Glaucomatous Progression in Series of Stereoscopic Photographs and Heidelberg Retina Tomograph Images

Neil O’Leary, MSc; David P. Crabb, PhD; Steven L. Mansberger, MD; Brad Fortune, PhD; Michael D. Twa, PhD; Michael J. Lloyd, MD; Aachal Kotecha, MD; David F. Garway-Heath, MD; George A. Cioffi, MD; Chris A. Johnson, PhD

Objective: To compare optic disc changes using automated analysis of Heidelberg retina tomograph (HRT) images with assessments, by glaucoma specialists, of change in stereoscopic photographs.

Methods: Baseline and follow-up stereophotographs and corresponding HRT 1 series of 91 eyes from 56 patients were selected. The selection criteria were sufficiently long, good-quality HRT series (7 visits in ≥70 months of follow-up) and follow-up photographs contemporaneous with the final HRT image. Topographic change analysis (TCA), statistic image mapping (SIM), and linear regression of rim area (RALR) across time were applied to HRT series. Glaucomatous change determined from stereophotographs by expert observers was used as the reference standard.

Results: Expert observers identified 33 eyes (36%) as exhibiting glaucomatous change. Altering HRT progression criteria such that 36% of eyes progressed according to each method resulted in concordance between HRT methods and stereophotograph assessment of 54% for TCA, 65% for SIM, and 67% for RALR (Cohen κ = 0.05, 0.23, and 0.30, respectively). Receiver operating characteristic curves of the HRT analyses revealed poor precision of HRT analyses to predict stereophotograph-assessed change: areas under the curve were 0.61 for TCA, 0.62 for SIM, and 0.66 for RALR.

Conclusions: Statistical methods for detecting structural changes in HRT images exhibit only moderate agreement with each other and have poor agreement with expert-assessed change in optic disc stereophotographs. Arch Ophthalmol. 2010;128(5):560-568
The aim of this study was to examine change in HRT image series identified by 3 automated statistical analytical methods: topographic change analysis (TCA), statistic image mapping (SIM), and ordinary least squares linear regression of rim area (RALR) against time of follow-up. We compared these methods with assessments by glaucoma specialists in optic disc stereophotographs from the same eyes and sought to determine which method had the highest concordance with expert assessment of stereophotographs. To control the sample specificity and sensitivity of HRT change detection analyses, criteria for change were varied in stringency.

METHODS

PATIENTS

Data from the Devers Eye Institute Perimetry and Psychophysics in Glaucoma study were used, and details of the investigation have been previously published. All of the patients provided voluntary written consent to participate and to allow their clinical measurements to be securely held for future data analysis. All of the procedures adhered to the tenets of the Declaration of Helsinki and were approved by the local ethics committee. Participants were recruited prospectively from the Devers Eye Institute or other ophthalmic practices in the Portland, Oregon, metropolitan area. At recruitment, all of the patients were considered to have either high-risk ocular hypertension or early glaucoma. All of the patients had a history of untreated intraocular pressure of at least 22 mm Hg in both eyes and at least 1 additional risk factor: a vertical cup-disc ratio of at least 0.6 in at least 1 eye or interocular cup-disc ratio asymmetry of at least 0.2; a positive family history of glaucoma; a personal history of migraine, Raynaud syndrome, or vasospasm; African-American ancestry; or age older than 70 years. All of the patients met the following criteria for both eyes: best-corrected visual acuity of 20/40 or better and spectacle refraction within ±3.00 diopter (D) sphere and ±2.00 D cylinder and reliable standard-automated perimetry results with mean deviation better than or equal to –6 dB. Patients were excluded if they had any other previous or current ocular or neurologic disease, previous ocular surgery (except uncomplicated cataract surgery), or diabetes mellitus requiring medication.

SELECTION CRITERIA

Data from an initial data set of both eyes of 168 patients (336 eyes) with follow-up of at least 4 years (median, 6.1 years) were evaluated. Figure 1 illustrates the selection criteria from this initial data set with the numbers of patients and eyes in the study as selection criteria were applied.

STEREO OPTIC DISC PHOTOGRAPHY AND GRADING

Photographs were obtained annually for all of the patients using a simultaneous stereoscopic camera (3-Dx; Nidek Co Ltd, Gamagori, Japan) after maximum pupil dilation. For each eye, the photographs obtained at baseline and at the most recent follow-up visit were randomly assigned to be labeled as A or B to mask the temporal order. All other information about the eye and the patient was masked from the graders, including the appearance of the fellow eye. Two fellowship-trained glaucoma specialists independently viewed the baseline and final follow-up photographs using a Stereo Viewer II (Asahi-Pentax, Tokyo, Japan) and graded them as “changing” or “stable,” indicating which photograph showed worse damage (A or B). If there was change, the type of change was recorded as 1 or more of the following: increased neuroretinal rim narrowing, increased excavation, new or increased retinal nerve fiber layer defect, or new notching. The location of change was recorded in 90° sectors as follows: (0°, 90°], (90°, 180°], (180°, 270°], and (270°, 360°]. Quality assessments of each image pair were recorded separately for clarity and for stereopsis as “excellent,” “adequate,” or “unacceptable.”

The reviewers mediated disagreements by reexamining the photographs together to reach a consensus; any continuing disagreements between these 2 graders were adjudicated by a third masked expert (S.L.M. or G.A.C.). Change identified in the correct temporal direction (ie, the follow-up photograph graded as worse) was labeled “true” (glaucomatous change). Change identified in the “wrong” temporal direction (ie, the baseline photograph graded as worse) was labeled as “false progression.”

Sample specificity, sample sensitivity, and the reproducibility of the assessment method were estimated by presenting the graders with 3 additional sets of photographs:

1. A second set of stereophotographs obtained on the same day of a subset of 10 cases from the larger study cohort were presented again. These 10 cases were randomly assigned a unique identification number and were inserted into the study set. Sample specificity (ie, the rate of correctly identifying no change) was defined as the proportion of these 10 eyes that the graders determined to have remained stable.

2. Two glaucoma experts (S.L.M. and G.A.C.) selected 10 examples of “definite” glaucomatous change from their private practices that were separate from the study cohort. Temporal or-

Figure 1. Selection criteria applied and the resulting sample sizes (patients and eyes) at each stage of selection. HRT indicates Heidelberg retina tomography.
Downloaded From: by a Non-Human Traffic (NHT) User on 11/01/2018

The graders were unaware that these 30 cases were not part of the study cohort.

**CONFOCAL SCANNING LASER OPHTHALMOSCOPY**

The CSLO images were obtained using the HRT Classic. Scans of angle $10^\circ \times 10^\circ$ centered on the optic disc were acquired, and the 3 best-quality images were combined to create a mean topography for each eye. Experienced technicians outlined the optic disc margin. Images were analyzed using the latest available software (version 1.6.1.0) but were not imported to HRT III software. The manual landmarking facility was used to correct obvious failures of the automatic alignment algorithm to adequately register images across time.

**STATISTIC IMAGE MAPPING**

SIM\(^9\) derives the significance of change at each pixel in the image by comparing the actual rate of change in height across time (proportional to the variability) with all the possible rates of change derived from 1000 random permutations of the data. The significance of a cluster of significantly changing (active) pixels is similarly obtained by comparing the observed maximum cluster size with the size of maximum clusters generated in the random permutations.

**TOPOGRAPHIC CHANGE ANALYSIS**

The primary method for assessing change using HRT software is TCA, a technique that compares the topographic height variability at superpixels (4 $\times$ 4 pixels) in a baseline examination with the height change between baseline and follow-up examinations.\(^20,22\) A change map of P values, indicating the probability of change at each superpixel, is created, and contiguous superpixels showing significant (P < .05) decreases in retinal height are clustered, thus allowing the generation of various TCA change summary variables describing the size and location of regions of change. Change across time is confirmed by comparing the most recent follow-up examination findings with those of the previous 2 examinations.

**LINEAR REGRESSION OF RIM AREA**

Previous works\(^30,32\) have shown that changes to rim area (RA) are likely to be a good measure of glaucoma progression, and studies\(^24,26\) on the development of RA progression analyses and their performance have been published. The RA analysis for this study was trend based. Global RA and RA for the 6 predefined sectors across time were analyzed using ordinary least squares linear regression of rim area (RALR). Areas under the curve are 0.62 for TCA, 0.61 for RA, and 0.66 for RALR.

**Figure 2.** Receiver operating characteristic curves for Heidelberg retinal tomograph progression algorithms using stereophotograph-assessed glaucomatous change as the reference standard for topographic change analysis (TCA), statistic image mapping (SIM), and ordinary least squares linear regression of rim area (RALR). Areas under the curve are 0.62 for TCA, 0.61 for RA, and 0.66 for RALR.

**ANALYSIS: MEASURES OF CHANGE AND SLIDING SCALE**

For SIM, the measure of change was the probability value of the largest cluster of “active” (red) pixels. In TCA, the measure of change was the area of largest cluster of red superpixels as a percentage of disc area. In the case of RA, the measure of change was the lowest P value (most significant) obtained by LR of the 7 (global and 6 sector) RA linear trends. Using expert-assessed stereophotographs as the reference standard, the aim was to vary the criteria for change for each method of HRT change analysis and compare proportions identified as changing. Thus, receiver operating characteristic (ROC) curves were generated to measure the diagnostic precision of each HRT change analysis method in predicting glaucomatous optic disc changes assessed on stereophotographs. Agreement between HRT methods and stereophotograph change, at equal rates of glaucomatous progression classification, was examined and illustrated using area proportional Venn diagrams. This entailed fixing discriminant criteria to classify the same number of eyes as changing in HRT analyses as in the stereophotograph assessment. The analysis was performed in MATLAB (release R2007a; The Mathworks Inc, Natick, Massachusetts).

**RESULTS**

Ninety-one eyes of 56 patients from the original 336 eyes of 168 patients met the chronological and quality criteria (Figure 1). Measurements from 7 annual mean HRT scans (composed of 3 single HRT scans) for each eye in this study were used for analysis. Mean patient age at base-
line was 56 years (age range, 35-82 years), and the male to female ratio was 52:48. The racial mix was as follows: 54 (96%) white, 1 Hispanic, and 1 American Indian.

In the patient data set, 33 eyes (36%) were assessed as exhibiting glaucomatous change using the stereophotograph reference standard. In 47 of 91 instances (52%) the assessment required adjudication by the third grader (Table). The mean interval between baseline stereophotograph and baseline HRT scan acquisition was 8 days, and the mean interval between follow-up stereophotograph and final follow-up HRT scan acquisition was 11 days.

Figure 2 shows the ROC curves for TCA, SIM, and RALR. Areas under the ROC curve (95% confidence intervals [CIs]) are as follows: 0.61 (0.56-0.66) for TCA, 0.62 (0.57-0.67) for SIM, and 0.66 (0.61-0.71) for RALR. Using the method of Hanley and McNeil to compare areas under the ROC curve resulted in \( P = .79 \), .26, and .24 for pairwise comparisons of TCA/SIM, TCA/RLR, and SIM/RLR, respectively. At a fixed specificity of 90% for all 3 methods, sensitivities were 25% for TCA, 27% for SIM, and 40% for RALR.

**Figure 3** shows the agreement of TCA-, SIM-, and RALR-identified change with stereophotograph change after rates of identified progression are matched to those of the stereophotograph assessment (36%). Concordances were 54% for TCA, 65% for SIM, and 67% for RALR, with Cohen \( \kappa \) values of 0.05, 0.23, and 0.30, respectively. **Figure 4** shows the agreement among the HRT change detection methods at equal rates of identified change (36%). This reveals concordance among the HRT change detection methods to be 60% and pairwise concordance to be 71% to 76%.

**Figures 5, 6, 7, and 8** show 4 cases to illustrate different levels of agreement between HRT analyses and stereophotograph assessment when criteria for progression are fixed for equal classification rates. The stereophotograph decisions were reached by consensus in cases 1 and 4 but required adjudication in cases 2 and 3.

Of the same-day stereophotograph set, 2 of 10 eyes were judged to be changing by graders, giving sample specificity of 80% (95% CI, 44%-98%). Of the definite glaucomatous change set, 8 of 10 eyes were judged to be changing. Of the repeated set, 2 of 10 stereophotograph pairs resulted in different assessments on repeated presentation. Thus, sample sensitivity and reproducibility were both estimated to be 80% (95% CI, 44%-98%).

**COMMENT**

The CLSO, as typified by the HRT, has been shown to give a repeatable measure of optic disc structure. The HRT does reasonably well in distinguishing glaucomatous eyes from healthy eyes, but the real promise of the technology may be in offering a reliable method for tracking structural change, potentially providing useful clinical management information about disease stability. A method for quantifying change is required to realize this potential, and there has been much research...
activity in developing an appropriate technique, but there is little evidence to suggest that one method is better than another.

Studies using functional progression (visual field deterioration) are confounded by aspects of the relationship between structural and functional changes. We have little idea of the relative proportions of associated and independent behavior, and the temporal sequence of structural and functional glaucomatous change is not well defined. Because change identified by structure and function do not seem closely related, we postulated that progression identified by glaucoma experts from optic disc stereophotographs would provide a better reference standard against which to assess the performance of another structural measurement for progression (CSLO images).

This study is one of few examining agreement of HRT change analyses with expert-assessed stereophotographs. We considered a variety of statistical methods for detecting change in HRT images, and we varied the stringency of the criteria for HRT change to give a measure of sample sensitivity across the full range of sample specificity. This study took advantage of data from a carefully collected prospective longitudinal study across a relatively long period, and strict image quality criteria were applied. Previous research has shown better between-grader agreement and better reproducibility from stereophotographs than from monoscopic photographs when discriminating between glaucomatous and healthy discs. The estimate of reproducibility for expert-assessed progression in stereophotographs in this study, at 80%, is comparable with that of previous studies that have obtained κ values of 0.62 to 0.89 and 0.80 to 1.00 for within-observer reproducibility.

The ROC analysis suggests that when using stereophotographs as the reference standard, automated HRT methods have only moderate precision to predict change. The ROC curves revealed poor sample sensitivities for clinically relevant regions of high sample specificity. At equal rates of classification, poor agreement was found between glaucoma expert–assessed stereophotographs and the HRT analyses. Stereophotographs and HRT images are assumed to give an accurate and repeatable measure of the structure of the optic disc, but the false-positive and false-negative rates of the reference standard and the HRT methods may largely explain this poor agreement. This is illustrated in cases 1 to 4 (Figures 5-8). The false-positive rate of photograph grading from this study (estimated to be 20%) is a major factor. This rate falls in the range of previous studies (0%-50% and 5%-77% depending on stringency of progression criteria and observers). Furthermore, grader agreement, even between

Figure 5. Case 1. Single baseline (April 1998) (A) and single final follow-up (April 2005) (D) photographs from stereophotograph pairs, with excavation and rim narrowing indicated superotemporally and superonasally (arrows). B, Baseline Heidelberg retina tomograph (HRT) mean image (April 1998). Final follow-up HRT mean images (April 2005) with topographic change analysis (progression flagged) (C) and statistic image mapping (progression flagged) (E) output (the dark red pixels are the largest cluster of pixels in the disc). F, Output for linear regression of rim area (red sectors: significant P values for negative trend of rim area).
“experts,” in other tasks, such as separating healthy and unhealthy optic discs and assessing progression events in series of optic disc stereophotographs or visual fields, is not good. Owing to the nature of the data set (generally early glaucoma), there is likely to be a wide range in the magnitude of changes, and agreement may be much worse when changes are of small magnitude, partly evidenced by the 52% of stereophotograph assessments that required adjudication by a third expert.

Further reasons for the difference may be that features sometimes implicitly attributed to glaucomatous change in stereophotographs (eg, color changes) may not be apparent in HRT images, which are simply estimates of the topographic height of the optic disc surface and surrounding areas. Other contributing factors are that certain optic disc configurations (such as hypoplastic and tilted discs) may present greater difficulty for either HRT analysis (image registration, contour line fitting, and RA calculation) or stereophotograph assessment.

Concordance among all 3 statistical methods for detecting HRT change was shown to be 60%, and pairwise concordance was 71% to 76%. It is not surprising that agreement among HRT analyses was better than agreement between stereophotograph assessment and HRT analyses. Differing statistical methods, such as those used in this study, will never have perfect concordance, even on the same data. However, the level of disagreement between HRT analysis methods is probably amplified by the low stringency of the change criteria (fixed to identify equivalent 36% proportions) and the related (likely) high false-positive rate. These criteria are less stringent than those used in previous studies.

The results of this study expose the limitation of using grader-assessed stereophotographs alone as a reference standard for structural glaucomatous progression. Future studies to assess the effectiveness of HRT change analysis methods may require a more innovative approach to establishing a reference standard. Using an accumulation of information from a variety of measurements (fields, intraocular pressure, and optic disc) and presenting this as a continuous scale for evidence of change might be useful. Another approach might be to simulate series of images with known properties using a virtual platform in which the CSLO image formation process and software treatment of images is simulated computationally. Automated analysis of the reconstructed optic disc from stereophotographs may also be useful. Imaging devices, such as spectral domain optical coherence tomography, which may give surrogate measurements closer to what is really required to detect real glaucomatous change, may also help refine reference standards for progression.
There were some limitations to this study. Varying criteria for progression in stereophotograph assessment would have been useful in determining various diagnostic strengths of the HRT analyses but may also have led to even less agreement between graders. The setting of a cutoff criterion to determine progression is not trivial: because the true specificity of a criterion cannot be known, setting the cutoff value to result in equal rates of progression provides an opportunity to compare the agreement between the HRT change detection methods when the “hit rate” is the same. Times to progression were not examined, in part due to stereophotographs being assessed only at baseline and final follow-up. Therefore, we do not have any estimates as to which HRT analyses detected change earliest, although one could question the value of such analyses when agreement between methods is poor. The number of eyes in the definite glaucomatous change stereophotograph set and the same-day (no glaucomatous change) set were low, resulting in large CIs for the estimates of sample specificity and sensitivity for the expert-assessed change. However, as already discussed, these estimates are similar to those of previous studies. Because only “depressed” change was examined in HRT topographies, glaucomatous change resulting in elevation of the optic disc surface (if it occurs) would have been overlooked. Moreover, the ability to accurately detect changes, or lack of changes, in HRT longitudinal image series will depend on the ability of the software to register the images appropriately, and, thus, the present results are limited by the constraint of the HRT alignment algorithm. Only HRT I (10° × 10°) images were available in the study, which, although wide enough to contain the optic disc in all cases, align less well in the latest software than do 15° × 15° HRT images.

In conclusion, this study revealed poor agreement between progression detection using a variety of HRT statistical methods and expert-assessed stereophotographs of the optic disc. Using stereophotograph-assessed change as the reference standard does not help determine which HRT change algorithm best identifies glaucomatous change in this group of patients with high-risk ocular hypertension and those with early glaucoma. This does not imply that stereophotographs are not integral to the assessment of glaucomatous change. Indeed, they are a clinically well-accepted standard that have been used in major clinical trials. However, the diagnostic precision associated with observer stereophotograph-assessed change precludes it from being a stand-alone benchmark by which to evaluate alternative change detection tools. The extent to which the HRT measures real change across time has yet to be established. However, the practical benefits of being able to observe change using automated or semiautomated digi-
tal image analysis, and other recent evidence, suggest that it is an important tool for assessing disease progression, especially if a statistical method for best detecting the change can be established.

Submitted for Publication: June 6, 2009; final revision received July 31, 2009; accepted August 26, 2009.

Correspondence: Neil O’Leary, MSc, Department of Optometry and Visual Science, City University London, Northampton Square, London EC1V 0HB, United Kingdom (oleary.neil@gmail.com).

Author Contributions: Mr O’Leary had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Financial Disclosure: Dr Garway-Heath has received research support from Heidelberg Engineering, Optovue, and Carl Zeiss Meditec and is a member of the advisory board for Carl Zeiss Meditec.

Funding/Support: This research was funded in part by unrestricted grants from Heidelberg Engineering, by The Moorfields Special Trustees, and by research grants K23-EYO16225 and R01-EYO3424 from the National Eye Institute, National Institutes of Health. Dr Garway-Heath received a proportion of his funding from the Department of Health’s National Institute for Health Research Biomedical Research Centre for Ophthalmology at Moorfields Eye Hospital National Health Service Foundation Trust and the University College London Institute of Ophthalmology.

Disclaimer: The views expressed in this publication are those of the authors and not necessarily those of the Department of Health.

REFERENCES


Figure 8. Case 4. Single baseline (July 1998) (A) and single final follow-up (July 2005) (D) photographs from stereophotograph pairs, with no observed change. B, Baseline Heidelberg retina tomograph (HRT) mean image (July 1998). Final follow-up HRT mean images (July 2005) with topographic change analysis (progression flagged) (C) and statistic image mapping (progression flagged) (E) output (the dark red pixels are the largest cluster of pixels in the disc). F, Output for linear regression of rim area (red sector: significant P value for negative trend of rim area).