Reliability of a Manual Procedure for Marking the EZ Endpoint Location in Patients with Retinitis Pigmentosa

Rithambara Ramachandran¹, Cindy X. Cai¹, Dongwon Lee¹, Benjamin C. Epstein¹, Kirsten G. Locke², David G. Birch²,³, and Donald C. Hood¹,⁴

¹ Department of Psychology, Columbia University, New York, NY, USA
² Retina Foundation of Southwest, Dallas, TX, USA
³ Department of Ophthalmology, University of Texas Southwestern Medical Center, Dallas, TX, USA
⁴ Department of Ophthalmology, Columbia University, New York, NY, USA

Correspondence: Donald C. Hood, Department of Psychology, 406 Schermerhorn Hall, 1190 Amsterdam Avenue, MC 5501, Columbia University, New York, NY 10027 USA; e-mail: dch3@columbia.edu

Purpose: We developed and evaluated a training procedure for marking the endpoints of the ellipsoid zone (EZ), also known as the inner segment/outer segment (IS/OS) border, on frequency domain optical coherence tomography (fdOCT) scans from patients with retinitis pigmentosa (RP).

Methods: A manual for marking EZ endpoints was developed and used to train 2 inexperienced graders. After training, an experienced grader and the 2 trained graders marked the endpoints on fdOCT horizontal line scans through the macula from 45 patients with RP. They marked the endpoints on these same scans again 1 month later.

Results: Intragrader agreement was excellent. The intraclass correlation coefficient (ICC) was 0.99, the average difference of endpoint locations (19.6 μm) was close to 0 μm, and the 95% limits were between −284 and 323 μm, approximately ±1.1°. Intergrader agreement also was excellent. The ICC values were 0.98 (time 1) and 0.97 (time 2), the average difference among graders was close to zero, and the 95% limits of these differences was less than 350 μm, approximately 1.2°, for both test times.

Conclusions: While automated algorithms are becoming increasingly accurate, EZ endpoints still have to be verified manually and corrected when necessary. With training, the inter- and intragrader agreement of manually marked endpoints is excellent.

Translational Relevance: For clinical studies, the EZ endpoints can be marked by hand if a training procedure, including a manual, is used. The endpoint confidence intervals, well under ±2.0°, are considerably smaller than the 6° spacing for the typically used static visual field.

Introduction

Retinitis pigmentosa (RP) is a group of heterogeneous inherited retinal disorders characterized by the degeneration of rod and cone photoreceptor cells and, in the most extreme cases, the retinal pigment epithelium (RPE). Patients first experience night blindness followed by midperipheral vision loss. Progressive constriction of the visual field eventually leads to central vision loss and, in some cases, complete blindness. While the time-scale may vary, progressive constriction of the useable field of vision is seen in all genetic forms of RP.¹⁻⁶

Traditionally, full-field electroretinogram (ERG) cone flicker, kinetic perimetry, static perimetry, and more recently multifocal (mf) ERG have been used to monitor disease progression. However, natural progression is slow relative to the variability inherent in these measures, making changes difficult to detect using conventional metrics. With the emergence of numerous new treatment strategies for RP, it is more important than ever to find ways to monitor progression that are robust but sensitive to change, relatively easy to administer, and widely available.

Optical coherence tomography (OCT) offers one possibility. The introduction of frequency domain (fd) OCT has now made it possible to visualize more clearly individual retinal layers affected by RP to follow disease progression.⁷⁻⁹ In particular, research has focused on the ellipsoid zone (EZ), also called the
inner segment/outer segment (IS/OS) border. This is a hyperreflective band clearly visible on OCT scans. While for some this signal is thought to come from the cilium connecting photoreceptor inner segment and outer segments, others argue that it is due to light scattered by the mitochondria in the ellipsoids of the distal inner segment. Regardless, disruption of this reflective EZ is a clinical marker of disease pathology.

The EZ disappears in the periphery early in the RP disease process. Rangaswamy et al., however, showed that for the EZ to disappear, the local visual field (VF) loss had to exceed approximately 8 dB. On the other hand, the edge of the EZ corresponds to the edge of the useable VF; that is, it corresponds to the precipitous drop in sensitivity seen on VFs of patients with some preservation of central visual sensitivity.

Birch et al., using an experienced grader to measure the distance between endpoints (EPs; EZ width) on two scans obtained on the same day, found that 95% of all test–retest differences for EZ width were less than 0.43°, far better than the standard VF test, which has test points spaced by 6°. Further, Birch et al. looked at the VF sensitivities inside and outside the EZ EPs and found that the region surrounding this edge is more sensitive in detecting progression compared to global sensitivity measures. In general, the evidence suggests that the EZ EP is a more sensitive measure than existing VF and ERG methods.

Following the edge of the EZ has distinct advantages over other measures. First, compared to conventional measures of VF and ERG, it is easier to administer and analyze. Second, it also has advantages over other measures of OCT scans. Ramachandran et al. showed that following the ends of the EZ on horizontal and vertical line scans taken through the fovea was as sensitive, if not more sensitive, in detecting annual changes compared to other metrics derived from a full macular cube scan including outer nuclear layer (ONL), outer segment (OS), and RPE volume. Thus, only one or two OCT line scans must be added to routine clinical protocols.

It remains open how best to standardize the marking of the edge of the EZ. For the EZ edge points (EPs) to be a viable outcome measure in RP clinical trials, inter- and intragraders agreement must be good. To this end, we developed a training procedure based upon a written manual and a training protocol. Here, we test this procedure by training two inexperienced “graders” to mark the EPs on a set of horizontal line scans. After training, intra- and intergrader agreement was assessed on a new set of horizontal scans from 45 patients with RP. Finally, we compared the results of the manually marked EPs to EP markings on the same set of scans 1 month later.

**Methods**

**Subjects**

**Graders**

Three graders participated in this study. Grader A, the first author, who had more than five years of segmentation experience, trained graders B and C, two inexperienced undergraduate students. Neither student had any experience with outer retinal segmentation outside of the training and test set grading for this study.

**Patients**

A total of 75 patients with RP of mixed genotypes was randomly selected from a larger database of patients with RP available at the Retina Foundation of the Southwest. Of these patients, 30 were included in the training sets and the remaining 45 (age, 35.4 ± 18.9 years; range, 8–72 years) were included in the study sets (see Fig. 1A). The 45 patients (15 autosomal dominant, 15 X-linked, and 15 autosomal recessive) in the study set had an average mean deviation on the 30-2 visual field of −22.5 dB (range, −2.2 to −33.6 dB), and their logMAR ranged from 0 to 0.8. The scans that were selected had an EZ that was clearly visible and had EPs within the 30° field. If more than one scan was available for a patient, the most recent scan that met the inclusion criteria was chosen. All procedures adhered to the Declaration of Helsinki and were approved by the Institutional Review Board at the University of Texas Southwestern Medical Center. Informed consent was obtained from all participants.

**Frequency Domain OCT Scans**

Horizontal fOCT line scans (Spectralis HRA-OCT; Heidelberg Engineering, Heidelberg, Germany) along the horizontal meridian were obtained from one eye per patient. Scans were approximately 9 mm in length. For the training session, there were three bands of interest (Fig. 1B): (1) outer limiting membrane (OLM), (2) EZ band, and (3) proximal border of the retinal pigment epithelium (pRPE). The EZ EP was defined as the point where the EZ converged with the pRPE (orange arrow in Fig. 1B). For each scan, the EP on the nasal and temporal sides of the fovea was marked. The fovea was marked by
Grader A and the same foveal point was used for graders B and C to make sure the reference for calculating the EP was the same for all scans.

**Study Design**

**The Manual**

Based upon previous experience segmenting fdOCT scans from patients with outer retinal disease, a training manual was written detailing how to mark the EP locations of the EZ (see Supplementary Fig. S1). The manual provided instructions on how to identify and segment the OLM, EZ band, and pRPE and illustrated examples of commonly encountered ambiguities.

In brief, the manual instructs the user to: (1) Step 1 – Decide the boundary lines for 3 bands: OLM, EZ, and the pRPE. These lines do not have to be actually drawn. (2) Step 2 – Mark the location where the EZ merges with the pRPE. This is the EP. (3) Step 3 – If the ending of the EZ is ambiguous, segment the OLM, EZ, and the pRPE based on mental outlines (Step 1). (4) Step 4 – If EZ EPs are still ambiguous, refer to the ‘Marking the EZ Edge’ flowchart provided on the last page of the manual.

**Training**

With the help of the written manual, Grader A conducted a training session for the two inexperienced graders (B and C) using the scans from the 30 patients in training set #1. To help familiarize the new graders with the anatomy of the outer retina and to detect obvious systemic errors, during training session #1, Grader A first segmented the three borders of interest (OLM, EZ, pRPE) on a healthy control and an RP patient, while graders B and C watched. Graders B and C then segmented the three borders on the remaining 13 RP patients’ scans, which covered a range of disease...
severity. The three graders discussed and compared segmentations at the end of this training session. The next day, grader A conducted training session #2, which discussed ambiguities in marking the EP. This time, the three graders were not required to do the full segmentation. Instead, they independently marked just the EP on the 15 patients in training set #2. The results were quantitatively analyzed and differences among the 3 graders were discussed.

**Test Set Time 1**

Shortly after the training session (1–5 days), the 3 graders independently marked the EPs from the 45 RP patients in the study set. All three graders had access to the manual, but were not able to consult with each other. No feedback was given at the end of this set.

**Test Set Time 2**

At 1 month after test set time 1, the same 45 scans were presented in a random order. The 3 graders again independently marked the EPs on these scans.

**Data Analysis**

To assess the degree of inter- and intragrader agreement, the signed and absolute differences of the following two measures were obtained: (1) $\Delta E_{\text{P intra}}$ – the difference in marked EP between time 1 ($E_{\text{P time1}}$) and time 2 ($E_{\text{P time2}}$), and (2) $\Delta E_{\text{P inter}}$ – the difference in marked EP between a grader ($E_{\text{P grader}}$) and the average EP marked by the other two graders ($E_{\text{P others}}$).

In addition, intraclass correlation coefficients (ICC) and the 95% limits of agreement using Bland-Altman plots were calculated.

**Results**

Data from the scans of 45 eyes with RP were analyzed to assess the reproducibility of marking the EZ edge points (EPs). Figure 2A shows a representative scan with the EPs marked by three graders at time 1 (solid lines in Fig. 2A) and then 1 month later.
at time 2 (dashed lines in Fig. 2B). The enlarged insets are from the same location at both time points. The experienced grader’s markings are shown in red (Fig. 2A), while the inexperienced graders’ markings are in blue and green (Figs. 2B, 2C, respectively). To put the deviations in perspective, the scale bar in the lower right of the Figure is 100 um (approximately 0.35°). Although the marked EPs are similar, there is variability within a single grader (intragrader) and between graders (intergrader).

**Intragrader Agreement**

In Figure 3A, the EPs measured for all 45 eyes at time 1 were plotted against the same locations at time 2 for each grader. The data points cluster around the black dashed ($y = x$) line, indicating that the EPs are
nearly the same at both time points. The best-fit lines are shown in color.

To obtain a quantitative measure of intragrader agreement for each grader, the mean signed and unsigned difference between EPs marked at each time point \( (\Delta EP_{\text{intra}}) \) was measured. These differences along with standard deviations (SDs) are provided in Table 1. The average signed and absolute mean \( \Delta EP_{\text{intra}} \) for all three graders was 19.6 and 80.8 \( \mu \)m, respectively. Pairwise comparisons of marked EPs between times 1 and 2 were not significant \( (P = 0.37) \) and the overall ICC value was 0.98, indicating excellent agreement. On the other hand, although the signed intragrader differences among the three graders were not significantly different (ANOVA; \( P = 0.229 \)), a post hoc test (Tukey HSD test) indicated a significant difference between the experienced (A) and the inexperienced (B, \( P < 0.01 \); C, \( P < 0.05 \)) graders. There was no significant difference between the two inexperienced graders. (Note that inexperienced graders were not exposed to OCT scans from patients with RP between the two testing periods so additional [implicit] training did not take place.)

Further, there was good agreement between EPs at the two time points as demonstrated by the Bland-Altman plot in Figure 3B and the coefficients of repeatability (CR) in Table 1. The average difference (19.6 \( \mu \)m, solid black line) was close to 0 \( \mu \)m (solid red line) and the 95% limits of agreement for \( \Delta EP_{\text{intra}} \) were between \(-284 \) and \(323 \) \( \mu \)m (dashed back lines), approximately \( \pm 1.1^\circ \).

### Intergrader Agreement

To obtain an analogous measure of intergrader agreement, the EP marked by one grader (\( EP_{\text{grader}} \)) was compared to the average EP marked by the other two graders (\( EP_{\text{others}} \)). Figure 4 shows the results at

### Table 1. Intragrader Differences

<table>
<thead>
<tr>
<th>Grader</th>
<th>Mean (SD)</th>
<th>( \Delta EP_{\text{intra}} )</th>
<th>( \Delta EP_{\text{intra}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>signed ( \mu )m</td>
<td>absolute ( \mu )m</td>
<td>CR ( \mu )m</td>
</tr>
<tr>
<td>A</td>
<td>3.74 (82.7)</td>
<td>42.3 (71.0)</td>
<td>270</td>
</tr>
<tr>
<td>B</td>
<td>13.2 (221)</td>
<td>108 (193)</td>
<td>611</td>
</tr>
<tr>
<td>C</td>
<td>41.8 (127)</td>
<td>91.8 (96.5)</td>
<td>368</td>
</tr>
<tr>
<td>Average</td>
<td>19.6 (155)</td>
<td>80.8 (133)</td>
<td>432</td>
</tr>
</tbody>
</table>

\( \Delta EP_{\text{intra}} \): measure of intragrader agreement; SD, standard deviation; CR, coefficient of repeatability.

Figure 4. (A) Scatterplot of EP markings of each grader (\( EP_{\text{grader}} \)) and the average marking of the other two graders (\( EP_{\text{others}} \)) at time 1. (B) Scatterplot of EP markings of each grader (\( EP_{\text{grader}} \)) and the average marking of the other two graders (\( EP_{\text{others}} \)) at time 2.
times 1 (Fig. 4A) and 2 (Fig. 4B). If the grader’s location equaled the average of the other two graders, then the points would fall along the black dashed line, which has a slope of 1.0. The data fall close to this line and show a strong correlation with $R^2 = 0.985$ and $R^2 = 0.991$ at times 1 and 2, respectively. Intraclass correlation coefficient values ranged from 0.97 to 1.00.

Plots analogous to Figure 3B are given in Figure 5, where the signed differences between $EP_{\text{grader}}$ and $EP_{\text{others}}$ ($\Delta EP_{\text{inter}}$) have been plotted against the average $EP$ marking of the three graders ($EP_{\text{all}}$). An $\Delta EP_{\text{inter}}$ of 0 μm (EP marked by grader is equal to the mean EP marked by the other two graders) would lie at $y = 0$ (red line). The average $\Delta EP_{\text{inter}}$ (time 1, $-1$ μm; time 2, 7 μm) is shown as the solid black line and the 95% limits of agreement are the dashed lines at $-295$ and $293$ μm (time 1) and at $-328$ and $342$ μm (time 2).

Figures 5A and 5B indicate that on average there was no consistent difference between any one grader’s EPs and the remaining two graders’ EPs when signed differences were compared. The average absolute difference ($\Delta EP_{\text{inter}}^{\text{absolute}}$) was $101$ (time 1) and $96$ (time 2) μm. The signed and absolute differences between these two markings ($\Delta EP_{\text{inter}}$) for each grader...
Table 2. Intergrader Differences (Time 1)

<table>
<thead>
<tr>
<th>Grader</th>
<th>(\Delta EP_{inter}) signed, (\mu m)</th>
<th>(\Delta EP_{inter}) absolute, (\mu m)</th>
<th>CR, (\mu m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-14.2 (144)</td>
<td>95.0 (108)</td>
<td>398</td>
</tr>
<tr>
<td>B</td>
<td>-60.4 (144)</td>
<td>99.3 (121)</td>
<td>431</td>
</tr>
<tr>
<td>C</td>
<td>70.5 (131)</td>
<td>110 (101)</td>
<td>411</td>
</tr>
<tr>
<td>Average</td>
<td>-1.4 (150)</td>
<td>101 (110)</td>
<td>414</td>
</tr>
</tbody>
</table>

Table 3. Intergrader Differences (Time 2)

<table>
<thead>
<tr>
<th>Grader</th>
<th>(\Delta EP_{inter}) signed, (\mu m)</th>
<th>(\Delta EP_{inter}) absolute, (\mu m)</th>
<th>CR, (\mu m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>22.3 (159)</td>
<td>89.5 (133)</td>
<td>470</td>
</tr>
<tr>
<td>B</td>
<td>-50.8 (168)</td>
<td>96.7 (146)</td>
<td>523</td>
</tr>
<tr>
<td>C</td>
<td>49.0 (171)</td>
<td>101 (145)</td>
<td>482</td>
</tr>
<tr>
<td>Average</td>
<td>6.83 (171)</td>
<td>95.9 (141)</td>
<td>493</td>
</tr>
</tbody>
</table>

Discussion

The purpose of this study was to use inexperienced graders to see whether, with proper training, we could achieve good reliability. We evaluated a training procedure, based upon a manual and a training protocol, for identifying the EZ EPs. The results indicated that, with our 2-tier training system and the guidance of a written manual, it is possible to train inexperienced graders to reliably identify the edge of the EZ in a relatively short period of time. Inter- and intragrader measurements were similar 1 month after training. The average difference in EP markings made by one grader compared to the other two graders was less than 2 \(\mu m\) at time 1 and less than 7 \(\mu m\) 1 month later, and the confidence intervals were less than 335 \(\mu m\) (approximately 1.2\(^{\circ}\)). Likewise, the average intragrader difference in markings was typically less than 20 \(\mu m\) (<0.1\(^{\circ}\)) and the 95% confidence intervals were approximately 304 \(\mu m\) (approximately 1.1\(^{\circ}\)). This is not to say that further training or more experience with OCT scans will not produce superior results. As noted above, the intragrader agreement of the experienced grader was better than that of the inexperienced graders, suggesting that precision may increase with experience.

In this context, it is appropriate to ask how experience may affect the ability to detect progression. Birch et al.\(^{19}\) reported a reduction in EZ width in patients with X-linked (x) RP of 248 \(\mu m\) per year. Under optimal conditions with a single experienced grader, this progression rate was higher than the 95% confidence interval for EZ width test–retest differences. Thus, most patients showed significant change over 1 year. With inexperienced graders in this study, the 95% confidence interval for marking the EZ EP was comparable with the rate of progression for one EP,\(^{19}\) thus reducing somewhat the sensitivity to significant progression. This estimate of variability includes, however, grader B, who by far had the weakest reliability. The confidence interval for test–retest differences of Graders A and C, on the other hand, fall well within the average rate of yearly progression.

Several aspects of our training procedure contributed to its relative success. First, in developing our manual, we reviewed OCT scans not used in the study to anticipate sources of variability. Second, we noted patterns of EZ band dropout and made use of typical RPE characteristics, such as localized reflection spots or thickening of the RPE, to decrease inter- and intragrader variability. Third, we also encouraged segmenting the full RPE and EZ band in particularly difficult cases.

Our approach has the advantage of requiring clinicians or graders, on most scans, to simply mark two points – a nasal and temporal EP – rather than to segment the entire line. In the future, automated algorithms may be able to mark these two points. However, it is likely that they will have to be checked carefully and corrected manually especially in the case of patients with severe damage where our experience indicates the algorithms have the most difficulty. To investigate the effects of disease severity with our procedure, we compared \(\Delta EP_{inter}\) absolute and \(\Delta EP_{inter}\) absolute in scans where the EZ edge fell inside the parafovea (central 2500 \(\mu m\)) to those where the EZ edge falls outside this region. Neither the \(\Delta EP_{inter}\) absolute nor the \(\Delta EP_{inter}\) absolute were significantly different between these two groups (2 sample \(t\)-test; \(\Delta EP_{inter}\) absolute, \(P = 0.061\); \(\Delta EP_{inter}\) absolute, time 1: \(P = 0.160\) and time 2: \(P = 0.582\)). Our results are provided along with SDs and CRs in Table 2 (time 1) and Table 3 (time 2). While the average signed difference is slightly better when compared to intragrader differences, absolute values suggest the opposite. In fact, the confidence intervals are similar for intra- and intergrader agreement, all falling within ± 340 \(\mu m\), or ±1.2\(^{\circ}\).
validate that in most RP patients, regardless of severity, our manually marked EP technique produces good agreement between and among graders.

Limitations and Caveats

There are limitations to the use of EZ EPs for measuring progression in RP. First, it is difficult to capture high quality OCT images from the midline for patients with poor fixation. Second, the method currently is limited to patients with a discernible EZ band within the central 30° and so precludes tracking patients with early stages or with very advanced stages of disease. This may change as wide-field OCT capabilities evolve. Third, the EZ band is not lost until visual field sensitivity has decreased by approximately 8 dB. Fourth, the method does not distinguish between selective loss of rods or cones as the loss of the EZ depends upon the loss of both.

Clinical Relevance

Previous work showed that the width of the EZ band, measured on high-quality midline OCT scans, is a sensitive, quick, and reliable method for measuring progression in patients with RP. In this study, we developed a short manual to train inexperienced graders to identify the edges of the EZ band. Our results have yielded excellent agreement between and among individual graders. In the future, this technique can be used to minimize variability within and among OCT segmentation centers participating in long-term RP clinical trials. Finally, it should be noted that the confidence intervals for inter- and intragrader agreement fell well within the 6° spacing of a 30-2 visual field, the most commonly used measure of progression of RP.

Conclusion

When human graders are trained, the intra- and intergrader agreement of the EZ band edges is very good.

Acknowledgments

Supported by National Institute of Health Grant EY-09076 (DCH, DGB) and Foundation Fighting Blindness (DGB).

Disclosure: R. Ramachandran, None; C.X. Cai, None; D. Lee, None; B.C. Epstein, None; K.G. Locke, None; D.G. Birch, None; D.C. Hood, None

References