Intraobserver and Interobserver Reproducibility in the Evaluation of Optic Disc Stereometric Parameters by Heidelberg Retina Tomograph

Stefano Miglior, MD,1 Elena Albé, MD,2 Magda Guareschi, MD,2 Luca Rossetti, MD,2 Nicola Orzalesi, MD2

Purpose: To assess intraobserver and interobserver reproducibility of the measurement of stereometric parameters of the optic disc by means of the Heidelberg Retina Tomograph I (HRT).

Study design: Observational study, with interobserver variability.

Participants: Fifty-five volunteers (healthy subjects and patients with glaucoma).

Methods: HRT examination of the optic disc was repeated on 3 consecutive days on 1 eye of each of the 55 subjects. During each session, five single images were randomly acquired by two independent observers. One mean topography image (MTI), based on three single images, was then built at each session. For the intraobserver interimage evaluation, the two observers traced their own contour line on one randomly chosen MTI. This procedure was repeated three times. For the intraobserver interimage and interobserver intra/interimage evaluations, the two observers traced their own contour line on the MTI of the first session, which was then automatically superimposed on the MTIs of the other two sessions.

Main Outcome Measures: Reproducibility of the 12 stereometric parameters was calculated for each observer by means of the intraclass correlation coefficient (ICC). The expected range of variability between two independent evaluations was calculated by the scatter-plots of each test—retest difference versus their mean. The standard deviation of the mean test—retest score differences was used to describe the spread of score differences.

Results: The ICC ranged between 0.79 and 0.99 for intraobserver interimage and between 0.56 and 1 for intraobserver interimage evaluation. The ICC ranged between 0.54 and 0.99 for interobserver interimage and between 0.65 and 0.97 for the interobserver interimage evaluation. ICC was almost perfect to perfect for planimetric measures (0.81 < ICC ≤ 1), substantial to almost perfect for volumetric and cup measures (0.61 < ICC ≤ 0.99), and moderate to almost perfect for retinal nerve fiber layer related measures (0.41 < ICC < 0.99). The expected variability was low (95% confidence interval, <±9%). Interimage evaluation showed a higher variability than interimage evaluation in both of interobserver (P = 0.012) and intraobserver evaluation (P = 0.028 and P = 0.031 for the two observers).

Conclusions: Measurement of optic disc stereometric parameters by HRT is highly reproducible. However, the use of retinal nerve fiber layer–related parameters should be taken cautiously. The image acquisition–induced variability seems larger than the operator-induced variability. Ophthalmology 2002;109:1072–1077 © 2002 by the American Academy of Ophthalmology.

Primary open-angle glaucoma consists of a progressive loss of retinal ganglion cells that leads to subsequent visual field damage. Retinal ganglion cell loss is clinically detectable as a structural modification of the optic disc or as a loss of retinal nerve fiber layer (RNFL) reflectivity, particularly when using a blue or green light.1 Because the optic disc and RNFL are usually evaluated by means of subjective techniques (ophthalmoscopy, retinography, etc.) that clearly rely on the examiner’s experience, it seems advantageous to introduce objective techniques to obtain quantitative measurements. The results of such quantitative analysis should improve the ability not only to distinguish a normal from an abnormal (glaucomatous) optic disc and/or RNFL but also to distinguish true from false progressive optic disc/RNFL changes, which is as important as making an accurate early diagnosis in the correct clinical management of the disease.

This has led to the development of a number of new...
devices over the last few years, some of which provide quantitative measures of a variety of optic disc parameters,2,3 and others of which are designed to measure only RNFL,4,5 or entire retinal thickness.5,6

The Heidelberg Retina Tomograph (HRT) is a scanning laser ophthalmoscope that allows three-dimensional topographic analysis of the optic disc and the retina by providing rapid measurements of optic disc topography on a pixel-by-pixel basis and analyzing a number of optic disc parameters. However, this requires manually defining the optic disc by tracing a contour line along the disc margin and thus introduces a potential source of variability that may negatively affect the result of the examination. It also needs to be borne in mind that various observers may be involved in the process of image acquisition and processing; this introduces another source of error that is particularly important during the long-term follow-up typical of glaucomatous patients.

For this reason caution is necessary when considering serial quantitative HRT data, unless reproducibility can be precisely estimated. Although the reproducibility of the topographic measures provided by HRT has been widely investigated, most of the studies have addressed intraobserver intraimage and interimage reproducibility,2,7-11 Only Orgu¨l et al12 have evaluated the intraobserver intrimage and interimage reproducibility, and they found a wide range of values for the coefficient of variation of the repeated measures (<0.01 –0.24, depending on the topographic parameter). Only Hatch et al13 have evaluated interobserver reproducibility. This last study was limited to the intratome assessment of a single original topography image, and they found a substantial to almost perfect agreement in the evaluation of intraclass correlation coefficient (ICC) (0.67–0.94, depending on the topographic parameter).

Because no comprehensive data have yet been published, concerning the reproducibility of the measurement of HRT stereometric parameters of the optic disc (including intraobserver and interobserver, as well as intratome and interatome evaluations)), this study was designed with the following aims: (1) to assess reproducibility of HRT optic disc measurements as they are usually made in the clinical setting (thus including all the possible sources of variability inherent in the procedure); and (2) to estimate the HRT measurements that can be expected in eyes believed not to have experienced optic disc progressive changes.

Material and Methods

The study population consisted of 55 volunteers aged between 25 and 79 years attending the Outpatient Service at the S. Paolo Hospital. The protocol of the study was approved by the local ethical review committee. After informed consent was obtained, one eye was randomly chosen for each subject. The sample included 11 eyes affected by primary open angle glaucoma at different stages and eyes with a wide range of disc, rim, and cup sizes and different morphologic characteristics. The glaucomatous eyes were pooled with normal eyes to have a group of eyes with completely different morphologic features (very narrow or very thick neurorretinal rims, enlarged cups or absence of cups, cups with temporal flat slopes, steep “punched-out” cups, etc.), which may affect the reproducibility of HRT topographic disc measurements.14,15

Inclusion criteria were 20/20 visual acuity, a clear lens, and normal retina. Exclusion criteria were myopia >−6 diopters, optic disc abnormality (i.e., drusen or tilted disc), and history of neuroophthalmologic diseases. All the patients included in the study underwent a complete ophthalmologic examination. Biomicroscopy of the anterior segment and intraocular pressure measurement by means of Goldmann tonometry were made before pupillary dilation. The indirect ophthalmoscopic fundus oculi evaluation and HRT examination were made in mydriasis. Visual field was determined using the standard 30/II Humphrey full-threshold strategy, and primary open angle glaucoma was defined on the basis of the criteria suggested by the European Glaucoma Society.16

The HRT (Heidelberg Instruments, Heidelberg, Germany—version 2.01) is a confocal scanning laser ophthalmoscope that uses a diode laser (wavelength 670 nm) to scan the retinal surface in three dimensions. A topographic image consists of a series of 32 confoical images (i.e., 32 optical sections) at 32 consecutive focal planes, each made up of 256 × 256 pixels. The HRT examination was performed using a 10° angle view. The reference plane was set at the standard value of 50 µm below the mean retinal surface at the temporal sector between 350° and 356°.

For study purposes, five image series were acquired each day for 3 consecutive days (a total of 15 image series per eye) by two independent and well-trained observers. All the acquired image series were recorded only after the built-in quality control of the HRT software confirmed that the optimal settings were found. Each day a mean topographic image (MTI) was then computed from three of the five original image series (a total of 3 MTIs per eye), by excluding the image series in which the optic disc was not well centered in the monitor.17

Intraobserver intratome and interimage reproducibility for each of the 12 predefined stereometric parameters was tested according to the following procedure:

- For the intratome comparison, each observer manually drew the contour line on the same MTI three times. This procedure was performed on one MTI only, randomly chosen from the three MTIs.
- For the interimage comparison, first observer E drew the contour line and then observer M did the same on MTI-1. The two contour lines were then automatically superimposed by the HRT software on the MTI-2 and MTI-3.

Interobserver intratome and interimage reproducibility for each of the 12 parameters was tested according to the following procedure, using the same contour lines previously drawn by both observers for the intraobserver/interimage evaluation:

- For the intratome comparison, we analyzed the three combinations obtained at the same sessions by observer E and observer M (E1–M1, E2–M2, E3–M3).
- For the interimage comparison, we analyzed the six combinations obtained during the different sessions by observer E and observer M (E1–M2, E1–M3, E2–M1, E2–M3, E3–M1, E3–M2).

Reproducibility was evaluated by means of the ICC using the formula:

$$r_i = \sigma_i^2 / (\sigma^2 + \sigma_i^2 + \sigma_e^2)$$

where: $r_i$ = ICC; $\sigma_i^2$ = QM(e) or variability intraclass; $\sigma_e^2 = [QM(a) – QM(e)]/n$ or (interclass–intraclass variability)/n² of classes.18–20

The ICC was calculated for each independent test-retest evaluation, after which the mean ICCs were calculated with their ranges. The agreement was judged to be poor (an ICC of between
deviation was 27.87

The study is based on a total of 165 MTIs, whose mean standard

Results

The expected range of variability between two independent
evaluations of each stereometric parameter was obtained by as-
sessing test-retest variability. This was calculated using the paired
score differences of the test minus the retest (i.e., first minus
second) for each patient. A scatterplot was then built by plotting
each test-retest difference versus its mean. The mean and standard
deviation of the test-retest score differences were then calculated.
The standard deviation from the mean test-retest score differences
was used to describe the spread of score differences, which is an
estimate of their dispersion between repeated measurements. The
95% confidence intervals of the mean difference are thus the
boundaries of the expected range of variability. Because the
mean difference between the two repeated measurements per-
formed on our 55 cases tends to zero, the 95% confidence intervals
are expressed as percent values around the mean measurement.
Statistical comparisons were performed by means of Wilcoxon’s
nonparametric test.

Table 1. Descriptive Statistics of Topographic Parameters. Mean ± Standard Deviation (Range).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Observer E</th>
<th>Observer M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disc area</td>
<td>1.856 ± 0.378 (0.917–2.796)</td>
<td>1.888 ± 0.398 (1.017–2.896)</td>
</tr>
<tr>
<td>Cup area</td>
<td>0.636 ± 0.434 (0.201–1.915)</td>
<td>0.654 ± 0.426 (0.200–1.857)</td>
</tr>
<tr>
<td>Cup/disc area</td>
<td>0.319 ± 0.184 (0.201–0.784)</td>
<td>0.323 ± 0.175 (0.201–0.737)</td>
</tr>
<tr>
<td>Rim area</td>
<td>1.220 ± 0.289 (0.526–1.890)</td>
<td>1.234 ± 0.274 (0.609–1.914)</td>
</tr>
<tr>
<td>Height variation contour</td>
<td>0.409 ± 0.107 (0.210–0.648)</td>
<td>0.402 ± 0.109 (0.208–0.621)</td>
</tr>
<tr>
<td>Cup volume</td>
<td>0.170 ± 0.174 (0.201–0.831)</td>
<td>0.146 ± 0.167 (0.300–0.786)</td>
</tr>
<tr>
<td>Rim volume</td>
<td>0.323 ± 0.143 (0.026–0.749)</td>
<td>0.318 ± 0.134 (0.083–0.707)</td>
</tr>
<tr>
<td>Mean cup depth</td>
<td>0.242 ± 0.107 (0.303–0.504)</td>
<td>0.241 ± 0.106 (0.223–0.487)</td>
</tr>
<tr>
<td>Maximum cup depth</td>
<td>0.617 ± 0.204 (0.085–1.069)</td>
<td>0.619 ± 0.204 (0.069–1.078)</td>
</tr>
<tr>
<td>Cup shape measure</td>
<td>−0.163 ± 0.083 (−0.360–0.004)</td>
<td>−0.166 ± 0.082 (−0.340–0.017)</td>
</tr>
<tr>
<td>Mean RNFL thickness</td>
<td>0.245 ± 0.079 (0.087–0.426)</td>
<td>0.242 ± 0.078 (0.097–0.432)</td>
</tr>
<tr>
<td>RNFL cross-sectional area</td>
<td>1.164 ± 0.377 (0.481–2.064)</td>
<td>1.143 ± 0.346 (0.524–1.875)</td>
</tr>
</tbody>
</table>

The descriptive statistics values of the 12 parameters for each
observer are summarized in Table 1; there was no statistically
significant difference between the two observers for any of the
parameters.

ICC

The analysis of the intraobserver intrainteval reproducibility
showed a mean ICC ranging from substantial to almost perfect
(between 0.79 and 0.99) for each of the 12 parameters (Table 2).
The analysis of the intraobserver interimage reproducibility showed a mean ICC ranging from moderate to perfect (between
0.56 and 1.00) for each of the 12 parameters (Table 2).
The analysis of the interobserver intrainteval reproducibility
showed a mean ICC ranging from moderate to almost perfect
(between 0.54 and 0.99) for each of the 12 parameters (Table 3).
The analysis of the interobserver interimage reproducibility
showed a mean ICC ranging from substantial to almost perfect
(between 0.65 and 0.97) for each of the 12 parameters (Table 3).

Scatterplots

The expected measurements for the intraobserver intrainteval
and interobserver interimage evaluations are reported in Tables 4
and 5. All the mean expected measurements were within ±10%,
most being within ±5%. Most of the expected measurements coming
from the single test–retest comparisons were within ±10%, with

RNFL = retinal nerve fiber layer.
only “cup volume” and “mean RNFL height” having a maximum value between ±10% and ±15%. The results arising from the intraobserver intraimage evaluation and interobserver intraimage evaluation were similar to those given in the tables.

The standard deviations of the differences of the intraobserver interimage evaluations of observers E and M were not significantly different (P = 0.436 Wilcoxon’s test). The standard deviations of the differences between the intraobserver intraimage and interimage evaluations were significantly different for both observers E (P = 0.028) and M (P = 0.031). The standard deviations of the differences between the intraimage and interimage analysis in the interobserver evaluation were significantly different (P = 0.0128).

## Discussion

Primary open-angle glaucoma patients are affected by optic nerve head modifications and retinal nerve fiber loss, as well as by progressive visual field alterations. To plan a therapy designed to preserve visual function for as long as possible, it seems clinically appropriate to use instruments capable of detecting progressive structural changes.

An important step in the follow-up of a chronic disease is to know whether the observed changes are related to test variability or to real disease progression. For this reason, it is useful to know the major sources of variability, which parameters may be negatively affected, and what value could be expected solely from the “background noise” of the measurement. This study evaluated the observer-related and image-related sources of HRT variability when two independent observers analyze the same (intraobserver intraimage) or different images (intraobserver interimage), as well as the reproducibility of the two independent observers (interobserver intraimage and interimage).

The reproducibility of the measurements was high for most of the stereometric parameters, with the mean ICC ranging from 0.81 and 0.99 for cup- and rim-related parameters, and from 0.54 and 0.96 for RNFL-related parameters (height variation contour, RNFL cross-sectional area and mean RNFL thickness). In the single test-retest evaluations, the ICC ranged from 0.60 and 1.0 for cup- and rim-related parameters, and from 0.54 and 0.96 for RNFL-related parameters. The mean expected variability, estimated by means of the scatterplot (at a confidence level of 95%), was between ±0.55% and ±8.98% for cup- and rim-related parameters, and between ±2.25% and ±9.65% for RNFL-related parameters, and from 0.54 and 0.96 for RNFL-related parameters.
related parameters. In relation to the single test-retest evaluations, the expected variability at the same confidence level ranged from \(\pm 0.40\%\) and \(\pm 14.05\%\) for cup- and rim-related parameters and from \(\pm 1.53\%\) and \(\pm 11.82\%\) for RNFL-related parameters.

Of the two major sources of variability (the tracing of the contour line and the mean image construction, which includes the acquisition of the three original image series and the software processing to build the mean image), the first was less influential insofar as the mean expected variability of all the stereometric parameters was lower for intrainage than for interimage evaluation. The difference was statistically significant in the case of both the intraobserver (\(P = 0.028\) for observer E; \(P = 0.031\) for observer M) and interobserver evaluation (\(P = 0.012\)). It should be noted that the mean expected variability in disc area was 0% in the case of the intraobserver interimage evaluation, because the first contour line was automatically superimposed on the next images (and thus theoretically did not affect their analysis); conversely, the mean expected variability in disc area ranged from \(\pm 0.72\%\) to \(\pm 1.28\%\) in the case of the intraobserver intrainage and interobserver evaluations, thus being extremely low. These results confirm the recent findings of Lester et al., who reported good agreement in the interobserver intrainage evaluation of disc area when four glaucoma experts independently drew their own contour line on the same HRT MTI (Kendall’s W coefficient = 0.92).

Reproducibility in the evaluation of HRT stereometric parameters has been tested in a number of previous studies, but these did not attempt to make a comprehensive assessment of the different sources of variability that may occur in a clinical setting. Our results are comparable with those of Azañera-Blanco et al. in terms of the intrainage evaluation and those provided by Hatch et al. in terms of the interobserver evaluation. We confirmed the previous findings of Orgül et al., who reported that the contour line tracing induced less variability than image acquisition/processing. The other previous studies used the coefficient of variation to evaluate reproducibility, which cannot appropriately be compared with the ICC we used.

The figures of the interimage evaluations, represent a common clinical situation in which the same observer may use the same contour line to analyze follow-up images or two or more different observers may use their own contour line for the same type of analysis. On the basis of these figures it seems possible to speculate that the “background noise” induced by test-retest variability is extremely low in the case of HRT stereometric parameters of those eyes believed not to have experienced any structural optic disc changes. This holds true for most of the parameters, the 95% confidence limits of variability of which are within 10% of the mean value.

New statistical analyses have recently been developed to assess longitudinal changes of the optic disc by HRT. Chauhan et al. have developed software that can adjust each follow-up image to the baseline image to detect subtle changes in the topography of each calculated “megapixel,” with the aim of avoiding the use of a standard reference plane (a potential major source of error) and standard stereometric parameters. This technique has greatly increased the specificity of detecting longitudinal optic disc changes. Although Chauhan’s approach is extremely interesting and valuable, other comparative methods for the longitudinal HRT follow-up of glaucoma are readily available, such as the direct comparison of baseline and subsequent follow-up stereometric parameters, which is simple and easy to perform, and allows investigators to work on mean topography images and standard parameters.

Greater care should be taken when acquiring the images, because this was the greatest source of variability in our HRT examination. On the basis of our figures relating to a series of 55 normal and glaucomatous eyes with a standard deviation of MTIs ranging from 11.5 to 66 \(\mu\)m, progressive changes of more than 10% of the baseline values may be considered statistically significant (at 95% confidence limit) for most of the standard stereometric parameters. To increase specificity, the variability of each patient should be taken into account and calculated using a simple test-retest analysis made over a short time span. Whether these statistically significant changes are also clinically significant can only be addressed by means of ongoing longitudinal clinical trials.

References