpinning EZ area have lower reliability than suggested by area analysis.

Translational Relevance: This study provides new reliability perspectives and logistical considerations for the manual measurement procedures that generate EZ area outcome measures.

Introduction

In spectral-domain optical coherence tomography (SD-OCT), the ellipsoid zone (EZ) appears as a solid hyperreflective layer throughout the normal macula.[1,2] Retinal degenerations can cause progressive EZ deterioration as photoreceptor integrity degrades. The EZ disruption in retinitis pigmentosa (RP), for example, begins in the periphery and advances centrally as the island of intact photoreceptors erodes.[3,4] This reduction in EZ extent has been associated with commensurate losses in visual field,[5–8] the primary outcome measure in many clinical trials. As potential structural surrogates for visual function, anatomy-based outcome measures, such as the area of intact EZ (EZ area) necessitate delineating the EZ boundary in OCT scans, which often is performed with a manual or semimanual procedure.

We investigated two common methods of EZ boundary delineation, hereby termed the profile and en face methods. Both methods produce an EZ area measurement from an OCT volume scan, are commercially available via vendor software, and require human raters to delineate the EZ boundary manually. With the profile method, raters identify EZ termini on

cross-sectional b-scans from the volume. With the en face method, raters trace the EZ boundary on a contrast-enhanced en face projection of the volume. Here, we determined the interrater and intrarater reliability of these manual measurements, the effect of rater experience, and other important operability characteristics.

We also investigated, for the first time to our knowledge, EZ measurement reliability in terms of shape. Previous studies have focused on the reliability of the EZ area[9,10] or EZ width[11,12] values, but these quantities are not the raw measurements; instead, they are only summary indicators of extent derived from the EZ boundary measurements. Reducing the boundary measurement to simply the area contained therein discards all morphologic components of measurement variability such as shape and position. Consequently, a reliability analysis of EZ area values cannot capture all of the variability in the original, precursory measurements and could yield overly optimistic estimates of endpoint consistency. We developed a shape-sensitive analysis framework that captures the morphologic variability of EZ boundary measurements. We report here the shape reliability using this framework and the area reliability from conventional analysis.

Methods

Participants

All participants were enrolled in the Trial of Oral Valproic Acid (VPA) for Retinitis Pigmentosa (NCT01233609),[13] a phase 2, multisite, randomized, placebo-controlled trial of VPA in a cohort of genetically confirmed autosomal dominant RP patients. The relevant trial inclusion criteria were 20/200 or better visual acuity and 18 years minimum age. Trial exclusion criteria were the presence of other retinal diseases, less than 5° of visual field, or unreliable perimetry measurements in both eyes.[14] Patients provided written informed consent in accordance with the Declaration of Helsinki tenets and an institutional review board (IRB) protocol implemented at each trial site. All patient records were anonymized and deidentified before analysis. This study was determined to be exempt from review by the Oregon Health & Science University IRB in accordance with the Department of Health & Human Services regulation 45 CFR 46.101(b)(4).

We analyzed SD-OCT volume scans (Spectralis OCT; Heidelberg Engineering, Heidelberg, Germany) of both eyes from baseline visits of 89 participants. Data from one of the 90 enrolled subjects were missing at analysis. Trial protocol set the OCT field of view (FOV) to 20° (approximately 6 mm) centered on the fovea with volume scans comprised of 97 horizontal b-scans, each with 512 a-lines, acquired with 5x averaging via automatic real-time tracking (ART). Before this study, we inspected all 178 available scans and visually identified intact EZ completely contained within the OCT FOV in 122. The scans from these 122 eyes from 64 patients served as the OCT data for this investigation.

Raters

Three raters certified in reading center procedures participated in this study. Two were senior raters with more than 4 years of OCT grading and analysis experience each, and the third had no prior OCT experience. All raters used identical calibrated display monitors and workstations installed with the vendor's HEYEX analysis software (Heidelberg Eye Explorer 1.9.10.0, Viewing Module 6.3.4.0), which provided the capability to perform the profile and en face delineation methods. Raters were trained and tested on both methods, which are summarized below and explained in detail in Supplementary Appendices S1 and S2. Raters worked independently and without knowledge of the others' results or their own prior measurements.

Profile Method for EZ Delineation

The profile method (Figs. 1A–C) consists of manually marking the EZ termini or endpoints on all b-scans within the volume in which the EZ is detected, a methodology used in studies by Ramachandran et al.[3] and Hariri et al.[9] We defined the EZ endpoint as the location where the EZ was no longer distinguishable from the retinal pigment epithelium (RPE) complex. Using HEYEX, each rater inspected all b-scans and marked the EZ endpoints on every b-scan where intact EZ was visually apparent. The EZ endpoint coordinates were exported in the native HEYEX .xml format and, using custom software, joined to create a digitized polygon representing the boundary of intact EZ.

En face Method for EZ Delineation

The en face method (Figs. 1D–F) consists of tracing the EZ boundary contour on an en face image with enhanced contrast between regions of intact and absent EZ. We followed the approach described by Hariri et al.,[10] adding certain missing details as

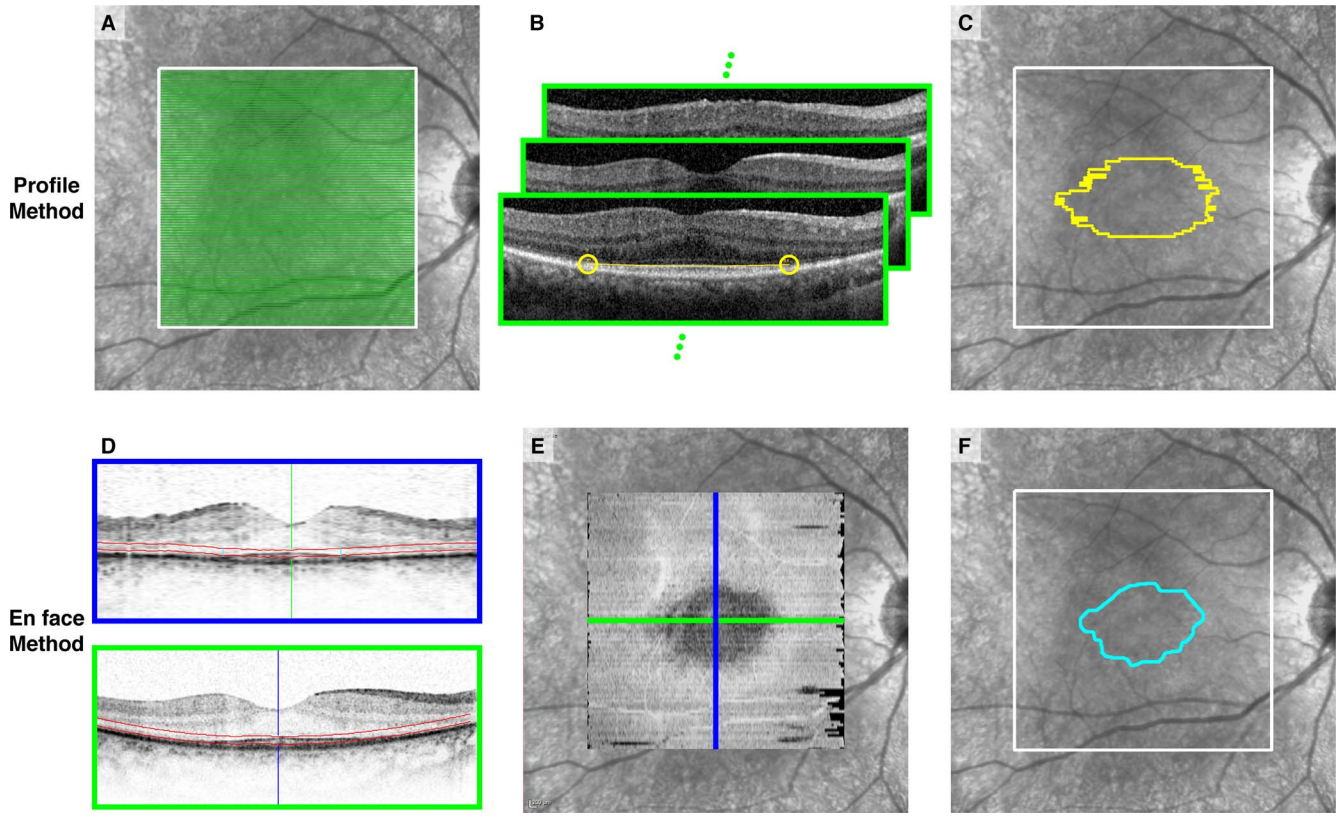

Figure 1. Illustration of the profile method (A–C) and en face method (D–F) of EZ boundary delineation. (A) For the profile method, all b-scans (*green lines*) in the OCT volume scan field of view (*white square*) are inspected. (B) Raters mark the EZ endpoints (*yellow circles*) on each b-scan where intact EZ is found. (C) The endpoints from all b-scans are joined in software to form a polygon representing the EZ boundary (*yellow*). (D) For the en face method, a conformal slab (*red lines*) following the RPE-BM boundary contour is positioned to include the EZ. The grayscale is inverted to setup the MIP operation. (E) The b-scan intensities through the slab are projected to form an MIP en face image with enhanced EZ contrast (the *darker central region*). (F) Raters outline the EZ to create the EZ boundary polygon (*cyan*). Supplementary Appendices S1 and S2 contain complete procedural details for both methods. BM, Bruch's membrane.

described in Appendix S2. Using HEYEX, each rater created an en face image from a minimum intensity projection (MIP) through a conformal cross-sectional slab excised from the volume scan. The upper and lower surfaces of the slab were defined by manually placing copies of the automatically segmented RPE/Bruch's membrane boundary above and below the EZ band. The slab location was adjustable up to a specified limit to allow each rater to optimize the EZ boundary contrast without significantly cropping its extent. Raters traced the EZ boundary on the en face MIP and exported the result. Using custom software, the traced boundary was automatically extracted from the exported image to create another digitized EZ boundary polygon.

Measurability and Timing

Each rater was instructed to delineate the boundary only where intact EZ could be identified and measured. For the profile method, this requirement meant EZ endpoints must be visible on at least one b-scan; for the en face method, this meant the EZ boundary contrast in the MIP image must be sufficient enough for delineation. For this study, we defined measurability as the percentage of OCT volume scans in which a rater determined the EZ boundary was conspicuous enough to delineate. The time required for EZ delineation of a scan was estimated by comparing the exported file timestamps from successive scans.

The raters analyzed all scans with the profile method first and then, masked to those results, they reanalyzed all scans with the en face method. Approximately 1 year after beginning this study, the raters were retrained on both methods and repeated the same sequence of operations. This second round was conducted on a subset of 40 eyes from 20 subjects randomly selected from those found to have measurable EZ by all raters during the first round.

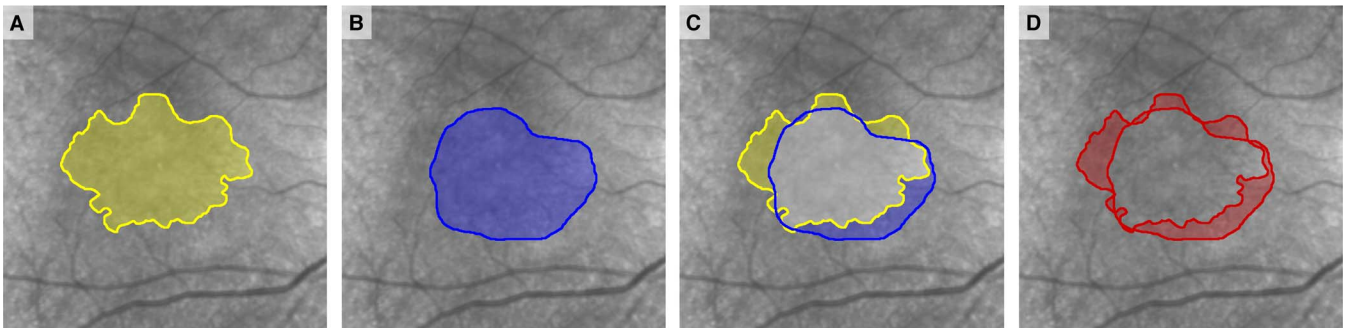

Figure 2. A geometric ambiguity arises when two polygons have different shapes but similar areas. In this example, the areas of the *yellow* (A) and *blue* (B) polygons are 5.65 and 5.69 mm^2, respectively. Knowing only the area values, these measurements would seem to be in good agreement with a small ΔArea of 0.04 mm^2. However, despite their overlap (C), there is a significant disparity between them (D). The area of the *red* region in (D) is the ASD, which quantifies this disagreement. The ASD is 2.19 mm^2, meaning that approximately 20% of the total polygon area is in disagreement. When comparing two EZ boundary measurements, the ASD is a more sensitive metric of comparison than ΔArea because it detects shape and position differences.

Area and Shape Measurements

Repeatability analysis hinges on the metric that quantifies the difference or disparity between two measurements. We analyzed the raters' measurements of EZ boundary polygons with two distinct metrics: the area difference (ΔArea) and the area of symmetric difference (ASD).

The ΔArea metric is the basis of a conventional analysis of EZ area values. The ΔArea is defined as the absolute difference between the areas of the two polygons. This approach has been used in prior studies of EZ area reliability,[9,10,15] and is conceptually similar to the difference in widths used in reliability studies of EZ width measurements from individual b-scans.[6,11,12]

The ΔArea metric has a significant limitation, though: it does not capture shape, orientation, and position differences between EZ boundary measurements. To determine EZ area, an EZ boundary measurement must be distilled down to a scalar summary value, thereby reducing the variability inherent in the original, higher-dimensional measurement. The sensitivity of ΔArea as a metric is impacted by the geometric ambiguity that two dissimilar EZ boundary polygons could have the same or similar EZ area, yielding a small ΔArea that belies their true disparity.

In comparison, the ASD metric is sensitive to morphologic differences between EZ boundary measurements (Fig. 2). The ASD of two polygons quantifies their total area of disagreement, and is defined as the area of the polygons' union minus the area of their intersection.[16–18] The ASD ranges in value from a minimum of zero (signifying the polygons are identical and exactly aligned) to a maximum equal to the sum of both polygons' areas (signifying the polygons are disjoint and overlap nowhere).

Statistical Analysis

We investigated the inter- and intrarater reliability and agreement and the intermethod agreement for both EZ boundary delineation methods.[19,20] Reliability was quantified by the intraclass correlation coefficient (ICC), a value ranging from 0.0 to 1.0 that quantifies the fraction of total measurement variance attributed to the inherent variation between subjects as opposed to measurement error, with larger values indicating better reliability. Several types of ICC are available; here, we used the criterion-reference reliability ICC, labeled as ICC(2,1) by Shrout and Fleiss,[21] which quantifies the degree of absolute agreement among the measurements.[22] This ICC uses a two-way repeated-measures analysis of variance (ANOVA) model with interpatient and interrater random effects, but no patient–rater interaction, which accounts for any systematic differences between the raters while treating them as samples from the population of all raters. We do not report an intermethod ICC because it would be difficult to interpret since the two methods are fixed and not random samples from a population.[23,24]

Agreement was quantified by the repeatability coefficient (RC), the value below which lies with 95% probability the difference between any two measurements, where lower values indicate better agreement.[20] The RC is estimated from the within-subject

Table 1. Duration and Measurability of the EZ Boundary Delineation Methods

	Rater 1	Rater 2	Rater 3	Expert Raters (2 and 3)	All Raters
Measurability, percentage[a] out of (122 eyes, 64 patients)					
Profile	100.0 (122, 64)	100.0 (122, 64)	99.2 (121, 64)	99.2 (121, 64)	99.2 (121, 64)
En face	95.9 (117, 62)	76.2 (93, 53)	84.4 (103, 57)	73.0 (89, 50)	73.0 (89, 50)
Delineation time, minutes, mean [95% CI]					
Profile	11.6 [10.5–12.6]	7.4 [6.0–8.8]	8.9 [7.7–10.1]	8.2 [7.3–9.1]	9.3 [8.6–10.0]
En face	5.5 [4.7–6.2]	4.0 [3.5–4.6]	4.2 [3.7–4.8]	4.1 [3.8–4.5]	4.7 [4.3–5.0]

Rater 1 is novice, raters 2 and 3 are expert.

[a] Measurability is the percentage of eyes in which the EZ boundary was discernible enough to delineate it.

standard deviation in the ANOVA model described above, and is similar to Bland and Altman's limits of agreement[25] and Beckerman's smallest real difference.[26] Whereas the ICC gives a statistical value of repeated measurement consistency relative to the total population variance, the RC provides an absolute assessment in the measurements' native units, which can be more clinically meaningful.

We calculated agreement and reliability coefficients twice, once with each metric. Using the ΔArea metric, we calculated area RC and area ICC. With the ASD metric, we computed shape variance and calculated shape RC and shape ICC, as described in our previous work.[18] Shape variance is formulated identically to area variance (the variance of the EZ area values), but with the ASD metric supplanting the subtraction operation (the ΔArea metric) in the variance formula. The resulting shape RC and shape ICC are directly comparable with the conventional area RC and area ICC, but capture the morphologic variability in the EZ measurements.

The ANOVA model assumes measurements are normally distributed and homoscedastic,[21,22] and the RC formulation assumes measurement differences are normal.[20] Because these assumptions were not intrinsically satisfied, we applied a square root transformation to EZ area measurements to improve distributional characteristics and reduce heteroscedasticity. We confirmed normality with Kolmogorov-Smirnov and Shapiro-Wilk tests and homoscedasticity with Bland-Altman plots.[25] We used the delta method to transform RCs back to original measurement units (mm^2) for ease of interpretation.[27] We determined P values of ICC differences with the modified Fisher's Z-test for dependent ICCs.[28] All quality control, data processing, and statistical analysis software was developed with MATLAB (R2017b, MathWorks, Natick, MA).

Results

Measurability and Timing

Table 1 lists the measurability and timing of each method. All three raters as a group measured the EZ boundary with the profile method in 32 (36%) more eyes than with the en face method. Furthermore, with the profile method there was better unanimity in measurability among raters, whereas the en face method showed less consensus in which eyes had measurable EZ. Supplementary Appendix S2 shows examples of EZ boundaries unmeasurable with the en face method. The en face method was faster ($P < 0.001$), requiring approximately half the time of the profile method, regardless of rater experience. En face method time was independent of the area measured ($P = 0.6$), whereas profile method time increased with area ($P < 0.001$).

Reliability and Agreement

Table 2 shows interrater, intrarater, and intermethod results for both metrics, separated by rater experience level. The profile method interrater area ICC with expert raters was 0.986, better than all other rater/method combinations ($P < 0.04$). The en face method yielded a still excellent interrater area ICC of 0.959 with expert raters, albeit with a much larger area RC. As expected, shape RCs and ICCs were worse than the area RC and ICC counterparts in all scenarios, often significantly so. Intermethod agreement was worse than expert interrater agreement, indicating that the two delineation methods were less interchangeable than two expert raters. Bland-Altman plots of the interrater, test–retest, and intermethod transformed area measurements are shown in Figure 3.

For the intrarater study, the random subset of 40 eyes (EZ area range, 0.54–13.61 mm^2 measured by the

Table 2. Reliability and Agreement Results

Study	Area Analysis: ΔArea, mm² Mean ± SD	Area Analysis: RC, mm²	Area Analysis: ICC [95% CI]	Shape Analysis: ASD, mm² Mean ± SD	Shape Analysis: RC, mm²	Shape Analysis: ICC [95% CI]	Area ICC vs. Shape ICC *P* Value
Interrater[a]							
One expert							
1 & 2: profile	−1.32 ± 1.12	2.96	0.763 [0.050–0.927]	1.51 ± 1.04	3.61	0.660 [0.371–0.741]	0.8
1 & 2: en face	−0.52 ± 0.87	1.78	0.906 [0.760–0.954]	0.84 ± 0.68	2.12	0.780 [0.673–0.853]	0.2
1 & 3: profile	−1.29 ± 1.12	2.87	0.776 [0.046–0.931]	1.47 ± 1.07	3.57	0.608 [0.384–0.746]	0.8
1 & 3: en face	−0.34 ± 0.65	1.32	0.945 [0.897–0.971]	0.63 ± 0.49	1.57	0.841 [0.765–0.893]	0.04
Two experts							
2 & 3: profile	0.037 ± 0.26	0.56	0.986 [0.980–0.990]	0.33 ± 0.21	0.77	0.906 [0.868–0.933]	<0.001
2 & 3: en face	0.18 ± 0.65	1.09	0.959 [0.931–0.975]	0.58 ± 0.48	1.49	0.845 [0.772–0.896]	0.005
All raters							
1–3: profile	−0.86 ± 1.12	2.30	0.831 [0.338–0.935]	1.10 ± 1.03	3.11	0.672 [0.540–0.769]	0.5
1–3: en face	−0.23 ± 0.79	1.42	0.935 [0.890–0.962]	0.68 ± 0.56	1.92	0.817 [0.750–0.870]	0.02
Intrarater[b]							
Novice raters							
1: profile	−0.79 ± 0.92	2.17	0.932 [0.583–0.978]	1.19 ± 0.92	2.97	0.770 [0.602–0.872]	0.3
1: en face	−0.23 ± 0.81	1.33	0.965 [0.934–0.981]	0.67 ± 0.64	1.84	0.855 [0.744–0.920]	0.02
Expert raters							
2: profile	−0.085 ± 0.42	0.83	0.983 [0.967–0.991]	0.42 ± 0.32	1.05	0.900 [0.820–0.946]	0.007
2: en face	−0.26 ± 0.74	1.16	0.965 [0.929–0.982]	0.54 ± 0.61	1.61	0.863 [0.757–0.925]	0.04
3: profile	−0.42 ± 0.26	0.94	0.977 [0.957–0.995]	0.47 ± 0.26	1.06	0.884 [0.788–0.937]	0.02
3: en face	−0.081 ± 0.44	0.76	0.987 [0.976–0.993]	0.43 ± 0.33	1.08	0.906 [0.829–0.949]	0.002
2 & 3: profile	−0.25 ± 0.39	0.88	0.980 [0.962–0.994]	0.44 ± 0.29	1.05	0.891 [0.840–0.937]	0.01
2 & 3: en face	−0.17 ± 0.61	0.98	0.976 [0.961–0.985]	0.49 ± 0.49	1.36	0.885 [0.826–0.924]	<0.001
All raters							
1–3: profile	−0.43 ± 0.66	1.33	0.964 [0.844–0.985]	0.69 ± 0.68	1.90	0.850 [0.788–0.895]	0.1
1–3: en face	−0.19 ± 0.68	1.09	0.972 [0.958–0.981]	0.55 ± 0.54	1.52	0.873 [0.823–0.910]	<0.001
Intermethod[c]							
Novice raters							
1	−0.58 ± 1.32	2.55	–	1.32 ± 0.99	3.24	–	–
Expert raters							
2	0.17 ± 0.52	1.11	–	0.58 ± 0.37	1.36	–	–
3	0.36 ± 0.87	1.50	–	0.71 ± 0.69	1.95	–	–
2 & 3	0.23 ± 0.63	1.20	–	0.63 ± 0.44	1.51	–	–
All raters							
1–3	−0.09 ± 1.00	1.69	–	0.88 ± 0.76	2.28	–	–

Rater 1 is novice, raters 2 and 3 are experts.

[a] For rater pairs, ΔArea is the second listed rater's area minus the first's. For example, ΔArea for "1 & 2" is the area measured by rater 2 minus the area measured by rater 1.

[b] ΔArea is the retest (second round) area minus the test (first round) area.

[c] ΔArea is the en face method area minus the profile method area.

en face method) was a good representation of all measurable eyes (EZ area range, 0.52–13.61 mm^2), with no significant difference ($P = 0.5$) and small effect size (Cohen's $d = 0.15$). The mean test–retest ΔArea was negative for every rater, regardless of method, meaning that on average the second area measurement was smaller than the first. These differences were generally small and significantly different from zero only for raters 1 and 3 using the profile method ($P < 0.001$ for both). The intrarater ICCs and RCs for area and shape were generally as good or better than the corresponding interrater values, which supports maintaining the same raters for the entirety of a study.

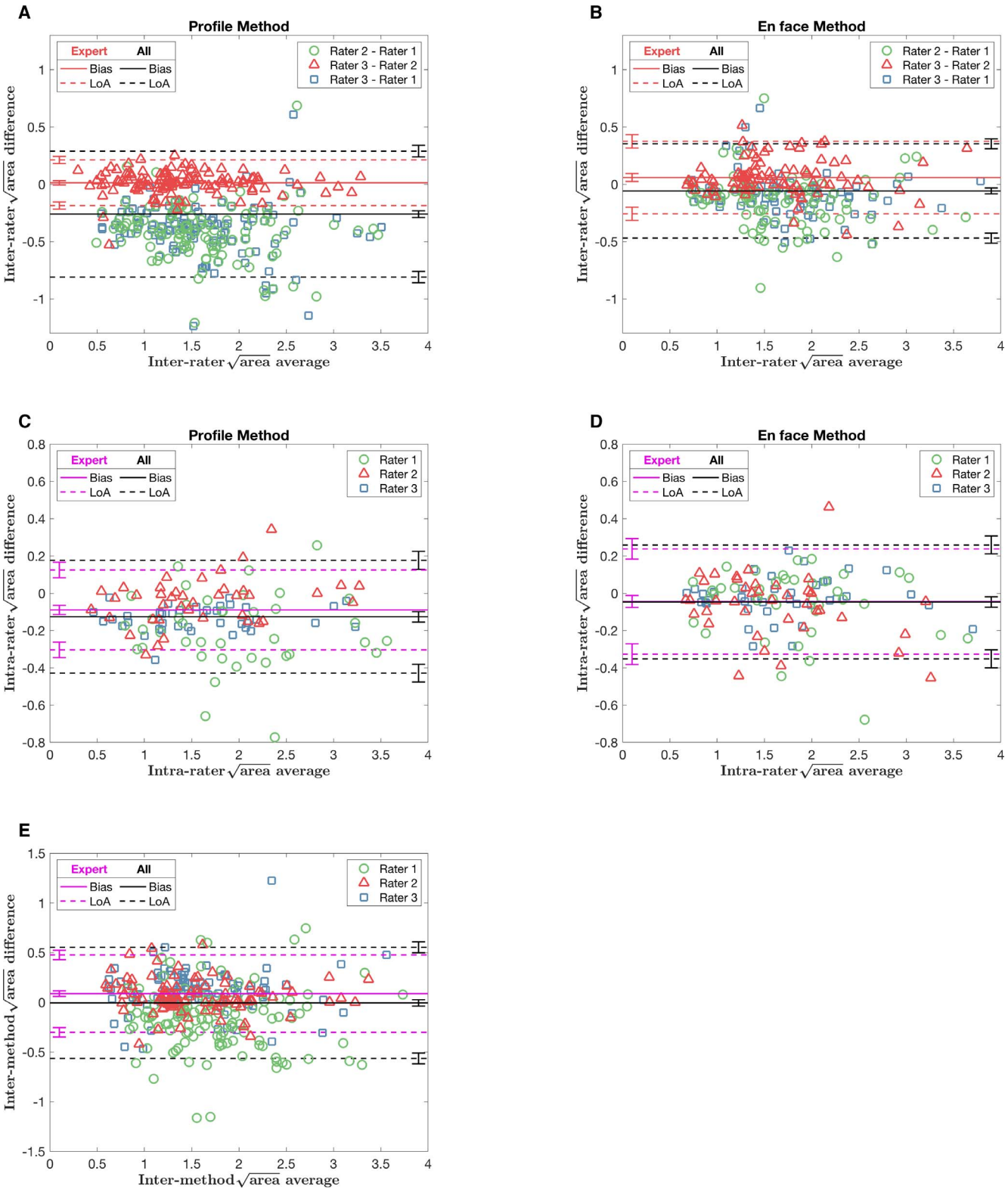

Figure 3. Bland-Altman plots for data after square root transformation. Plotted are interrater differences for the (A) profile method and (B) en face method, intrarater differences (second measurement minus first measurement) for the (C) profile method and (D) en face method, and (E) intermethod differences (en face measurement minus profile measurement). *Solid lines* show the mean difference (bias) and *dashed lines* show the mean ± 1.96 standard deviations (95% limits of agreement [LoA]), shown for all raters and for expert raters only. The 95% CIs for the bias and LoA are shown for expert raters on the *left side* and for all raters on the *right side* of each plot. Rater 1, novice; Rater 2, expert; Rater 3, expert.

Discussion

The underlying boundary measurements do not agree as well as a conventional analysis of EZ area might indicate. After accounting for morphologic measurement variability, EZ area reliability appears more likely in the approximate range of 0.80 to 0.90 as determined by shape analysis than the 0.95 to 0.99 range indicated by area analysis. Qualitatively categorizing ICC ranges of 0.75 to 1.00 and 0.60 to 0.74 as indicative of, respectively, excellent and good reliability,[29] shape ICCs still were excellent for the most part despite often being significantly smaller than their area ICC counterparts.

The profile method expert interrater shape RC was 0.77 mm^2, nearly half that of the en face method. Individual intrarater shape RCs for both methods ranged from 1.05 to 1.61 mm^2. These values are larger than previously reported mean progression rates of 0.27 mm^2/year,[10] 0.64 mm^2/year,[3] and 0.67 mm^2/year[15] from various types of RP cohorts, which suggests reliable detection of EZ area change with these methods may require follow-up times of more than 1 year.

Measurability, timing, and rater experience are important considerations when comparing delineation methods. Although twice as slow as the en face method, the profile method — accelerated by marking only the EZ termini instead of the full EZ band contour — was not prohibitively time-consuming as has been suggested,[10] and it identified the EZ boundary in roughly a third more eyes. En face method measurability could be improved with more sophisticated approaches to generating en face EZ contrast than a simple volumetric MIP. En face method reliability was largely independent of rater experience, whereas profile method reliability was often significantly better with expert raters. Ramachandran et al.,[11] studying EZ width reliability, also found that profile method reliability improved with rater experience.

En face method repeatability in RP cohorts has been previously investigated. The intrarater area RCs of 0.65 to 1.04 mm^2 described by Tee et al.[15] for a single rater are comparable to the 0.76 to 1.16 mm^2 expert rater area RCs reported here. With two raters analyzing a subset of the same VPA Trial data used here, Hariri et al.[10] observed an interrater area ICC of 0.996 (95% confidence interval [CI], 0.995–0.997), much better than the expert rater area ICC of 0.959 we found. Their study did not report checking for very large subject sample heterogeneity,[20,30] the type of ICC used, or confirmation of the underlying ANOVA assumptions — any of which could generate misleadingly high area ICC values. The lower ICC in our study could also be attributable to the larger sample size (89 vs. 45 eyes) or the inclusion of challenging eyes, as described below. As an experiment, when we selectively removed measurements, keeping only the 45 eyes with the best expert interrater agreement, the en face method area ICC improved to 0.997 (95% CI, 0.994–0.998).

There are several limitations to this work. As a retrospective study, our analysis pertains only to the rater measurements and does not capture the variability inherent to the OCT acquisition itself, which would require replicate scans to estimate, but is likely much smaller than the rater variability. Another consideration is the inclusion of only autosomal dominant RP patients, which makes the findings directly applicable to other RP and Usher syndrome populations, yet still relevant for any cohort with similar EZ boundary conspicuity. Only one junior rater was available, but two would have balanced the study and better represented novice raters. As a result of making every effort to include all trial data in this study, there were image quality realisms to contend with, such as cystoid macular edema, vascular shadowing, and disintegrating EZ reflectivity profiles, all of which added complexity during delineation that likely impacted reliability. One final consideration about the large sample size in our study: because raters worked for long periods with many scans waiting in their queues, the results may not reflect a different scenario in which raters intermittently analyze only a few scans.

To our knowledge, this is the first intrarater study with a 1-year test–retest interval — a duration chosen to approximate a trial follow-up schedule. For both methods, the second round of area measurements tended to be smaller than the first. Rater fatigue may have contributed to this bias. With the profile method, for example, smaller areas would result from overlooking entire b-scans near the EZ boundary along the slow-scan direction. Fatigue also could skew en face method measurements toward smaller boundaries, because they would be faster to delineate. Rater knowledge of EZ dynamics among RP patients possibly could have led to unintentionally smaller retest measurements, despite this not being a longitudinal study.

As with any potential endpoint, the reliability and agreement findings are relevant to clinical trial design

and outcome assessment. By capturing more EZ measurement variability, the shape-sensitive approach we described may provide more accurate estimates of EZ area repeatability. Both delineation methods generally showed excellent reliability, but for the slower profile method with better measurability, that excellence required expert raters whereas en face method reliability was more consistent across experience levels.

Acknowledgments

The authors thank the VPA Clinical Trial investigators for their contributions to the design, execution, and management of the VPA Trial. The VPA Clinical Trial investigators are: Paul S. Bernstein, MD, PhD; David G. Birch, PhD; Judith Chiostri, MS; Laura R. Erker, PhD; John R. Heckenlively, MD; Alessandro Iannaccone, MD, MS; Byron L. Lam, MD; Jennifer B. McCormack, MS; Mark E. Pennesi, MD, PhD; and David J. Wilson, MD.

Supported by grant P30 EY010572 from the National Institutes of Health (Bethesda, MD) and by unrestricted departmental funding from Research to Prevent Blindness (New York, NY). The VPA Clinical Trial was co-sponsored by the Foundation Fighting Blindness Clinical Research Institute and the Department of Defense (W81XWH-09-2-0189).

Disclosure: **T.B. Smith**, Applied Genetic Technologies Corporation (C); **M.A. Parker**, None; **P.N. Steinkamp**, None; **A. Romo**, None; **L.R. Erker**, None; **B.J. Lujan**, Genentech/Roche (C), BioTime/Cell Cure (C), Carl Zeiss Meditec (R) P; **N. Smith**, None

References

1. Spaide RF, Curcio CA. Anatomical correlates to the bands seen in the outer retina by optical coherence tomography: literature review and model. *Retina*. 2011;31:1609–1619.
2. Staurenghi G, Sadda S, Chakravarthy U, Spaide RF, Panel INfOCT. Proposed lexicon for anatomic landmarks in normal posterior segment spectral-domain optical coherence tomography: the IN*OCT consensus. *Ophthalmology*. 2014; 121:1572–1578.
3. Ramachandran R, Zhou L, Locke KG, Birch DG, Hood DC. A comparison of methods for tracking progression in X-linked retinitis pigmentosa using frequency domain OCT. *Transl Vis Sci Technol*. 2013;2:5.
4. Cai CX, Locke KG, Ramachandran R, Birch DG, Hood DC. A comparison of progressive loss of the ellipsoid zone (EZ) band in autosomal dominant and x-linked retinitis pigmentosa. *Invest Ophthalmol Vis Sci*. 2014;55:7417–7422.
5. Rangaswamy NV, Patel HM, Locke KG, Hood DC, Birch DG. A comparison of visual field sensitivity to photoreceptor thickness in retinitis pigmentosa. *Invest Ophthalmol Vis Sci*. 2010;51: 4213–4219.
6. Birch DG, Locke KG, Wen Y, Locke KI, Hoffman DR, Hood DC. Spectral-domain optical coherence tomography measures of outer segment layer progression in patients with X-linked retinitis pigmentosa. *JAMA Ophthalmol*. 2013;131:1143–1150.
7. Birch DG, Locke KG, Felius J, et al. Rates of decline in regions of the visual field defined by frequency-domain optical coherence tomography in patients with RPGR-mediated X-linked retinitis pigmentosa. *Ophthalmology*. 2015;122:833–839.
8. Smith TB, Parker M, Steinkamp PN, et al. Structure-function modeling of optical coherence tomography and standard automated perimetry in the retina of patients with autosomal dominant retinitis pigmentosa. *PLoS One*. 2016; 11:e0148022.
9. Hariri AH, Velaga SB, Girach A, et al. Measurement and reproducibility of preserved ellipsoid zone area and preserved retinal pigment epithelium area in eyes with choroideremia. *Am J Ophthalmol*. 2017;179:110–117.
10. Hariri AH, Zhang HY, Ho A, et al. Quantification of ellipsoid zone changes in retinitis pigmentosa using en face spectral domain-optical coherence tomography. *JAMA Ophthalmol*. 2016;134:628–635.
11. Ramachandran R, Cai CX, Lee D, et al. Reliability of a manual procedure for marking the EZ endpoint location in patients with retinitis pigmentosa. *Transl Vis Sci Technol*. 2016;5:6.
12. Strampe MR, Huckenpahler AL, Higgins BP, et al. Intraobserver repeatability and interobserver reproducibility of ellipsoid zone measurements in retinitis pigmentosa. *Transl Vis Sci Technol*. 2018; 7:13–13.
13. ClinicalTrials.gov. Trial of oral valproic acid for retinitis pigmentosa. NCT01233609. Available at: https://clinicaltrials.gov/ct2/show/NCT01233609; Accessed September 7, 2018.

14. Birch DG, Bernstein PS, Iannacone A, et al. Effect of oral valproic acid vs placebo for vision loss in patients with autosomal dominant retinitis pigmentosa: a randomized phase 2 multicenter placebo-controlled clinical trial. *JAMA Ophthalmol*. 2018;136:849–856.
15. Tee JJL, Yang Y, Kalitzeos A, Webster A, Bainbridge J, Michaelides M. Natural history study of retinal structure, progression, and symmetry using ellipzoid zone metrics in RPGR-associated retinopathy. *Am J Ophthalmol*. 2019;198:111–123.
16. Alt H, Fuchs U, Rote G, Weber G. Matching convex shapes with respect to the symmetric difference. *Algorithmica*. 1998;21:89–103.
17. Veltkamp RC, Hagedoorn M. State of the art in shape matching. *Prin Vis Information Ret*. 2001: 87–119.
18. Smith TB, Smith N. Agreement and reliability statistics for shapes. *PLoS One*. 2018;13: e0202087.
19. Kottner J, Audige L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J Clin Epidemiol*. 2011; 64:96–106.
20. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. *Ultrasound Obstet Gynecol*. 2008;31:466–475.
21. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979; 86:420–428.
22. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods*. 1996;1:30–46.
23. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Comput Biol Med*. 1990;20:337–340.
24. Barnhart HX, Haber MJ, Lin LI. An overview on assessing agreement with continuous measurements. *J Biopharm Stat*. 2007;17:529–569.
25. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1:307–310.
26. Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek AL. Smallest real difference, a link between reproducibility and responsiveness. *Qual Life Res*. 2001;10:571–578.
27. Oehlert GW. A note on the delta method. *Am Stat*. 1992;46:27–29.
28. Donner A, Zou G. Testing the equality of dependent intraclass correlation coefficients. *J Royal Stat Soc Series D*. 2002;51:367–379.
29. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess*. 1994;6:284–290.
30. Raunig DL, McShane LM, Pennello G, et al. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. *Stat Methods Med Res*. 2015;24:27–67.