Evaluation of Automated Segmentation Algorithms for Optic Nerve Head Structures in Optical Coherence Tomography Images

Xiangyun J. Duan, Joan L. Jefferys, and Harry A. Quigley

The Glaucoma Center of Excellence, Wilmer Ophthalmological Institute, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States

Correspondence: Harry A. Quigley, Wilmer 122, Johns Hopkins Hospital, Baltimore, MD 21287, USA; hquigley@jhmi.edu.
Submitted: April 2, 2018
Accepted: July 2, 2018
Citation: Duan XJ, Jefferys JL, Quigley HA. Evaluation of automated segmentation algorithms for optic nerve head structures in optical coherence tomography images. Invest Ophthalmol Vis Sci. 2018;59:3816–3826. https://doi.org/10.1167/iovs.18-24469

PURPOSE. To compare the identification of optic nerve head (ONH) structures in optical coherence tomography images by observers and automated algorithms.

METHODS. ONH images in 24 radial scan sets by optical coherence tomography were obtained in 51 eyes of 29 glaucoma patients and suspects. Masked intraobserver and interobserver comparisons were made of marked endpoints of Bruch’s membrane opening (BMO) and the anterior lamina cribrosa (LC). BMO and LC positional markings were compared between observer and automated algorithm. Repeated analysis on 20 eyes by the algorithm was compared. Regional ONH data were derived from the algorithms.

RESULTS. Intraobserver difference in BMO width was not significantly different from zero (P = 0.52) and the difference in LC position was less than 1% different (P = 0.04). Interobserver were slightly larger than intraobserver differences, but interobserver BMO width difference was 0.56% (P = 0.65). Mean interobserver difference in LC position was 14.74 µm (P = 0.004), 3% of the typical anterior lamina depth (ALD). Between observer and algorithm, BMO width differed by 1.85% (P = 0.23) and mean LC position was not significantly different (3.77 µm, P = 0.77). Repeat algorithmic analysis had a mean difference in BMO area of 0.58% (P = 0.47) and mean ALD difference of 0.54 ± 0.72%. Regional ALD had greater variability in the horizontal ONH regions. Some individual outlier images were not validly marked by either observers or algorithm.

CONCLUSIONS. Automated identification of ONH structures is comparable to observer markings for BMO and anterior LC position, making BMO a practical reference plane for algorithmic analysis.

Keywords: glaucoma, imaging, optic disc, optic nerve head, lamina cribrosa, optical coherence tomography

Current clinical management of glaucoma seeks to detect and to assess the progression of damage using both structural measurements of the optic nerve head (ONH) and nerve fiber layer and functional testing of the visual field sensitivity. Imaging of the ONH using optical coherence tomography (OCT) has become the dominant technique for structural assessment of glaucoma damage, due to its ability to quantify specific features of the retinal ganglion cells, their cell bodies and axons, and the associated tissues that support them, including connective tissues and vasculature.¹ ²

The biomechanical response of ONH tissue is recognized as an important risk factor in glaucoma damage.³–⁵ OCT imaging of this response has sequentially improved with better resolution of the position of the surface of the ONH and the anterior border of the lamina cribrosa (LC). In OCT images taken at two levels of IOP, the anterior LC moves relative to reference positions, such as Bruch’s membrane opening (BMO) or the choroidal—scleral interface (CSI).³–⁵ More rapid image acquisition in both spectral-domain and swept-source OCT allows multiple images to be acquired efficiently at each clock hour of the ONH. However, investigations that compare change within OCT images have been limited by the need to mark anatomic landmarks manually, including LC position. Doing so for large numbers of images is prohibitively time consuming and prior reports have often used only one or a small number of marked locations to indicate changes in anterior lamina depth (ALD). Both research and clinical applications of the study of ALD and its changes would benefit from accurate, automated identification of the features of interest.

Automatic identification of BMO and anterior lamina cribrosa depth (ALD) was recently developed for the Heidelberg Spectralis OCT instrument and is available for research purposes in scientific software (Heidelberg Engineering, Heidelberg, Germany). The software algorithms implemented to assess image data could eliminate a bottleneck in clinical glaucoma research, but their accuracy as compared to observer identification of the same anatomic features had not been independently tested. This investigation evaluates the performance of the automated algorithms for segmenting BMO and LC position compared to two observers who marked structural positions manually in normal and glaucoma ONHs.

METHODS

Images of adult POAG and glaucoma suspect patients of the Wilmer Glaucoma Center of Excellence were taken during
Table 1. Demographic Data of Patients

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total eyes</td>
<td>51</td>
</tr>
<tr>
<td>Total patients</td>
<td>29</td>
</tr>
<tr>
<td>Patient age (mean ± SD)</td>
<td>64.9 ± 10.7</td>
</tr>
<tr>
<td>Sex, by patient</td>
<td>13 M, 16 F</td>
</tr>
<tr>
<td>Sex, by eye</td>
<td>20 M, 51 F</td>
</tr>
<tr>
<td>Race</td>
<td>12 AA, 14 W, 3 A</td>
</tr>
<tr>
<td>K readings, by eye (mean ± SD)</td>
<td>7.77 ± 0.35</td>
</tr>
<tr>
<td>Axial length, by eye (mean ± SD)</td>
<td>24.36 ± 1.62</td>
</tr>
</tbody>
</table>

AA, African derived; W, European derived; A, Asian derived; M, male; F, female.

Routine office visits. The study was approved and monitored by the institutional review board of the Johns Hopkins School of Medicine, written informed consent was obtained, and the study abided by the Declaration of Helsinki. Glaucoma was defined by a standard system, including an abnormal glaucoma hemifield test with three test points abnormal in one hemisphere at a P < 5% level (HFA2i; Zeiss Meditec, Dublin, CA, USA), and no other clinical finding that could explain these features. Demographic data on the included eyes is in Table 1. The range of ages among persons studied was 49 to 87 years.

Prior to OCT imaging, axial length and keratometry measurements were taken with the IOL Master instrument (Carl Zeiss Meditec AG, Jena, Germany). Keratometry readings were entered into an integrated patient database software package (Heidelberg Eye Explorer; Heidelberg Engineering) to estimate optical magnification and to present data in micrometers with appropriate scaling.

OCT images were taken with the Heidelberg Spectralis from one or both eyes of each patient. For each patient, optic disc images were obtained from 24 high-resolution radial sectors, therefore comprising images every 15 minutes of the clock (7.5° between scans; Fig. 1). At each location, 25 B-scans were averaged, with 768 A-scans per B-scan. The imaging rate was 40,000 A-scans per second. Images were acquired in enhanced-depth imaging mode and by using a software-based anatomic positioning algorithm, which aligns the 9 o'clock temporal scan to the line spanning from the BMO center to the fovea. Up to 3 sets of ONH images were acquired consecutively from each patient to determine reproducibility.

OCT scans were taken in 51 eyes from 29 patients, and statistical methods were used to account for interocular correlation in analyses that included both eyes of a subject. During image grading, the investigators were masked to all identifiers and information about the subject. We imaged typical patients with glaucoma or suspicion of glaucoma. As such, there were persons with physical frailties or unusual optic disc structure that were not able to undergo satisfactory imaging. Images were inspected for quality by the study coordinator immediately after acquisition and automated segmentations of ONH anatomic structures were computed by the Heidelberg Spectralis SPX research software (VWM version 6.4.8.116). Poor quality was determined by examining all 24 scans and ensuring that the image was centered on the ONH and included the LC and retinal tissue on both sides of the B scan. In images where the imaging range was too superficial and there was insufficient depth to include a majority of the LC, or where retinal tissue was not imaged because the imaging range was too deep or the retina was skewed relative to the imaging plane, the image was discarded and re-acquired. This was particularly true when the examiner had less experience. The 24 scans in each image set were inspected to categorize them as having most or all of the 24 scans with the important parameters marked. A modest number of other image sets were not included when multiple scans among the 24 in a set failed to identify key structures. From a group of 10 image sets taken after the operator had become experienced with the process, only 3 had such algorithm failure to identify a structure; in each case it was a failure to identify the position of the CSI. Because the position of the CSI has not been optimized in automated algorithms, its data will not be included in this report.

To automatically delineate the anterior LC position and the BMO, the segmentation algorithm first analyses the OCT image gray values to highlight regions with a similar textural appearance as the respective target anatomy. This means that it examines the appropriate zone of the image in the expected location relative to BMO for a border of dark/light junction at the typical position of the anterior LC position. The final segmentation of LC position is identified by consolidating and connecting the individual positions of likely target regions under additional three-dimensional geometric regularization constraints, including expected curvature and surface area. This results in a most plausible, continuous contour of the anatomic structure of interest as defined by a spline fit to the identified locations of the anterior LC position.

For intraobserver-, interobserver-, and observer-algorithm comparisons, scans were selected at random from all acceptable image sets, with equal numbers of scans from right and left eyes and from men and women. From these images, two scans were selected from 10 eyes (20 total scans) for intraobserver and interobserver comparisons. For the observer-algorithm comparison, we compared two scans each from five eyes, six scans each from two additional eyes, and all 24 scans from three more eyes.

Images and horizontal and vertical scaling factors correcting for magnification were exported from Heidelberg Eye Explorer to generate a micrometer per pixel scale used to correct each image set. Images were exported at 100% zoom with 1:1-μm scaling. Using MATLAB (MathWorks, Inc., Natick, MA, USA),...
These points were then collected into data sets using Python in pixels, where the upper left corner of the image was (0, 0). (Fig. 2). Points were exported from ImageJ as (x, y) coordinates public domain by the National Institutes of Health, Bethesda, ImageJ software (http://imagej.nih.gov/ij/; provided in the identity and medical history of the subject. Graders used de-identified and randomized, so that graders were masked to the scan, and the two sets of data were compared after conversion from pixel positions to micrometers. The algorithm using the contrast difference between a darker anterior rim of the ONH and the point where collagenous beams of the LC produce a whiter zone marked the anterior border of the LC. The algorithm generates a continuous surface at the positions that it identifies as the anterior LC border and extending horizontally over the full BMO area. The resulting automated segmentation was exported from the Spectralis SP-X software, giving the depth of the LC every pixel from one side of the ONH to the other. ALD is measured relative to the connecting line between the two BMO endpoints in each B-scan.

It was not always obvious to observers where the position of the anterior LC was in some nasal areas of the ONH when the scan had shadowing from overlying blood vessels. Hence, in the comparisons to be presented, the data for observer–algorithm comparison are given only for the horizontal extent of the anterior LC border in which the observers felt that they could reliably mark it. Their 20 to 40 marked points per scan were converted to cubic spline fits and to sixth order polynomial fits and these were compared with the positions of the algorithm markings. There appeared to be minimal difference between the spline and polynomial fits, so the former will be presented here. The images compared in the observer–algorithm portion comprised 232 sets of markings of 9, 10) and multiple marks in yellow. The algorithms (bottom) marked the internal limiting membrane in red (ILM), the BMO endpoints in red dots and the anterior lamina position in solid yellow line (LCA). This is an example of substantial agreement between observer and algorithm.

Representative image with markings by observer and algorithm. These images of a single scan are shown unmarked (top), as marked by observer (middle), and as marked by the algorithm (bottom). The observer marks are at BMO (yellow numbers 9, 10) and anterior lamina (multiple marks in yellow). The algorithms (bottom) marked the internal limiting membrane in red (ILM), the BMO endpoints in red dots and the anterior lamina position in solid yellow line (LCA). This is an example of substantial agreement between observer and algorithm.

The images were cropped to include only B-scans. The images were de-identified and randomized, so that graders were masked to the identity and medical history of the subject. Graders used ImageJ software (http://imagej.nih.gov/ij/) provided in the public domain by the National Institutes of Health, Bethesda, MD, USA) to mark points of interest in the exported images (Fig. 2). Points were exported from ImageJ as (x, y) coordinates in pixels, where the upper left corner of the image was (0, 0). These points were then collected into data sets using Python (Python Software Foundation, Beaverton, OR, USA), MATLAB, and Excel (Microsoft, Redmond, WA, USA), rounded to the nearest pixel, and converted to micrometers using the scaling factors provided by Heidelberg Eye Explorer for analysis.

First, we marked on images the positions of the BMO endpoints (right and left) and the anterior LC position, using individual scan images. There were the following 4 comparative assessments: (1) a observer with himself (intraobserver), (2) between two observers (interobserver), (3) an observer compared to the algorithm (observer–algorithm), and (4) algorithm reproducibility between two sets of scans of the same ONH. The positions identified by observers were made from their markings on downloaded single scans within the ImageJ program. To do this, pictorial scans from randomly selected clock hours were exported with no overlying marks by the algorithm (blank images). The BMO right and left end points and LC anterior border positions were marked manually on the images. This was the approach used for intraobserver and interobserver comparison, with data converted from pixels in ImageJ to microns using appropriate scaling factors. For the intraobserver and interobserver comparisons, two scans from each of 10 optic nerves were compared (20 scans were marked by each observer).

Second, to compare observers to the algorithm, scans were again downloaded as blank images and the same scan was also downloaded with the algorithm markings of the positions of the BMO left and right endpoints and its marking of the anterior LC. The observers marked the stated positions while masked to the algorithm markings, then at a later time, the observers marked the algorithm marked positions, overlaid on the scan, and the two sets of data were compared after conversion from pixel positions to micrometers. The algorithm using the contrast difference between a darker anterior rim of the ONH and the point where collagenous beams of the LC produce a whiter zone marked the anterior border of the LC. The algorithm generates a continuous surface at the positions that it identifies as the anterior LC border and extending horizontally over the full BMO area. The resulting automated segmentation was exported from the Spectralis SP-X software, giving the depth of the LC every pixel from one side of the ONH to the other. ALD is measured relative to the connecting line between the two BMO endpoints in each B-scan.

To compare BMO endpoint markings, we calculated the difference in two-dimensional, horizontal, and vertical BMO position between markings for both left and right sides. We also calculated the difference in BMO width, using the scaling factor from pixels to micrometers and the diagonal distance between the endpoints (by Pythagorean theorem). For the ALD, we generated the mean depth difference across the LC, as well as the variance of ALD difference. Because the mean ALD difference could average differences that were in either direction (plus and minus), and this might minimize the true variability between measures, we also calculated the mean differences as absolute values. For this analysis, ALD was judged relative to the reference plane of a line joining the two
BMO endpoints, in both the observer and algorithm calculations.

Finally, we compared the data derived by the algorithm on the first and third sequential image sets of 24 images from 20 eyes of 20 persons (algorithm–algorithm). The software calculated automatically the mean ALD for each eye using all 24 sectors, using the BMO endpoint reference plane. The line joining the two BMO positions in each scan was defined as the reference position for calculating ALD. From this position, the distance to the spline fitted values for anterior lamina position were calculated as a mean value across the identified area(s) of the lamina.

We then compared the outcome parameters from the two image sets, calculating the differences in mean ALD and in a series of subregions of the ONH. These regional comparisons were performed for each of the 24 sector scans, for superior compared with inferior scan data, nasal compared with temporal data, central compared with peripheral data (dividing each disc into its central 50% area), and comparing four quadrants of the ONH by an X-shaped cross into superior, nasal, inferior, and temporal quadrants (Fig. 1). These regional data are not part of the software provided by the company, but were generated on separate programs based on the A-scan-wise exported values for ALD.

### Statistical Methods

For the statistical analysis of data, the outcome variables were the differences in the major parameters (BMO, LC anterior position, or ALD) between the following two pairs of methods: (1) intraobserver–, interobserver–, observer–algorithm, and (2) algorithm–algorithm. The P value for an outcome tested the hypothesis that the true mean difference for the outcome was zero, or, when comparing interobserver difference to observer–algorithm difference, the hypothesis was that the two methods had no significant difference. Estimates and P values were derived from generalized estimating equation (GEE) linear models that took into account correlations in outcome between scans from images of the same eye when more than one scan from an eye is included in an analysis. Outcomes from scans from an individual eye were assumed to have a compound symmetry correlation. An individual pair of outcomes was considered an outlier if its difference for that parameter was greater than the third quartile + 1.5 times the interquartile range or less than the first quartile minus 1.5 times the interquartile range. An outcome was considered an extreme outlier if it was greater than the third quartile + 3 times the interquartile range or less than the first quartile minus 3 times the interquartile range.

### Results

The degree of damage among the 51 eyes included ranged from a mean deviation (MD) value of zero or positive in 13 eyes (all with normal glaucoma hemifield testing), to 11 eyes with zero to –2 decibels in MD (all with normal hemifield tests), to 13 eyes with from –2 to –6 decibels MD (9/13 with abnormal hemifield tests), 7 with MD between –6 and –12, and 7 with MD worse than –12. The range of MD was from +2.66 to –29.88, spanning the full range from nearly half with normal function to those with advanced glaucoma. Twenty of the eyes were being followed without eye drop therapy as glaucoma suspects.

The intraobserver comparison for the positions of key landmarks showed very small differences between markings of the same scan (Table 2). In comparison to the actual BMO width, for example, the estimated mean difference in markings by the observer was less than 1% of the measurement (difference from zero, P = 0.34). A Bland-Altman plot for difference in the marking of the BMO between observers is shown in Figure 3. A linear regression between the mean difference and the lamina diameter had moderately positive slope indicating greater difference in wider laminas (linear regression: y = –0.0043 × –6.60; Pearson r = 0.649, P = 0.0019).

---

**Table 2. Intraobserver Differences in Key Parameters**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean Difference (95% CI)</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bruch’s membrane opening</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difference in X coordinate of left endpoint</td>
<td>5.34 (–6.72, 17.40)</td>
<td>0.39</td>
</tr>
<tr>
<td>Difference in Y coordinate of left endpoint</td>
<td>2.22 (–6.14, 10.59)</td>
<td>0.60</td>
</tr>
<tr>
<td>Distance between left endpoint</td>
<td>19.69 (9.68, 29.69)</td>
<td>0.0001</td>
</tr>
<tr>
<td>Difference in X coordinate of right endpoint</td>
<td>–11.84 (–36.66, 12.98)</td>
<td>0.35</td>
</tr>
<tr>
<td>Difference in Y coordinate of right endpoint</td>
<td>6.75 (0.20, 13.50)</td>
<td>0.04</td>
</tr>
<tr>
<td>Distance between right endpoint</td>
<td>28.18 (3.84, 52.51)</td>
<td>0.02</td>
</tr>
<tr>
<td>Difference in BMO width</td>
<td>–17.14 (–50.72, 16.44)</td>
<td>0.32</td>
</tr>
<tr>
<td>Percent difference in BMO width</td>
<td>–0.70 (–2.14, 0.74)</td>
<td>0.34</td>
</tr>
<tr>
<td>Anterior lamina cribosa position</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean difference in spline curve</td>
<td>–5.12 (–10.08, –0.15)</td>
<td>0.32</td>
</tr>
<tr>
<td>Mean absolute difference in spline curve</td>
<td>15.58 (9.59, 21.57)</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

Bolded values indicate statistical significance. Measures are in micrometers. Twenty scans were marked twice by the observer.

**Figure 3.** Interobserver comparison of BMO width. The difference in marking of the BMO width by two observers. In a GEE model, the difference between observers was not statistically significant (P = 0.93; Table 3).
The difference in repeated measurement of the LC position as defined by the spline curve that was fitted to the marked points was not significantly different between markings ($P = 0.32$) and represented less than 1% of the typical depth of the lamina (500 μm).

The interobserver differences in marking were also small (Table 3) and were approximately twice as large as those of the intraobserver differences in magnitude for location of the BMO parameter. While the positions of the left and right endpoints of the BMO were statistically different between observers, when these were used to calculate the difference this would represent between the BMO widths for each observer’s measurements, that difference was not significant ($P = 0.93$). Likewise, the percent difference in BMO width between observers was less than 1%. The LC anterior border was estimated by fitting a spline curve to the many points along the anterior LC marked by the observer. Both the mean difference and the mean absolute difference were significant between observer and algorithm, but the latter was only twice as large in magnitude as the interobserver value. Furthermore, as given in percent of the typical ALD of 500 μm, the mean difference of approximately 15 μm is a variation of 3%. Examples of marking of LC by two observers are seen in Figure 4 and comparative spline curves of LC position in Figure 5.

There were statistically significant differences between the observer and algorithm markings of the BMO endpoint points; however, these differences did not lead to statistically significant differences in the calculated BMO width, or to significant percent difference in BMO width, which varied less than 2% between observer and algorithm (Fig. 5, Table 4). The LC position was not significantly different in marking between observer and algorithm, when calculated as a mean difference (Table 4, Fig. 6). This mean difference of less than 4 μm is approximately 1% of the normal ALD, as referenced to the BMO. The absolute value of the mean difference was significantly greater than zero, indicating that more extreme values of difference existed that in effect were canceled out by the mean difference alone. Note that this comparison comprised only the zones marked by both observer and algorithm, while the algorithm continued its LC position estimation to the point below the BMO endpoint on either side in every case.

Further analysis of the comparison of human to algorithm marking of the position of the anterior lamina border is seen in Figures 7 and 8, where the mean position of the lamina in a spline fit performed by the observer and algorithm are compared and a Bland-Altman plot is shown for the comparison. Models of the comparative marking by observer and algorithm showed that the differences between them in BMO width and in mean anterior lamina position were not significantly related to age of the person who was imaged nor to the MD of the visual field of the eye that was imaged. For BMO width, the change (and 95% CI) in outcome per year increase in age was −0.88 μm (−2.25, 0.49; $P = 0.21$) and the change per decibel in MD was −3.08 μm (−9.33, 3.16; $P = 0.35$). For mean lamina position, the change with year of age was 0.26 μm (−0.31, 0.83; $P = 0.37$), while its change per decibel of MD was 1.88 μm (−0.94, 4.69; $P = 0.19$).

To compare the reproducibility of the algorithm in identifying the BMO width and ALD in repeat images of the same person, we took multiple 24 scan image sets of the same eye and compared the first and the third sets taken on the same day at the same sitting. There were 20 eyes of 20 persons included in this analysis, with mean age 64.6 ± 12.4-years old (range, 36–87 years). The algorithms mark a line from the fovea to the center of the BMO as the “horizontal” position for each eye, and the angle between this line and a true horizontal (relative to gravity) is called the FoBMOc Angle. The instrument aligns major landmarks in an initial image of an eye before taking a second image set, in order to assure that images of the same eye are aligned for longitudinal comparisons over time. In the 20 eyes in this test set, the mean for FoBMOc was −6.63° and the mean difference between two images of the same eye was 0.004°, or −0.6 ± 3.0% difference (not different from zero, $P = 0.38$, $t$-test). For this study, we used the position of the line joining the two endpoints of BMO in each scan as the reference zero for ALD. While the mean ALD was 481 ± 12 μm, the range was from 292 to 854 μm.

The ALD was slightly and significantly deeper when measured in the first scan than in the subsequent, repeated scan when the mean of all 24 sectors was calculated for the 20 pairs (Table 5, $P = 0.02$, $t$-test). However, this difference of 2.6 μm represented only a 0.54 ± 0.73% difference in the median ALD. Using 2 SD of these normally distributed data as an estimate for true change, a mean difference of 10 μm or more would be statistically greater than 97.5 percentile for the eyes imaged here. With the mean ALD of 480.9 μm, a 10 μm change would be equivalent to a change of 2.1%.

Data here are presented both as the actual measure in micrometers and change as a percentage of the baseline depth. The mean ALD values for individual scans in a 24 scan set are shallower in the horizontal zones (#7 thru #18) and deeper in the vertical zones (e.g., #19–#24, #1–#6). There

**Table 3.** Interobserver Differences in Key Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean Difference (95% CI)</th>
<th>$P$ Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bruch’s membrane opening</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difference in X coordinate of left endpoint</td>
<td>1.81 (−5.66, 9.28)</td>
<td>0.64</td>
</tr>
<tr>
<td>Difference in Y coordinate of left endpoint</td>
<td>7.58 (0.003, 15.15)</td>
<td>0.05</td>
</tr>
<tr>
<td>Distance between left endpoints</td>
<td>16.02 (9.07, 22.97)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Difference in X coordinate of right endpoint</td>
<td>0.19 (−28.18, 28.56)</td>
<td>0.99</td>
</tr>
<tr>
<td>Difference in Y coordinate of right endpoint</td>
<td>2.39 (−4.80, 9.57)</td>
<td>0.51</td>
</tr>
<tr>
<td>Distance between right endpoints</td>
<td>34.85 (13.83, 55.87)</td>
<td>0.001</td>
</tr>
<tr>
<td>Difference in BMO width</td>
<td>21.36 (−32.65, 29.92)</td>
<td>0.93</td>
</tr>
<tr>
<td>Percent difference in BMO width</td>
<td>−0.36 (−1.85, 1.12)</td>
<td>0.63</td>
</tr>
<tr>
<td>Mean difference in spline curve</td>
<td>14.74 (4.68, 24.80)</td>
<td>0.004</td>
</tr>
<tr>
<td>Mean Absolute difference in spline curve</td>
<td>21.99 (11.92, 32.06)</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Bolded values indicate statistical significance. Measures are in micrometers. Twenty scans were marked by each observer.
was a trend toward greater differences between scan sets in the horizontal zones as seen in mean, median, and SD of mean differences. Table 5 gives the mean ± SD of ALD, median ALD for each scan region, then in its right three columns, the differences between paired scans of the same region for the same three parameters. The horizontal 12 scans had a significantly greater SD in their mean difference between scans than did the vertical 12 scans (11.9 ± 1.2 compared with 9.0 ± 1.5, *P* = 0.0001, *t*-test).

We further divided the 24 scans into data that represented regional ALD zones. In doing this, we converting all data to the equivalent of a right eye, because the instrument acquires numbered scans clockwise in both right and left eyes (Table 6). In the baseline (first) image set, the superior half of the LC was significantly deeper than the inferior half, while the temporal quadrant was significantly shallower than the superior quadrant in baseline depth. The regional differences between two image sets of each eye in mean ALD by region were small (right two columns, Table 6), with even quadrant level mean differences less than 6 μm, representing 1.1% or less of the baseline ALD thickness (none significant, lowest *P* value = 0.43). The SD of mean difference in regions varied from 13.5 to 19.0 μm. These data indicate that to exceed the 97.5 percentile for difference, a mean change in a quadrant or a region would be significant if it were greater than 27 to 38 μm. This would indicate that a change of 8% in a region would be outside interimage variability. These differences should be taken simply as examples of how regional data may be acquired and not as either typical for normal or characteristic of glaucoma eyes, because this was a mixed group of undamaged and damaged eyes.

**DISCUSSION**

Diagnostic testing for glaucoma must efficiently and validly identify those parameters that are related to disease risk and...
progressive change. Early research in quantifying the nature of
glaucoma damage used the cup-to-disc ratio in color photographs, which has poor interobserver reliability. Digital imaging and quantitative topography of ONH rim area with the Topcon Imagenet and the Heidelberg Retinal Tomograph provided measures that were reproducible, automated, and quantitative. Clinical assessment of nerve fiber layer defects was shown to predict later visual field loss. Quantitative measurement of nerve fiber layer thickness and ONH structure with OCT now permits detailed examination of LC substructure and its positional movements with change in IOP and intracranial pressure. For digital analysis to be useful in clinical practice and research, automated methods are needed that require minimal observer input.

The automated algorithms studied here were generally acceptable in identifying the position of the BMO, anterior LC, and ALD, when compared with intraobserver and interobserver differences in marking these structures. There was a mean difference of 2% in BMO width between observer and algorithm. Given the high degree of variability in ONH structures among eyes, the reproducibility in ALD, with a 95% confidence limit of 10 μm, is reasonable. The actual range

---

**FIGURE 5.** Spline curve fits for two observers in marking of anterior lamina position. Three examples of spline fitting to marked locations of the anterior lamina by observer one (red) and observer two (blue) for same scans shown in Figure 3. Upper example shows close agreement, middle example shows divergence at right side of lamina, and lower example shows substantial difference between observers. Y axis is depth of lamina in micrometers, using a BMO endpoint reference plane and X axis is distance from left to right side of marked lamina (pixels).
TABLE 4. Observer–Algorithm Differences in Key Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean Difference (95% CI)</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bruch’s membrane opening</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difference in X coordinate of left endpoint</td>
<td>-7.63 (-25.04, 9.78)</td>
<td>0.39</td>
</tr>
<tr>
<td>Difference in Y coordinate of left endpoint</td>
<td>9.50 (0.60, 18.40)</td>
<td>0.04</td>
</tr>
<tr>
<td>Distance between left endpoint points</td>
<td>39.20 (25.73, 54.66)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Difference in X coordinate of right endpoint</td>
<td>25.91 (-13.34, 65.15)</td>
<td>0.20</td>
</tr>
<tr>
<td>Difference in Y coordinate of right endpoint</td>
<td>3.05 (-6.37, 12.46)</td>
<td>0.53</td>
</tr>
<tr>
<td>Distance between right endpoint points</td>
<td>47.90 (11.21, 84.60)</td>
<td>0.01</td>
</tr>
<tr>
<td>Difference in BMO width</td>
<td>34.92 (-24.67, 94.52)</td>
<td>0.25</td>
</tr>
<tr>
<td>Percent difference in BMO width</td>
<td>1.85 (-1.19, 4.90)</td>
<td>0.25</td>
</tr>
<tr>
<td>Anterior lamina cribosa position</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean difference in spline curve</td>
<td>3.77 (-21.61, 29.14)</td>
<td>0.77</td>
</tr>
<tr>
<td>Mean absolute difference in spline curve</td>
<td>40.30 (14.03, 66.57)</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Bolded values indicate statistical significance. Measures are in micrometers. Ninety scans were marked each by observer and algorithm.

FIGURE 6. Examples of observer compared with algorithm marking of lamina. Scans comparing identification of LC position by observer and algorithm. For each of the three scans, the unmarked image is on the left, the observer markings are in the middle, and algorithm markings are on the right. The upper set shows good agreement between observer and algorithm. The middle set shows an example where the observer marked the mid-ONH zone where lamina was definitely identifiable, while the algorithm extended its spline curve through an area masked by overlying vessels. The bottom set shows a major difference in both BMO endpoint marking and LC position between observer and algorithm.
of the 95% confidence interval for the mean absolute difference between human and algorithm in LC position was from 14.03 to 66.57 μm, so that values greater than 67 μm would occur 2.5% of the time.

While the actual LC position may be occasionally difficult to determine, observers and the algorithm agree about its location sufficiently often to study use of the algorithm further. It is still necessary to inspect individual scans, because outlying values are not uncommon and must be identified so that images may be repeated or excluded. Some portions of the LC are masked by overlying blood vessels, which leads to difficulty in identification of LC position and may explain the higher variance of algorithmic marking of LC position in scans along the horizontal axis. Many past reports have used horizontally oriented (as opposed to radial) scans and find that the superior pole and inferior pole horizontal scans are so narrow that the BMO is difficult to identify. As a result, only more central, horizontal scans were used. Radial scanning is less susceptible to this problem and should be preferred.

The ability to analyze LC position with automation will facilitate study of ONH biomechanics relevant to glaucoma. The LC has a complex three-dimensional shape, which was reduced to a single index number in some reports. We previously showed that greater LC deformation with IOP change occurred in clock-hour regions in which there were more remaining nerve fibers. Software to assess lamina change can and should provide regional data as illustrated here, to give a fuller picture of LC remodeling. It will be important to have both normative and glaucoma-specific databases of ALD to assess true abnormality and change. Swept-source OCT that may detect structures more fully, but must be implemented in an effective, automated approach. Software will benefit from showing quality measures. Some recent reports have attempted to evaluate the structures in the optic nerve head using machine learning or artificial intelligence methods, in which various test set to validation set approaches have identified some aspects of structure with minimal human intervention. As yet, these methods do not identify the position of the

**Figure 7.** Observer compared with algorithm in mean lamina position. The mean position of the anterior lamina in 116 scans as identified in a spline fit are compared as performed by the human observer (D1) and the algorithm. All values are compared with a line of agreement.

**Figure 8.** Bland-Altman Plot comparing observer and algorithm in lamina position. A Bland-Altman–type plot is shown, with the broad hashed lines representing the 95% confidence limit for the mean difference from GEE models that were used to accommodate the use of more than one scan from an eye.
anterior LC and have “a tendency to always identify LC insertions into the sclera that were not always present in the manual segmentations.” Further development of such methods may improve the automated measurements needed.

It has been proposed that the CSI, when marked by observers, is a better reference position for ALD measurement than BMO endpoints, as the latter may change with age. The BMO endpoints are a more useful reference position for automated ALD measurement, because there is, as yet, no validated means to identify CSI by algorithm. It may improve the automated measurements needed. However, the sclera changes in axial length and curvature at the CSI with IOP change, 28 so the CSI may be no more stable a reference than BMO endpoints.

In summary, an automated method for identifying key parameters at the ONH performed well in reproducibility and validity by comparison to observers for the BMO endpoints and anterior LC position. The BMO endpoints are a reliable reference plane for ALD estimation. Outlying values in both individual scans within a 24 scan set, or, in all images of unusual ONHs need to be accounted in use of a fully automated analysis mode.

Acknowledgments

The authors thank Christopher Lee and his colleagues at Heidelberg Engineering for provision of instruments, experimental software, and expert consultation.

Supported in part by donations to the Glaucoma Center (Baltimore, MD, USA) by William Forrester, Saranne and Livingston Kosberg, and Troy Williams.

Disclosure: X.J. Duan, None; J.L. Jefferys, None; H.A. Quigley, None.

References


