Optic Disc Progression in Glaucoma: Comparison of Confocal Scanning Laser Tomography to Optic Disc Photographs in a Prospective Study

Balwantray C. Chauhan,1 Donna M. Hutchison,1 Paul H. Artes,1 Joseph Caprioli,2 Jost B. Jonas,3 Raymond P. LeBlanc,1 and Marcelo T. Nicolela1

Purpose. To compare confocal scanning laser tomography (CSLT) to optic disc photography for detecting progressive glaucomatous disc changes.

Methods. Four expert observers first examined a training set of 40 pairs of monoscopic disc photographs to help set a criterion for change. They then classified 44 pairs of monoscopic photographs of patients followed prospectively as having either definitely no change, probably no change, probably change, or definitely change (score, 0, 1, 2, or 3, respectively). Observers were unaware that 10 pairs were photographs of fellow eyes obtained the same day to measure specificity. These patients were imaged with the Heidelberg Retina Tomograph every 6 months. Topographic change analysis (TCA) software with three criteria (yielding specificity of 81%, 94%, or 97%) was used. Comparisons between the TCA results and those of observers were made for a range of liberal to conservative criteria and for individual- and combined-observer scores.

Results. The median follow-up was 9.0 years with 18 CSLT images. Observer specificity varied from 50% to 90% (using score 0) and 60% to 100% (using score 0 or 1). Interobserver κ ranged from 0.22 to 0.38 from the most liberal to conservative criterion and was generally higher than corresponding TCA-observer agreement. For most criteria, observers had either a high hit rate with low specificity or high specificity with a low hit rate compared with the TCA. Similar results were obtained when observer scores were combined to derive a broader range of criteria.

Conclusions. The results indicate that the TCA performs at least as well as either the individual or best combination of observer classifications of disc photographs. (Invest Ophthalmol Vis Sci. 2009;50:1682–1691) DOI:10.1167/iovs.08-2457

Copyright © Association for Research in Vision and Ophthalmology

From the 1Department of Ophthalmology and Visual Sciences, Dalhousie University, Halifax, Nova Scotia, Canada; the 2Jules Stein Eye Institute, University of California Los Angeles, Los Angeles, California; and the 3Department of Ophthalmology, Faculty of Clinical Medicine, University of Heidelberg, Mannheim, Germany.

Supported by Grant MOP-11357 from the Canadian Institutes of Health Research.

Submitted for publication June 19, 2008; revised August 8, September 8, and October 20, 2008; accepted February 4, 2009.

Disclosure: B.C. Chauhan, Heidelberg Engineering (F); D.M. Hutchison, None; P.H. Artes, None; J. Caprioli, None; J.B. Jonas, None; R.P. LeBlanc, None; M.T. Nicolela, None

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Corresponding author: Balwantray C. Chauhan, Department of Ophthalmology and Visual Sciences, Dalhousie University, 2nd Floor Centennial Building, Capital Health District Authority, 1278 Tower Road, Halifax, NS, Canada B3H 2Y9; bal@dal.ca.

Confocal scanning laser tomography (CSLT) is a noninvasive technique for measuring surface topography of posterior pole structures. Since its introduction almost 20 years ago, CSLT has been used increasingly in ophthalmology, primarily for imaging the optic disc in glaucoma.

There are several diagnostic tools for cross-sectional analyses of single CSLT examinations that help clinicians discriminate between normal and glaucomatous optic discs. Depending on study samples, these tests generally yield a sensitivity of approximately 70% at fixed specificities of 90% to 95%. This relatively low sensitivity is a result of the large variation in optic disc appearance in the normal and glaucomatous populations resulting in an overlap in the measured parameters. The ability of glaucoma expert clinicians examining conventional disc photographs to discriminate between normal and glaucomatous discs is comparable to CSLT and, therefore, the task of distinguishing a glaucomatous disc in a single observation, irrespective of the mode of examination, is hampered by the variation in optic disc appearance. Since the intraindividual retest variability of CSLT can be accounted for either in event-based or trend-based serial analyses and because images can be obtained relatively quickly and easily, the greatest utility of CSLT is probably in the determination of changes in the optic disc over time.

Several reports on the ability of CSLT to detect glaucomatous optic disc changes have been published; however, most have compared findings to visual field test results. There are fewer reports that compared serial CSLT findings to serial optic disc photographs. In experimentally induced intraocular pressure (IOP) changes to the optic nerve in monkeys, CSLT detected serial changes and performed at least as well as glaucoma specialists who evaluated disc photographs. In two clinical studies, reasonable agreement was found between CSLT images analyzed with the Topographic Change Analysis (TCA) program of the Heidelberg Retina Tomograph (HRT; Heidelberg Engineering GmbH, Heidelberg, Germany) and disc photography. However, neither of these studies provided adequate control for parameters such as TCA specificity, appropriate criteria for progression, or observer specificity. Furthermore, only the size but not depth of disc change was considered.

The purpose of this investigation was to evaluate the ability of TCA to detect optic disc progression in patients with glaucoma observed longitudinally with both CSLT and conventional disc photography. We used various criteria for both the TCA and expert observer evaluation and provide comparisons with the observers at various degrees of TCA specificity.

Methods

Patients

The patients in this study were drawn from a longitudinal prospective study on glaucomatous visual field and optic disc progression. They...
were recruited consecutively from the glaucoma clinics of Capital Health District Authority (Halifax, Nova Scotia, Canada), with the following inclusion criteria: (1) clinical diagnosis of open-angle glaucoma with notching or progressive thinning of the neuroretinal rim, typically recorded photographically; (2) baseline visual field with a mean deviation index between −2 and −10 dB; (3) open angles by gonioscopy, and (4) best corrected visual acuity ≥ 6/12. Exclusion criteria were (1) concomitant ocular disease; (2) systemic disease or systemic medication known to affect the visual field; (3) refractive error exceeding 5 D (equivalent sphere) of myopia or hyperopia or 3 D of astigmatism; and (4) contact lens wear.

In accordance with the Declaration of Helsinki, all patients gave informed consent to participate in the study, and the protocol was approved by the Capital Health Research Ethics Board.

Testing Protocol
The full testing protocol has been described elsewhere, and only the pertinent tests for this report will be described. At baseline and every 6 months thereafter, CSLT was performed with the HRT 1. At each session, several images with the optic disc centered in the image frame were obtained with the 10° scan angle. After the image quality was examined carefully for evenness in illumination, optic disc centration, and fixation stability, three of the best-quality images were used to compute mean topography. Mean topography was not computed for inclusion in the study if the image quality was inadequate. A contour line was drawn in the baseline image and automatically imported to the follow-up images for analysis using HRT 3 software (Heidelberg Engineering GmbH). The quality of alignment of the longitudinal image series was checked with the movie feature of the software. If there were misaligned images, they were removed. Fewer than 5% of the images were removed for misalignment.

Optic disc photography was conducted throughout the follow-up at various intervals and at the last patient visit. Monoscopic color slides were obtained with a fundus camera (FF3; Carl Zeiss Meditec, Dublin, CA) and 64 ASA film (Ektachrome; Eastman Kodak, Toronto, ON, Canada).

Data Sets
The disc photograph closest in time to the baseline CSLT examination and the last photograph were coded. Only patients who had a difference in baseline CSLT and disc photograph of less than 3 months were included in the study. Poorly focused or exposed photographs were excluded. From a group of 40 patients with available disc photographs, 34 eyes of 34 patients met these criteria and were included in the study. In addition, two photographs of 10 fellow eyes taken on the same day were included to estimate observer specificity. The observers were not aware of this addition and were told there were 44 pairs of photographs of one eye of 44 patients at baseline and last visits to evaluate. This set of photographs was termed the evaluation set.

Another set, termed the training set, of 40 pairs (baseline and last) of optic disc photographs used for the preparation of the Reading Centre for the European Glaucoma Prevention Study20 was also used. As this set contained stereo disc photographs, only one photograph per eye was included. The classifications (change or no change obtained by consensus from another panel of expert observers) were provided. The use of the training set was an attempt to establish a criterion for progression when assessing the evaluation set and minimize the interobserver variability.

The sequence of CSLT examinations with the closest overlap in time with the disc photographs in the evaluation set were identified for analysis with the TCA software.

Expert Observer Classifications
The expert observers were four glaucoma specialists (observers A–D) who had independently reviewed the training set. The observers then assessed the evaluation set, and for each of the 44 pairs indicated, assigned a score of 0 (definitely no change), 1 (probably no change), 2 (probably change), or 3 (definitely change). They were also asked to indicate the location where change had occurred (for the right optic disc: superior-temporal: 90–135°; temporal: 135–225°; inferior-temporal: 225–270°; inferior-nasal: 270–315°; nasal: 315–45°; and superior-nasal: 45–90°).

Topographic Change Analysis
The TCA is an event-based technique for detecting topographic changes in the optic nerve head used in conjunction with CSLT and the HRT software. The basic TCA methodology has been described elsewhere, as have the performance characteristics of an earlier version of the software. Briefly, the 256 × 256-pixel array of topographic height values are divided into a 64 × 64-superpixel array (in later versions of the HRT, the arrays are 384 × 384 and 96 × 96, respectively). An analysis of variance is performed to compare each follow-up examination to the baseline examination. Probability symbols (red and green, respectively, for locations with significantly decreasing and increasing topographic heights) are overlaid on the reflectance images, indicating locations with statistically significant change. In the latest version of the TCA software, height changes in the significant locations are also color coded for depth, with the darkest red or green superpixels showing locations with the largest depth changes.
Statistical Analyses

A range of liberal (more sensitive and less specific) to conservative (less sensitive and more specific) criteria for progression with both the TCA and disc photographs were used to compare performance characteristics of these two imaging modalities. These criteria are summarized in Figure 1.

For the optic disc photographs, the three criteria were liberal: probably no change, probably change, or definitely change (score, 1, 2, or 3); moderate: probably change or definitely change (score, 2 or 3); conservative: definitely change (score, 3). Hence for the most liberal criterion, all classifications with the exception of definitely no change (score, 0) were considered a classification of progression (Fig. 1).

For the TCA, criteria for progression took into account the size (as a percentage of the optic disc area) and depth of the largest cluster of superpixels with statistically significant change within the optic disc. The change had to occur in at least three of four consecutive follow-up examinations. The liberal criterion required a cluster of $\geq 0.5\%$ of the disc area and depth of $\geq 20\, \mu m$; the moderate criterion a cluster of $\geq 1\%$ of the disc area and depth of $\geq 50\, \mu m$; and the conservative criterion a cluster of $\geq 2\%$ of the disc area and depth of $\geq 100\, \mu m$. In a group of 34 eyes of 34 normal control subjects who have been followed identically with CSET, the specificity (i.e., 1 – [incidence of progression or false-positive rate]) at 5 years is 81% (95% confidence interval [CI], 67.8%–94.2%) for the liberal criterion, 94% (95% CI, 86.0%–100%) for the moderate criterion, and 97% (95% CI, 91.3%–100%) for the conservative criterion (Artes PH, et al. IOVS 2006;47: ARVO E-Abstract 4349).

The agreement between the TCA and observer scores were computed for the different criteria. The interobserver $\kappa$ agreement statistics were compared to the observer-TCA $\kappa$ agreement. Since specificities were available for both the TCA and observers, for each technique the proportion of optic discs identified as progressing (or hit rate) was evaluated as a function of specificity.

RESULTS

The median (interquartile range, IQR) baseline age of the 34 patients was 57.6 (50.2, 61.9) years. The median (IQR) follow-up was 9.0 (7.8, 10.1) years with 18 (15, 20) HRT images. The median (IQR) absolute time difference between the base-
line HRT examination and baseline disc photograph was 0 (0, 70) days, and the last HRT and last disc photograph was 0 (0, 1) days. The median (IQR) baseline age of the 34 normal control subjects used to derive the TCA specificity estimates was 56.0 (45.4, 64.5) years. They were followed for 10.2 (7.7, 10.8) years with 14 (11, 17) HRT images. The median (IQR) mean pixel height standard deviation (MPHSD) of the HRT images was 21 (16, 27) μm in patients and 19 (14, 29) μm in control subjects. There were no statistically significant differences between patients and control subjects in the baseline age (P = 0.52, Mann–Whitney test) or length of follow-up (P = 0.08); however, patients had more examinations (P = 0.01) and a higher MPHSD (P = 0.01).

Observer scores for the 10 optic disc photographs used for the specificity estimates are shown in Figure 2. There were three patients (cases 6, 7, and 9; Fig. 2) in whom all four observers scored 0 (definitely no change) and 9 patients in whom at least two observers scored 0 (all except case 3, Fig. 2). A score of 0 (definitely no change) yielded specificities ranging from 50% (observer B) to 90% (observer A). A score of either 0 or 1 (definitely no change or probably no change) yielded specificities ranging from 60% (observer B) to 100% (observer A).

For the remaining 34 photographs in the test set, the agreement between observers was fair according to Landis and Koch’s descriptors with κ values of 0.22 (95% CI, 0.18–0.26) when a score of 1, 2, or 3 was used for defining progression, 0.24 (95% CI, 0.11–0.38) when a score of 2 or 3 was used, and 0.38 (95% CI, 0.34–0.43) when a score of 3 was used.
The agreement rates between the TCA and observers depended on the criterion used to define progression. For the moderate TCA criterion, the overall agreement decreased with increasingly conservative observer criteria (Fig. 3). Agreement was highest for observer B who classified the highest number of discs as progressing and lowest for observer A who classified the lowest number of discs as progressing. For the conservative TCA criterion, the overall agreement generally increased with increasingly conservative observer criteria (Fig. 4) and was highest for observer A and lowest for observer C. The $\kappa$ statistics of agreement between the TCA and observers are shown in Figure 5 for the moderate and conservative TCA criteria. These data indicate that the agreement between the TCA and observers was generally poorer than that between the observers (Fig. 5); however, there was at least one observer criterion where the TCA-observer agreement was close to or better than the equivalent interobserver agreement. In some instances, $\kappa$ was $<0$, indicating worse than chance agreement between the TCA and observers.

Illustrative cases (Figs. 6, 7, 8) show the spectrum of agreement between the two methods. For case 1 (right eye), a 54-year-old patient at baseline followed for 10 years, all four observers gave a classification of definitely change (score, 3) with all observers noting changes in the superior temporal, temporal, and inferior temporal sectors of the disc (Fig. 6). The TCA showed progression with the conservative criterion with increasing changes throughout the disc (Fig. 6 and Movie S1; all movies are online at http://www.iovs.org/cgi/content/full/50/4/1682/DC1). In case 2 (left eye), a 58-year-old patient at baseline followed for 8 years, there were two observer classifications of probably no change (score, 1) and two of definitely change (score, 3) in the superior and inferior temporal quadrants (Fig. 7). The TCA showed progression with the conservative criterion with significant changes
in the superior temporal, superior nasal, and inferior temporal quadrants (Fig. 7 and Movie S2). Finally for case 3 (right eye), a 45-year-old patient at baseline followed for 11 years, there was one classification of definitely no change (score, 0), two of probably change (score, 2), and one of definitely change (score, 3) in the temporal and inferior temporal sectors (Fig. 8). There was significant progression with the conservative TCA with overall widening of the cup with additional surface changes in the neuroretinal rim (Fig. 8 and Movie S3).

The TCA hit rates (proportion of progressing cases) for the three criteria—liberal (specificity, 81%), moderate (specificity, 94%), and conservative (specificity, 97%)—were 94%, 77%, and 35%, respectively (Fig. 9). Each observer's hit rates for the three criteria used—probably no change, probably change, or definitely change (score, 1, 2, or 3); probably change and definitely change (score, 2 or 3); and definitely change (score, 3)—are also shown in Figure 9. These data show substantial differences among the observers. For example, observer A yielded high specificity, but even the most conservative TCA criterion yielded a higher hit rate with a slightly reduced specificity for the two more conservative observer criteria. Although observer B yielded comparable hit rates to the TCA, the specificity was notably poorer. The responses of observers C and D were between those of observers A and B.

Using combined responses from the observers allowed comparison with the TCA over a wide range of criteria, from the most liberal (progression defined as score of 1, 2, or 3 from only 1 observer) to the most conservative (progression defined as a score of 3 from all four observers). These data show that for each of the three observer criteria, there was one of four possible combined responses where the observer performance was similar or better than that of the TCA; however, in all other cases, the observer specificity and hit rates were lower than those of the TCA (Fig. 10).

**FIGURE 7.** Left optic disc of a 58-year-old patient at baseline, who was observed from July 1997 to September 2005 with optic disc photography (top). Two expert observers gave a classification of probably no change (score, 1) and the other two of definitely change (score, 3). The TCA (bottom) of the follow-up images showed progression with the conservative criteria with significant changes in the superior temporal, superior nasal, and inferior temporal quadrants. Movie S2, http://www iovs.org/cgi/content/full/50/4/1682/DC1, loops between the aligned baseline and final HRT reflectivity images.
DISCUSSION

CSLT is an important technique for imaging the optic disc in glaucoma whose greatest potential is in the detection of changes in disc topography over time. Most of the published experience to date with CSLT deals with cross-sectional analysis for case detection,\(^3,4,6,7,28–30\) prediction of visual field progression from ocular hypertension,\(^31\) or correlation of CSLT changes with visual field progression.\(^19,16\) The present study adds to the small number of published reports comparing the performance of CSLT with longitudinal analysis methods with that of conventional disc photography.\(^18,21,22\) This comparison is likely most relevant as validation of an imaging technique with visual field progression as the reference (or gold) standard is probably less pertinent because of the substantial independence of identified visual field and optic disc progression during the relatively short time periods of clinical studies compared with disease duration.\(^19,16\) We compared the robustness of the techniques for identifying progression across a range of liberal to conservative criteria.

The agreement between expert observers determining optic disc progression with conventional photography varies among the published reports.\(^21,26,32–34\) In our study, the agreement among the four observers making independent classifications was only fair according to the descriptors proposed by Landis and Koch.\(^27\) There was a large variation in the observer specificity; however, there was a relationship between specificity and hit rate. For example, observer A had high specificity, but low hit rate, whereas observer B had low specificity, but a high hit rate. This finding underscores the importance of individual observer criteria for making decisions on progression. However, despite using a training set before assessing the evaluation set to attempt standardization of the criterion for progression, this criterion varied among the observers.

**Figure 8.** Left optic disc of a 45-year-old patient at baseline, who was observed from July 1995 to June 2005 with optic disc photography (top). One expert observer gave a classification of definitely no change (score, 0), two of probably change (score, 2), and one of definitely change (score, 3). The TCA (bottom) of the follow-up images showed progression with the conservative criteria with overall widening of the optic cup and depression of the neuroretinal rim surface height. Movie S3, http://www.iovs.org/cgi/content/full/50/4/1682/DC1, loops between the aligned baseline and final HRT reflectivity images.
The agreement between the TCA and observer evaluation of disc photographs depended critically on criteria and observer. With the moderate TCA criterion, more cases were classified as progressing with CSLT alone. The number of cases progressing with disc photographs alone decreased as the observer criteria for progression became more conservative (Fig. 3). Similarly, with the conservative TCA criterion, more cases were classified as progressing with disc photography when the most liberal criterion was used; however, with more conservative criteria, this number decreased (Fig. 4). Quantitatively, the agreement between observers was generally better than the agreement between the TCA and observers (Fig. 5), though these results also depended on the criteria used.

The use of a single criterion for optic disc progression in glaucoma may be restrictive and overly simplistic, given the absence of a sufficiently independent and valid reference standard for progression. This approach was a limitation of previous research in this area.\(^\text{18}\) In the present study, we used a cumulative scale from liberal to conservative to obtain a more global assessment of performance. In addition, we used approaches that combined responses from the observers; hence, the most liberal criterion for progression required a classification of probably no change, probably change, or definitely change from only one observer, whereas the most conservative criterion required a classification of definitely change from all four observers. Because we had specificity estimates for the TCA and observers, we were able to directly compare performance characteristics of the two methods without reliance on a reference standard (Figs. 9, 10). These data show that combining responses from the observers may yield better performance than individual observers alone. We did not seek arbitration in cases of disagreement and therefore it is unknown whether consensus agreement or a combination approach would lead to better performance characteristics. Our results indicate that the TCA may have comparable, and in many cases better, combination of hit rate and specificity compared to the observers.

Patients had a relatively long follow-up (median, 9 years) with frequent CLST examinations (median, 18 examinations). There was near complete overlap in follow-up time with CLST and disc photography; therefore, the two techniques could be compared accurately. However, this study has several important limitations. The disc photographs were monoscopic. Although studies comparing observer agreements judging progression using monoscopic and stereoscopic photographs have not been published, observer agreement rates in a one-time assessment of glaucomatous discs was reported to be higher with stereo photographs in one study\(^\text{35}\) but not in another.\(^\text{36}\) A recent study of agreement among three expert observers using stereo disc photographs\(^\text{37}\) yielded generally poorer agreement than that reported in this study. However, it is possible that had stereo disc photographs been available, our observer agreement, hit rates, and specificities would have been higher. We were unable to determine whether changes with CSLT were detected earlier than disc photography as the experimental protocol would have required photography at the same time intervals as CLST examinations. Observers were provided with only the first and last photographs in the follow-up. For the TCA, specificity estimates were derived from hit rates in a parallel group of normal subjects with identical follow-ups. Alternative estimates could have been derived from test–retest values in images from patients obtained several times over the same day or from a group of stable glaucoma population examined over a relatively short period. Evidence in normal subjects suggests that the test–retest variability estimates with these two approaches is comparable in normal subjects\(^\text{38}\); however, more comprehensive study in patients with glaucoma is required. Permutation analyses of follow-up data where the order of examinations is randomized to simulate variability\(^\text{19,15}\) has also been described for the this purpose. Each of these estimates of specificity have strengths and limitations that deserve further study. We elected to take advantage of data from the parallel normal cohort, as it was felt to be the data set with the least number of assumptions. Normal
to determine whether the changes detected by CLST have subsequent functional consequences.

References


