Variability between experts in defining the edge and area of the optic nerve head

S. Alayon, M. Gonzalez de la Rosa, F.J. Fumero, J.F. Sigut Saavedra, J.L. Sanchez

Departamento de Ingeniería de Sistemas y Automática y Arquitectura y Tecnología de Computadores, Universidad de La Laguna, La Laguna, Spain

Hospital Universitario de Canarias, Departamento de Oftalmología, Universidad de La Laguna, La Laguna, Spain

ABSTRACT

Objective: Estimation of the error rate in the subjective determination of the optic nerve head edge and area.

Method: (1) 169 images of optic nerve disk were evaluated by five experts for the defining of the edges in 8 positions (every 45°). (2) The estimated areas of 26 cases were compared with the measurements of the Cirrus Optical Coherence Tomography (OCT-Cirrus).

Results: (1) The mean variation of the estimated radius was ±5.2%, with no significant differences between sectors. Specific differences were found between the 5 experts (p < 0.001), each one compared with the others. (2) The disk area measured by the OCT-Cirrus was 1.78 mm² (SD = 0.27). The results corresponding to the experts who detected smaller areas were better correlated to the area detected by the OCT-Cirrus (r = 0.77–0.88) than the results corresponding to larger areas (r = 0.61–0.69) (p < 0.05 in extreme cases).

Conclusions: There are specific patterns in each expert for defining the disk edges and involve 20% variation in the estimation of the optic nerve area. The experts who detected smaller areas have a higher agreement with the objective method used. A web tool is proposed for self-assessment and training in this task.

© 2011 Sociedad Española de Oftalmología. Published by Elsevier España, S.L. All rights reserved.

RESUMEN

Objetivo: Estimar el grado de error en la determinación subjetiva del límite papilar.

Método: (1) Fueron evaluadas 169 imágenes papilares por cinco expertos para delimitar los bordes papilares en 8 posiciones (cada 45°). (2) Las áreas estimadas en 26 casos se compararon con las medidas mediante tomógrafo de coherencia óptica (OCT-Cirrus).

Resultados: (1) La variación media del radio papilar estimado fue de ±5.2%, sin diferencias significativas entre sectores. Entre los cinco expertos existieron diferencias específicas...
Introduction

Glaucoma is the most frequent cause of irreversible blindness in the population of industrialized countries. The best-known methods for diagnosing and controlling the evolution of glaucoma are optic nerve head (ONH) morphology assessment, establishing the thickness of the retinal optic nerve fiber layer and the subjective analysis of retinal sensitivity in the visual field of the patient.

The inter-patient variability of the ONH appearance is very high. Considerable experience is required not only to differentiate a normal ONH from a diseased one but also to precisely define its anatomic limits. The methods for analyzing the surface morphology of the papilla by means of direct observation, such as confocal scanning laser (HRT) or stereoscopic images, to not allow a view of the internal structure of the ONH and therefore subjective operator assessments are applied. However, observation by means of optic coherence tomography (OCT) is able to define the internal edges of the scleral channel.

It would be desirable to establish criteria to develop a closer match between the results of subjective procedures and the objective OCT measurements or, in the best of cases, to perform an automatic definition of the papillary contour. Precise ONH segmentations are essential to achieve this purpose and therefore many research papers have focused on this issue from the medical viewpoint as well as from that of software developers.

Interobserver variability cannot be entirely removed although the identification of its causes and a quantitative assessment could make a decisive contribution for improving the quality of clinical practice and research. The study of this type of interobserver variability in clinical practice has been addressed in numerous medical activities such as cardiology, radiology, neonatology or in the domain of ophthalmology.

The interest in achieving precise and reproducible definitions of ONH limits has prompted the authors to measure interobserver variability and the patents thereof against the objective measurements provided by OCT. In parallel and also as a consequence of this work, we have aimed at designing a computer tool, which was given the name of RIM-ONE, for self-assessment and training in manual definition of papillary edges. This has been developed as part of a joint ONH analysis research project by 3 Spanish hospitals, i.e. the Canarias University Hospital, the San Carlos Clinic Hospital and the Miguel Servet University Hospital.

Subjects, materials and methods

This study utilized 169 ONH images (118 healthy volunteers, 11 ocular hypertensive patients, 12 incipient glaucomas, 14 moderate and 14 advanced glaucomas) taken by means of various non-mydriatic retinographs at the 3 above mentioned hospitals, thus guaranteeing a representative and heterogeneous sampling. All the retinographs were captured with specific flash intensities to avoid color saturation.

The images were collected in accordance with the ethical standards established in the 1964 Helsinki declaration. Ethical committee approval was obtained and the patients were informed about the objectives of this study. The utilized images are available in the RIM-ONE website although said database does not provide detailed clinical information about the photographed subjects.

The reference taken (gold standard) was the middle position of the manual limits of the papillary edge performed by 5 experts. To this end, the mean centroid of all the contours was calculated and taken as a reference point (Fig. 1A). On the basis of this point, 8 axes were drawn (at 45° intervals) measuring the average radius (AR) or the mean distance to the intersection of each of the 5 contours (Fig. 1B, d1–d5 and C, AR1–AR5).

Subsequently, the deviation of each expert against the pattern was compared as a measure of inter-observer variability. For each definition, the variability percentage against the patterns in each direction (i) was calculated by means of the variation coefficient (VC), i.e.:

$$CV_i = \frac{\text{abs}(R_i - AR_i)}{AR_i} \times 100, \quad 1 \leq i \leq 8 \%$$

The RIM-ONE website includes the Matlab programming language code (Mathworks, Massachusetts, USA) which automatically calculates said measurements, comparing the results of each manual ONH definition with the pattern contour of the image.

After obtaining the deviations of each expert against the pattern, interobserver variability was studied comparing deviations with each other (mean and standard deviation [SD]).

In the second part of our study, the areas estimated on the basis of the number of pixels inside the contour defined by
Fig. 1 – Preparing the pattern edge. (A) Reference point and directional axes 1–8; (B) measuring distances and (C) pattern: mean radii AR1–AR8 and reference point.

Results

The mean variability results between each expert and the pattern for all images and redirections are shown in Table 1. Positive or negative values indicate differences against the pattern in one direction or another, i.e. the radius marked by the user is smaller or larger than the mean radius. The differences in the size of the chosen radius against the mean value are of 5.2%, without significant differences between sectors. There are significant mean differences between each expert and the other ones (p < 0.05), except between the second and
Fig. 2 – Approximate definition of ONH with an ellipse.

Fig. 3 – Precise adjustment of the papillary edge with uniformly distributed radii.
the third, as well as between all the cases in some of the radii \( (p < 0.001) \).

Table 2 illustrates the mean and variability of each radius between all the experts (for each direction) in absolute value. In this context, the mean value is the value of the mean differences of the radii marked by the users with the mean radius (pattern). The source from which the differences are measured is located at the end of the pattern radius (mean radius) and not in the center of the papilla. The SD values clearly show that the adjustments made by each individual against the mean value is not uniform in all radii, which indicates a high degree of subjectivity in the form of carrying out the definition. Variability was lower in upper sectors (R1, R2 and R8), with significant differences against lower sectors (R4, R5 and R6, \( p < 0.005 \)) and of these vis-à-vis horizontal sectors (nasal R3 and temporal R7, \( p < 0.005 \)) which exhibited the highest variability.

The variation coefficient of the subjective measurement of the papillary area was of 9.8%. For OCT, the mean papillary size was of 1.79 mm\(^2\) (SD = 0.27). Table 3 shows that the correlation between the estimated area and that measured by OCT exhibited differences between experts (which were significant for extreme values).

In the absolute comparison of the area value highly significant differences were observed between some experts \( (p < 0.0001) \) but not between others.

It must be noted that experts 4 and 5 exhibited the best correlations with the OCT area and also were the ones which defined the smallest areas. Expert 2 defined the maximum area and exhibited the worst correlation. Expert 5 is one of the experts which exhibited the lowest variability vis-à-vis the mean radius.

### Discussion

This study is the result of the drive for achieving precise and reproducible definitions of ONH limits.

A new computer tool, accessible through the web, has been proposed for training and self-assessment of experts in this task. Said tool is an initial prototype as there are no references to similar systems in specialized literature. The tool proposed for ONH segmentation could be very useful for those who are in a learning period as well as for experts because the tool facilitates a more objective determination of ONH as it is always focused on a mean point with equidistant measurements.

In what concerns its limitations, the graphic interface of the tool could be improved as well as the interpolation between points to improve visibility (by using curves instead of straight lines to join the points marked by the experts). A further improvement would consist in displaying a comparison of results between the points marked by the user and the gold standard established for that image. On the other hand, it would be convenient to enable inputting OCT data into the application so that it could gain experience and improve the gold standard. All these potential improvements are included in the future development of the application.

The experience of this study gives rise to the conclusion that there is a tendency to oversize the ONH area when estimated by surface observation. Accordingly, this tendency should be reverted possibly not taking tissue color change as reference but the slightly more internal area where the vessels begin to bend. Consequently, the initial pattern we have chosen should be pondered toward the personal criteria followed by observer 5 when performing ONH definition as this expert is the one exhibiting the highest match with OCT.

It also seems likely that the best match of experts in the task of defining the nasal and temporal contours is because in these areas papillary atrophies, pigmentation areas and other related areas appear with greater frequency and constitute obstacles for the precise detection of the edges.

In the future, we expect to complete the training tool available in the RIM-ONE website, to propose a similar measure for intraobserver variability and to carry out a study of the

---

### Table 1 – Mean variability percentage between manual definitions of each expert and the pattern.

<table>
<thead>
<tr>
<th>Expert</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>–1.0</td>
<td>3.7</td>
<td>3.9</td>
<td>–4.0</td>
<td>–2.5</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>3.4</td>
<td>4.2</td>
<td>3.5</td>
<td>5.2</td>
<td>3.8</td>
</tr>
</tbody>
</table>

### Table 2 – Mean variability percentage of all experts in each area.

<table>
<thead>
<tr>
<th>Radius</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
<th>R5</th>
<th>R6</th>
<th>R7</th>
<th>R8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.5</td>
<td>3.7</td>
<td>4.2</td>
<td>3.8</td>
<td>3.8</td>
<td>3.7</td>
<td>4.0</td>
<td>3.3</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>2.9</td>
<td>3.4</td>
<td>3.8</td>
<td>3.4</td>
<td>3.5</td>
<td>4.0</td>
<td>4.2</td>
<td>3.3</td>
</tr>
</tbody>
</table>

R1, superior vertical (90°); R2, superior nasal (45°); R3, horizontal nasal (0°); R4, inferior nasal (315°); R5, inferior vertical (270°); R6, inferior temporal (225°); R7, horizontal temporal (180°); R8, superior temporal (135°).

### Table 3 – Correlation between the papillary area estimated by the expert and that measured by optical coherence tomography and the mean variation of the area determined by each expert.

<table>
<thead>
<tr>
<th>Correl. coefficient (r)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of papillary area OCT</td>
<td>98.0</td>
<td>109.9</td>
<td>108.5</td>
<td>91.1</td>
<td>92.5</td>
<td>100.0</td>
</tr>
</tbody>
</table>
relationship between contour definition variability and the papillary area.

**Funding**

This paper has been partially funded by the following projects: SolSubC200801000097 of the Research Agency of the Canary Islands, Innovation and Information Society and PI09/0601 of the Health Research Fund of the Carlos III Institute, both with FEDER funds.

**Conflict of interests**

No conflict of interests has been declared by the authors.

**Acknowledgments**

The authors wish to express their gratitude to Marta González Hernandez and Dr. Tinguaro Díaz-Alemán (Canarias University Hospital), Dr. Julián García Feijoo (San Carlos Clinic Hospital) and Dr. Susana Pérez Oliván (Miguel Servet University Hospital) for their invaluable contribution in the development of the manual definitions of ONH and the classification of images.

**References**