Comparing Glaucomatous Disc Change Using Stereo Disc Viewing and the MatchedFlicker Software Program in Ophthalmologists-in-Training

JAMIE L. SCHAEFER, ZACHARY L. LUKOWSKI, ALISSA M. MEYER, ANTHONY J. LEONCAVALLO, ANTHONY GREER, GINA M. MARTORANA, BAIMING ZOU, JONATHAN J. SHUSTER, AND MARK B. SHERWOOD

PURPOSE: To compare the accuracy and speed of using the computerized MatchedFlicker software program (EyeIC Inc, Narberth, Pennsylvania, USA) to evaluate glaucomatous optic disc change against the traditional gold standard of manually examining stereoscopic disc photographs.

DESIGN: A prospective evaluation of diagnostic technology.

METHODS: Two resident ophthalmologists and 1 glaucoma fellow at the University of Florida independently evaluated 140 image pairs from 100 glaucomatous/ocular hypertensive patient eyes using a handheld stereo viewer and the MatchedFlicker program. Fifty had progression to glaucoma as determined by the Ocular Hypertension Treatment Study (OHTS) Optic Disc Reading Group and the OHTS Endpoint Committee in the OHTS, and 50 more had photographs taken a few minutes apart, which were negative controls with no progression. Twenty photograph pairs from each group were duplicated to determine reviewer variability. Photographs were examined in alternating blocks of 70 photograph pairs for each method, with the starting viewing method randomized. Reviewer accuracy and time to review for each method were measured.

RESULTS: Using the handheld stereo viewer, the reviewers correctly identified progression or nonprogression in 76.0% of the slide pairs. Using the MatchedFlicker software, 87.6% were correctly identified (P = .011). Evaluator speed averaged 34.1 seconds per image pair with the stereo viewer vs 24.9 seconds with the MatchedFlicker program (P = .044). Overall, Flicker was significantly more specific but less sensitive than stereo slides. Trainees appeared more reluctant to identify glaucoma progression from slides than from Flicker. For the 2 less experienced trainees Flicker was significantly more accurate.

CONCLUSION: The MatchedFlicker software had a greater accuracy and was quicker to perform than using a handheld stereoscopic viewer. (Am J Ophthalmol 2016;167:88–95. © 2016 Published by Elsevier Inc.)
and stereo viewer (although it is possible to obtain less standardized stereoscopic photographs with a nonsimultaneous camera), and the difficulty and time-consuming nature of the stereo photograph evaluation. With the increased use of electronic medical records (EMR), it is important to have a disc imaging technique that is accurate at detecting glaucoma progression and is also well adapted to the new electronic format.

This study was designed to compare the accuracy and speed of the new MatchedFlicker (EyeIC Inc, Narberth, Pennsylvania, USA), technique using a computer screen with the traditional gold-standard method of examining slides of stereoscopic disc photographs in a group of nonexpert observers who were in various stages of ophthalmology training.

METHODS

ALL RESEARCH IN THIS EXPERIMENTAL STUDY WAS performed in accordance with the University of Florida's Institutional Review Board (IRB) with protocol approval prior to initiation of the study. Approval was given for the collection of optic nerve photographs from patient records and from the Bascom Palmer Eye Institute Optic Disc Reading Center, as well as for the participation of the photograph evaluators. A full waiver of informed consent was obtained for the use of de-identified patient optic nerve photographs, because no risk was posed to any patient with their use. This study adhered to the tenets of the Declaration of Helsinki and all federal and state laws.

From February to May of 2013, 2 ophthalmology residents (postgraduate year [PGY]-2 and PGY-3) and 1 glaucoma fellow (PGY-5) at the University of Florida independently evaluated stereoscopic pairs of disc photographs of 100 eyes taken at 2 time points. Fifty eyes were identified from patients in the OHTS study that showed glaucoma progression. This progression was previously determined by the OHTS Optic Disc Reading Group of the Bascom Palmer Eye Institute and confirmed as glaucomatous change by the OHTS Endpoint Committee. The patient's comparison photographs were taken between 1 and 5 years apart, depending on when the disc progression was first noted by the study committee. Another group of 50 eyes were obtained that were known to show no progression since the photographs were taken at the University of Florida using a Topcon TRC-50DX simultaneous stereo camera system (Topcon, Oakland, New Jersey, USA) just a few minutes apart. Duplicate photographs of 20 of the 50 eyes that showed progression and 20 of the 50 eyes with no progression were randomly selected to allow for assessment of intra- and interobserver variability in detecting progression. Thus, a total of 140 image pairs (280 photographs) were examined by each observer.

Two different examination methods were used by each observer for judging the optic disc photographs for glaucomatous progression: handheld stereo viewers and the computerized MatchedFlicker program (identified further in the study as “Flicker”). Using 2 Pentax handheld stereo viewers (Pentax, Tokyo, Japan) (Figure 1), 35 mm simultaneous-stereoscopic color slides of optic disc photographs were reviewed with the slide from the first time point in one stereo viewer and the slide from the second time point in the other stereo viewer. No poor-quality photographs were used, but photographs were of naturally slightly varying quality to reflect what a practitioner would encounter in real-world clinical practice.

The evaluators reviewed the same sets of photographs using the Flicker software. For this computerized analysis, stereoscopic images are not required. The right image of the stereo pair from each of the 35 mm stereo photographs was digitized using a Nikon Super Coolscan 5000 (Nikon, Tokyo, Japan) and uploaded into the Flicker alignment and presentation software. In 4 eyes the right image was blurred and the left image of the stereo pair was selected instead. The right image at the second time point was also digitized and uploaded into the Flicker software. The Flicker program took these 2 digitized images, registered that they were photographs of the same eye taken at 2 time points, and aligned the object (the optic disc) in the 2 images so that they were superimposed (Figure 2). The program then “flickered” rapidly between the 2 images, which simulated a sensation of movement if any structural alterations had occurred. In this way, any features that underwent a potentially glaucomatous structural change between the photographs would be detected as movement. A 16 inch Samsung Galaxy laptop (Suwon, Korea) computer screen was used for the Flicker presentations.

Using the 2 different methodologies, the evaluators assessed the optic nerve head (ONH) images for evidence of progression including disc rim thinning (focal or diffuse), vessel movement related to increased cupping, or detection of new or enlarged RNFL defects.
Prior to examining the study photographs, each evaluator was given a practice session with 10 additional photograph pairs to familiarize them with each technique. For 2 of the examiners this practice with the Flicker technique was their only exposure to the program, as they had not used it previously. All 3 had some experience reviewing disc stereo photographs with a handheld stereo viewer.

The method of examining the disc pairs was divided into 2 alternate blocks of 70 paired images, using 1 method for the entirety of that block. The method that was used first was randomized. All 280 images were reviewed separately by each of the 3 observers in a single session. A forced decision on progression or nonprogression was required for each image, even if the observer was not certain. A study coordinator recorded the evaluator’s assessment of disc progression for each photograph and recorded the time taken for the overall disc determination using each method.

The time taken for each evaluator to complete the set of photographs using the 2 different techniques was divided by the number of photograph pairs viewed to give a mean “time-per-assessment” measurement.

- **STATISTICAL ANALYSIS:** The difference in accuracy between the Flicker and slides method was initially analyzed with a 2-sided paired t test comparing each reviewer’s percent accuracy for both the Flicker and slides method. A more precise analysis, in which the image pairs were scored for both Flicker and slides and the paired differences in score compared, is shown in Supplemental Table 1 (Supplemental Material at AJO.com). Each subject slide or image was scored 100% (3 of 3 reviewers correct), 66.7% (2 of 3 correct), 33.3% (1 of 3 correct), or 0% (0 of 3 correct) and the P value assessed by a 2-sided paired t test for all 100 image pairs.

The difference in the average evaluation time per assessment between the 2 methods was also analyzed with a 2-sided paired t test comparing the total time of each reviewer divided by all 140 slides in the block for each method.

Information on each reviewer’s sensitivity, specificity, and overall accuracy was compared between the 2 methods using the following logistic regression model:

\[
\logit[\Pr(y_{ijk} = 1)] = \log(\text{odds}(y_{ijk} = 1: y_{ijk} = 0)) = \alpha_0 + \alpha_1 I(i = 1) + \alpha_2 I(k = 2) + \alpha_3 I(k = 3)
\]

where \(y_{ijk}\) is the binary response (ie, 1 for correct and 0 for incorrect) of subject \(j\), reviewed by Reviewer \(k = 1, 2, or 3\) using method \(i = 1\) for Flicker and \(= 0\) for slides). Reviewer 1 (PGY-5) is used as the reference. \(I(x = \text{value})\) is the indicator function (ie, \(= 1\) if \(x = \text{value}\) and 0 otherwise). This produces for each of the groups (a, b, and c) an adjusted odds ratio for Flicker to slide, with values above 1.0 favoring Flicker and values below 1.0 favoring slides.

Sensitivity was determined by the percentage of progression-positive diagnoses that were correctly attributed to truly glaucoma-positive (ie, “true-positive” images).

**FIGURE 2.** Images produced by the MatchedFlicker software. (EyeIC Inc., Narberth, Pennsylvania, USA) used in assessing accuracy of detecting glaucomatous progression with the MatchedFlicker software technique.
The logistic regression models were analyzed using the open source R statistical analysis package (GNU Software), while the McNemar test was conducted in SAS (SAS Institute), using PROC UNIVARIATE, which provides exact P values. All tests were 2-sided and statistical significance was qualified by whether or not \( P < .05 \).

**RESULTS**

For all 3 reviewers, the number of correct responses using Flicker was significantly higher (\( P = .011 \)) than that achieved while using the stereo-slide viewer (Table 1 and Supplemental Table 1).

The observed time with Flicker was 9.2 seconds (27%) faster per assessment, on average, than those using conventional stereo slides (\( P = .04 \) (Table 2).

Specificity is the percentage of negative diagnoses that are truly negative, and sensitivity is the percentage of those diagnosed with true glaucoma progression. As a whole, reviewers were more specific and more accurate with Flicker than with slides (Table 3). Reviewers were also less sensitive with Flicker than with slides. While using Flicker, the reviewers assigned both positive and negative responses at almost an equal rate, identifying 49–54 of the 100 image pairs as negative. However, reviewers identified 66–79 of the 100 image pairs they viewed as negative for glaucomatous progression while using the slides.

According to Table 4, reviewers were significantly more accurate overall (\( P = .001 \)) and more likely to correctly diagnose the group of 50 image pairs depicting glaucomatous progression (\( P < .001 \)) when using the Flicker method. However, the group of 50 image pairs depicting no
glaucomatous progression were significantly more likely to be designated correctly by reviewers using the slides than by those using Flicker (P = .008) (Table 4). This difference from Table 3 is driven, as mentioned above, by the relative reluctance of the reviewers to call a positive progression with slide viewing, so that although they correctly called nearly all of the true negatives, there was also a group of true positives that were designated as negative by the reviewers, thus decreasing the percentage of negatives that were actually correctly called.

Reviewers 2 and 3 (PGY-2 and PGY-3, respectively) had significantly more instances where they could correctly decide whether there was glaucomatous progression between a set of image pairs while using Flicker, but were unable to decide correctly while using the slides for the same image pair (P = .043 and P = .026 for Reviewer 1 and Reviewer 2, respectively) (Supplemental Table 2, Supplemental Material at AJO.com). While results were not significant, Reviewer 1 (PGY-5) also had more instances of correctly deciding whether or not an image pair showed progression when using Flicker.

Analysis of the discordant pairs between any 2 reviewers while using Flicker revealed no cases of 1 reviewer having significantly more correct decisions on glaucomatous progression, with Reviewer 1 (PGY-5) also having more instances of correctly deciding whether there was progression in image pairs than his cohort (Table 5). However, it should be noted that when analyzing the cases of discordant responses while using slides, Reviewer 1 and Reviewer 2 (PGY-5 and PGY-2) differed significantly in the number of correct decisions on glaucomatous progression, with Reviewer 1 making a significantly greater amount of correct decisions (P = .012).

Intraobserver agreement was generally high between the 2 techniques when analyzing the 40 duplicates, with a mean of 30 and 33.67 of the 40 repeated images (75% and 84%) being given the same designation with the use of Flicker and slides, respectively (Table 6). While using the slides, there were 18 cases out of 42 agreements where the wrong decision on the presence of glaucomatous progression was made twice. In all 18 cases, the reviewers determined that there was no progression when there actually was. While using Flicker, there was only 1 case of 43 agreements where the wrong determination of glaucomatous progression was chosen twice. The error made was the same as with the slides: the reviewer saw no progression between the 2 image pairs, when progression had actually taken place.
A significant problem in the management of glaucoma is the difficulty in detecting progression so that adequate therapy can be instituted. Xin and associates found the accuracy for diagnosing glaucomatous disc progression over time using photographs with a known status in regard to progression. The negative controls in which simultaneous stereo photographs were taken within a few minutes of each other can reliably be stated to show no progression. It is more difficult to be certain that subtle disc progression has definitely occurred, but by using the results of a well-recognized NEI study disc reading center, our study aimed to get as close as possible to being certain that there is real progression. Indeed, this system of having a disc reading center review stereoscopic disc photographs is the current gold standard for detecting disc progression for NIH-sponsored glaucoma clinical trials. We considered using slides of nonprogressing discs from the OHTS study disc reading center as our negative control in view of the concern regarding increased lighting and other photographic differences when multiple clinics take photographs over longer time periods. However, we felt that there was some risk that discs that were graded in OHTS as not showing progression by the disc reading center might in fact have a minute number of subtle signs of progression, thus introducing a type II error (false negative). Although there are pros and cons of each negative control, on balance, we felt that using slides taken a few minutes apart and that were certain to have no progression would be better.

A recent study by Radcliffe and associates viewed stereo photograph image sets from patients with glaucoma using an automated alternation flicker (the Flicker software program) and found retinal blood vessel shifts in 26.4% of the longitudinally followed glaucomatous eyes. In addition, they found a statistically significant association between blood vessel movement and increased rate of visual field progression and neuroretinal rim loss.

Our study compared the new Flicker software program viewed on a standard computer screen with the traditional technique of examining a 35 mm slide in a handheld stereo viewer, for detecting glaucomatous optic disc structural progression. Our 3 observers included 2 resident ophthalmologists at different stages of their training and a glaucoma fellow at the University of Florida. The primary endpoint was the accurate identification of disc progression over time. A higher percentage of correct answers was found using the Flicker computerized program for all 3 reviewers (P = .01) (Table 1, Supplemental Table 1). There were no statistically significant differences in accuracy between the trainees with different levels of experience using Flicker, but when using slides the resident with the least experience examining stereo photographs was significantly less accurate than the glaucoma fellow (Table 5). This may reflect the fact that all 3 observers had a similar, limited experience using the Flicker program with just the same 10 training photograph pairs.
As mentioned previously in the statistical analysis section, specificity is the percentage of negative designations that are truly negative and sensitivity is the proportion of positive diagnoses that are truly positive (ie, have true glaucomatous progression). In our opinion, both are important, but perhaps having high specificity is of particular importance so that patients who are progressing are detected and can receive treatment. According to Table 3, the Flicker method produced more specific results (negative designations that are truly negative) than did slides. The slides method produced more sensitive results (positive designations that are truly positive) than Flicker, owing in part to the low number of positive designations that were made. The low number of image pairs that were identified as showing glaucomatous progression (Table 3) suggests that reviewers were less inclined to call the glaucomatous progression between a set of stereoscopic image pairs while using the slides method. For all 3 trainee observers, Flicker was more accurate for the true-positive image pairs (P = .001), less accurate for the true-negative image pairs (P = .008), and, regarding the full set of 100 unique image pairs, more accurate overall than stereoscopic slide review (P = .001) (Table 4). The test-retest reproducibility of detecting optic disc deterioration from stereo photographs by the certified graders at the Optic Disc Reading Center in the Ocular Hypertension Treatment Study was reported to show a specificity of 98%–100% and a sensitivity ranging from 64% to 81%. Our 3 trainee reviewers had a similar specificity of 95%–100% and a sensitivity of 65%–75% for the 40 duplicate slides they examined (Table 6).

With the demand for the clinician to see higher volumes of patients, the time taken to perform an accurate analysis is important. All 3 observers using Flicker were faster than using 35 mm slides (P = .04) (Table 2). Integration with the EMR will also be increasingly important over time. Scanning of digitized stereo images into the EMR and stereo viewing of these images usually requires specialized viewing devices and a specialized camera to obtain a simultaneous stereo image. The Flicker software has the advantage of using monoscopic images and does not require the use of stereo photography.

The current study only compared physicians at different stages of training. A future study comparing Flicker vs viewing of stereo photographs among glaucoma experts (subspecialists with multiple years of experience using stereo images) to determine the accuracy of detecting glaucoma progression would be useful to fully explore the role of Flicker analysis in clinical practice.

Our study with ophthalmologists-in-training demonstrated that the use of MatchedFlicker software is comparable or even better in accuracy than the traditional methods of stereo image optic disc viewing and was faster to perform. By using a technique for evaluating optic discs that is easier, faster, and potentially more accurate than the current gold standard, and in particular does not require stereo imaging and is suitable for EMR use, it is felt that ophthalmology residents would be more likely to regularly track the optic disc structure of their glaucoma patients, as well as visual fields and RNFL OCT. This technique has the potential to assimilate more easily into a clinical practice setting with EMR, and in conjunction with other functional and structural tests would be a good complement to the current methods of glaucoma progression detection.

REFERENCES


