APPRAISAL OF OPTIC DISC STEREO PHOTOS PRE- AND POST-TRAINING SESSION

CALLEWAERT S., FIEUWS S.**, STALMANS I.*, ZEYEN T.**

ABSTRACT

Purpose : To determine whether the diagnostic accuracy of judging serial optic disc color stereo photographs for glaucomatous change by non-expert ophthalmologists changed after a training session.

Methods : 24 ophthalmologists in training at the University Hospitals Leuven classified 50 eyes with varying severity of glaucoma as stable or progressing based on the appraisal of serial optic disc stereo photographs. A comparison between the diagnostic accuracy of residents (n=18) and trainees (n=6) was made before and after a training session.

Results: The mean agreement (κ) with the reference standard before training was lower for the trainees than for the residents. The mean κ before training was 0.37 for the residents and 0.29 for the trainees ($p = 0.18$). The mean agreement with the reference standard improved significantly after a training (from 0.29 to 0.56 [$p = 0.03$] for the trainees, and from 0.37 to 0.48 for the residents [$p = 0.005$]). The overall mean κ was 0.35 pre-training and 0.50 post-training ($p < 0.001$).

Conclusions: The agreement and diagnostic accuracy of residents in training in appraising serial optic disc photos improved significantly after a training session.

KEYWORDS

Non-expert ophthalmologists, optic disc stereo photographs, training session

.....

* *Department of Ophthalmology, University Hospitals, Leuven*
** *Leuven Biostatistics and Statistical Bioinformatics Centre (L-BioStat)*

INTRODUCTION

Glaucoma is a chronic neuropathy which is characterized by progressive ganglion cell loss, with specific optic disc excavation and corresponding visual field defects. It is the second leading cause of blindness worldwide (1).

Assessment of both the optic disc and visual field is of crucial importance in the diagnosis and follow-up of the disease (2). During the follow-up it is essential to determine whether there is progression of the disease. The optic nerve head (ONH) is the site at which the dropout of retinal ganglion cells is most easily identified with ophthalmoscopy. Since these changes can precede visual field defects, evaluating the optic disc is of the outmost importance (3). In the evaluation of the ONH and the retinal nerve fiber layer (RNFL), qualitative and quantitative features should be assessed (4). The qualitative features include the shape and the width of the neuroretinal rim, the evaluation of peripapillary atrophy and of the RNFL, and the presence of optic disc haemorrhages. The quantitative features are the optic disc size (vertical disc diameter), rim width and RNFL thickness. The poor agreement, even among expert observers, for subjective assessment of the optic disc has driven the development of clinical imaging devices such as the Scanning Laser Ophthalmoscope (HRT), the Scanning Laser Polarimeter (GDx), and Optical Coherence Tomograph (OCT) (5-7). A recent study suggests that the imaging devices outperform clinicians in classifying optic discs as normal or glaucomatous (8). A limitation of these imaging devices is that they tend to miss early glaucomatous damage more often than advanced damage. In addition, they have a limited accuracy for correctly classifying optic discs with an anatomic variation (e.g. tilted and myopic discs). Evaluation of images of suboptimal quality (e.g. motion artifacts and media opacities) should also be interpreted with caution. Although digital imaging devices have the potential to determine the rates of progression, pertinent and reliable studies related to this concern are not yet readily available. Furthermore, those machines are expensive and thus not available in every practice.

For these reasons it is important to emphasize that imaging devices should support rather than replace the clinical examination of the optic disc. Evaluating sequential stereo photos of the optic-disc still remains the gold standard for the evaluation of optic disc change (9). The advantages of stereo photography are a permanent recording of the optic disc status, especially useful for serial evaluation, an immediate availability of the photos when taken digitally, and a relatively fast examination without pupil dilation (using non-mydriatic fundus cameras).

Non-expert ophthalmologists are not as familiar as glaucoma experts with the evaluation of optic disc stereo photographs. The purpose of this study was to determine whether the diagnostic accuracy of judging serial optic disc stereo photographs for glaucomatous change by non-experts (residents in training) could be improved by a training session.

MATHERIALS AND METHODS

This study did not need to be submitted to our institutional review board, according to the policies of our ethical committee. This study adhered to the tenets of the Declaration of Helsinki.

A set of 50 color optic disc stereo photographs was selected by an experienced glaucoma specialist (TZ) out of data from the University Hospitals of Leuven. This set included data provided by J. Caprioli, MD, University of California, Los Angeles. The selection was made, based on the following criteria. All images had a high resolution. Only one eye per patient was selected. The optic disc stereo photographs of each patient had to be taken at least five years apart. All patients needed to have had at least five reliable automated visual field analyses and at least three HRTs during the follow-up period. This was done in order to allow further investigations; comparing structural to functional change.

The optic disc stereo photographs were presented in a chronological order. An example of one set of optic disc stereo photographs is shown in Figure 1. Each patient was assigned as “changed” (i.e. progressing, n = 21) or “unchanged” (i.e. stable, n = 29). The expert subdivided the progressing eyes into two groups: obvious change and subtle change. Four pho-

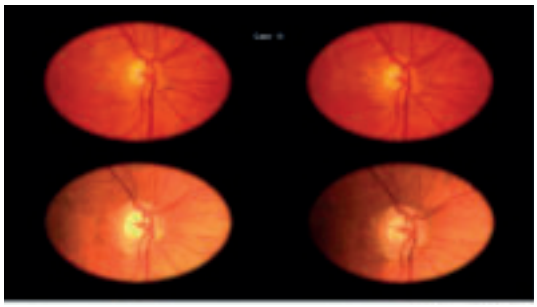

Fig. 1

tographs were randomly presented twice to check the intraobserver agreement during the test and as an informal check for agreement with the reference standard based on coincidence. So this maked a total of 54 photographs which were presented to the residents and trainees.

The study consisted of three parts. First, the 54 sets of serial optic disc stereo photographs were judged by each participant as changed or unchanged. The score was displayed and stored after the last patient. Then a short training course on the evaluation of optic disc stereo photographs was given by a glaucoma expert (TZ), using a different set of optic disc stereo photos. The explained characteristics of the optic disc were: location of rim loss, pallor and pseudo-pallor, optic disc haemorrhages and vessel positioning changes. After this training session, each participant was asked to judge the set of 54 stereo photos a second time. The answer was recorded and scored. After each second judgment the correct answer was revealed and a short explanation provided.

Twenty-four non-experts (ophthalmologists in training at the University Hospitals Leuven) have performed the test. The non-experts were subdivided into two groups: residents in training (n=18) and trainees who are a doing one year internship before starting their residency program (n=6). The participants evaluated the optic disc stereo photos on a normal computer screen (17 inches with a resolution of 1.440 x 900) using a stereo viewer (Figure 2).

STATISTICAL ANALYSIS

The 50 optic disc stereo photographs were assessed on the presence of change or no change

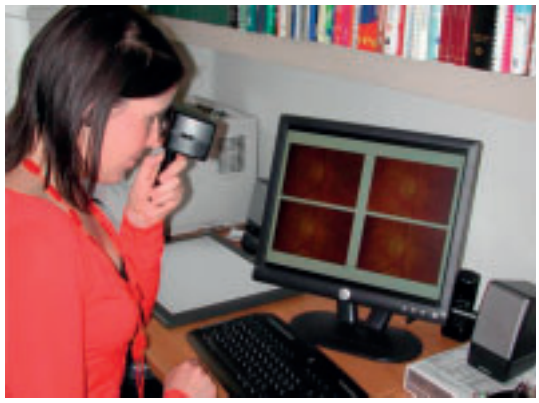

Fig. 2

by 24 participants before and after a training session. The interobserver agreement was assessed on both occasions using a kappa coefficient (κ) for multiple participants. A 95% confidence interval was constructed for the difference between the interobserver agreement pre- and post training.

Two approaches were used to study the agreement with the reference standard (the expert opinion was considered to be the reference standard) and to evaluate if this agreement changed as a function of training and type of participant (resident versus trainee).

In the first approach, indices were calculated for each participant separately before and after the training. Considered indices were Cohen's kappa, sensitivity, specificity, negative predictive (NPV) value and positive predictive value (PPV). Wilcoxon signed-rank tests were used to compare these indices between both sessions. Mann-Whitney U tests were used to compare these indices between residents and trainees.

In a second approach, the diagnostic indices were not summarized on participant-level but evaluated directly. Since all indices were proportions, binary logistic regression models were used and the clustering (multiple participants, two points in time) was taken into account using Generalised Estimating Equations (GEE). Specifically for sensitivity, it was verified if subtle changes were more difficult to detect than obvious changes.

All analyses have been performed using SAS software, version 9.2 of the SAS System for Windows. Copyright © 2002 SAS Institute Inc.

Table 1: *Pre- and post training comparisons.*

% Confidence intervals (V= value, LL=lower limit, UL=upper limit) and p-values are based on Generalised Estimating Equations (GEE)

		Pre-training			Post-training			
Setting	**Index**	**V**	**LL**	**UL**	**V**	**LL**	**UL**	***P*-value**
All observers	Sensitivity	0.601	0.492	0.701	0.734	0.621	0.823	<.0001
Residents	Sensitivity	0.595	0.480	0.701	0.717	0.600	0.811	<.0001
Trainees	Sensitivity	0.619	0.509	0.718	0.786	0.671	0.868	<.0001
All observers	Specificity	0.743	0.689	0.790	0.769	0.705	0.822	0.2140
Residents	Specificity	0.766	0.712	0.813	0.764	0.696	0.821	0.9286
Trainees	Specificity	0.672	0.586	0.749	0.782	0.704	0.843	0.0029
All observers	PPV	0.629	0.477	0.759	0.697	0.550	0.812	0.2932
Residents	PPV	0.648	0.495	0.776	0.688	0.538	0.807	<.0001
Trainees	PPV	0.578	0.420	0.721	0.723	0.574	0.834	0.0002
All observers	NPV	0.720	0.579	0.828	0.800	0.668	0.888	0.0012
Residents	NPV	0.723	0.582	0.831	0.789	0.654	0.880	0.0047
Trainees	NPV	0.709	0.563	0.822	0.834	0.708	0.913	<.0001

RESULTS

The mean agreement with the reference standard before training was lower for the trainees than for the residents. The mean κ before training was 0.29 for the trainees and 0.37 for the residents ($p = 0.18$). The trainees had pre-training $\kappa = 0.29$ and post-training $\kappa = 0.56$ (Wilcoxon signed-rank test, $p = 0.031$). The residents had pre-training $\kappa = 0.37$ and post-training $\kappa = 0.48$ (Wilcoxon signed-rank test, $p = 0.005$). The overall mean κ agreement with the reference standard was pre-training 0.35 and post-training 0.50 (Wilcoxon signed-rank test, $p < 0.001$).

The overall interobserver agreement significantly improved after the training (from $\kappa = 0.24$ to $\kappa = 0.38$ [$p < 0.0001$]). For the group of residents, the interobserver κ for difference in κ was 0.264 pre-training and 0.370 post-training session ($p < 0.0001$). For the group of trainees, the interobserver κ for difference in κ was 0.168 pre-training and 0.390 post-training session ($p < 0.0001$).

There was a significant change in sensitivity, positive predictive value (PPV) and negative predictive value (NPV) after the training session (Table 1). This holds for the group as a whole, as well as for residents and trainees separately (except for the PPV of the residents and the sensitivity of the trainees). There was no significant change in specificity after the training session. For none of the considered indices, there was evidence for a difference between residents and trainees. However, for specificity, PPV and NPV, the difference between pre- and post training depended on type of participant (i.e. the interaction between time and type of participant was significant). P was for the specificity 0.02, for the PPV 0.001 and for the NPV 0.01. More specifically, the evidence for an improvement after a training was stronger for trainees compared to residents.

The sensitivity was significantly ($p = 0.013$) higher when the changes were obvious as compared to subtle. The sensitivity for obvious change was 0.733 pre-training and 0.850 post-training. The sensitivity for subtle change was 0.481 pre-training and 0.629 post-training.

From all presented identical pairs of slides, 87% was evaluated consistent.

DISCUSSION

The purpose of this study was to evaluate the diagnostic accuracy of non-expert ophthalmologists (residents and trainees) in evaluating serial optic disc stereo photographs for change and to assess if a training session could improve their score.

The mean agreement with the reference standard was lower for the trainees than for the residents and this difference was statistically significant. This change can be explained by the lack of knowledge of the trainees before a training session. It also implies the effectiveness of the training session for both residents and trainees.

There was a significant improvement in the overall interobserver agreement after the training session for the group as a whole and also for residents and trainees separately.

The specificity and the PPV were significantly different between residents and trainees before the training (Table 1). Looking at the actual values of both indices, this implies that the impact of the training session is higher for the group of trainees, making them more similar to the residents after the training. Since the trainees have less experience grading the stereo photographs, it is not surprising that they have better results overall after the training session.

We found that the sensitivity was significantly higher when the changes were obvious as compared to subtle. *P*-values based on GEE for obvious change were 0.733 pre-training and 0.850 post-training, whereas *p*-values based on GEE for subtle change were 0.481 pre-training and 0.629 post-training. There was evidence that this difference depended on time (i.e. pre-/ post-training).

The likelihood to grade more photographs as "changed" is greater when the chronology is known. This was shown by Altangerel et al. (10). They found that the number of cases identified as having progressed was significantly higher (101 vs. 54) when the observer knew the chronological order in which the photographs were taken ($p=0.007$). Interobserver agreement was higher when the chronology of photographs was known ($\kappa = 0.68$ for unmasked evaluation vs. $\kappa = 0.30$ for masked evaluation). In our study we opted to display the optic disc stereo photographs in chronological order because this is how it happens in real life. We found that the sensitivity increased after the training session, but this was not at the expense of a lower specificity. This means that the likelihood to grade photographs as "changed" has increased, but this was not at the expense of the likelihood to grade the photographs as "not changed". It also implies the effectiveness of the training session

The intra-observer agreement was 87%. This compares favorably with the intraobserver agreements shown in other studies (11). A limitation of this study might be the high number of photographs used. Most participants complained of a lack of concentration during the second evaluation. This may explain why some participants scored lower on the second viewing and why the specificity and PPV did not improve after the training session.

Most studies have reported the performance of expert observers in identifying progression looking at serial optic disc stereo photographs. Little is known about the ability of general ophthalmologists to grade serial stereo photographs for change. In this study we assimilated residents in training to general ophthalmologists.

In order to obtain more data about the performance of general ophthalmologists in grading serial optic disc stereo photographs, we plan to run the same study online comparing the diagnostic accuracy of a larger group of non-expert ophthalmologists to the 2/3 agreement of glaucoma experts. Afterwards, the website can be used as a training tool or to reassess the diagnostic accuracy of the same group after 6 months.

In conclusion, we found, in this pilot study, that the diagnostic accuracy of residents in training to grade serial optic disc stereo photos improved significantly after a training session.

REFERENCES

(1) Resnikoff S, Pascolini D, Etya'ale D, et al. – Global data on visual impairment in the year 2002. Bull World Health Organ 2004; 82: 844-51.

(2) Giangiacomo A, Garway-Heath D, Caprioli J – Diagnosing glaucoma progression: current practice and promising technologies. Curr Opin Ophthalmol 2006; 17: 153-62.

(3) Sommer A, Katz J, Quigley HA, et al. – Clinically detectable nerve fiber atrophy precedes the onset of glaucomatous field loss. Arch Ophthalmol 1991; 109: 77-83.

(4) Caprioli J, Prum B, Zeyen T – Comparison of methods to evaluate the optic nerve head and nerve fiber layer for glaucomatous change. Am J Ophthalmol 1996; 121: 659-67.

(5) Medeiros FA, Alencar LM, Zangwill LM, et al. – Detection of progressive retinal nerve fiber layer

loss in glaucoma using scanning laser polarimetry with variable corneal compensation. Invest Ophthalmol Vis Sci 2009; 50: 1675-81.
(6) Kourkoutas D, Buys YM, Flanagan JG, Hatch WV, Balian C, Trope GE – Comparison of glaucoma progression evaluated with Heidelberg retina tomograph II versus optic nerve head stereophotographs. Can J Ophthalmol 2007; 42: 82-8.
(7) Medeiros FA, Zangwill LM, Bowd C, Weinreb RN – Comparison of the GDx VCC scanning laser polarimeter, HRT II confocal scanning laser ophthalmoscope, and stratus OCT optical coherence tomograph for the detection of glaucoma. Arch Ophthalmol 2004; 122: 827-37.
(8) Reus NJ, Lemij HG, Garway-Heath DF, et al. – Clinical Assessment of Stereoscopic Optic Disc Photographs for Glaucoma: The European Optic Disc Assessment Trial. Ophthalmology 2010; 117: 717-23.
(9) Medeiros FA, Zangwill LM, Bowd C, Sample PA, Weinreb RN – Use of progressive glaucomatous optic disk change as the reference standard for evaluation of diagnostic tests in glaucoma. Am J Ophthalmol 2005; 139: 1010-8.
(10) Altangerel U, Bayer A, Henderer JD, Katz LJ, Steinmann WC, Spaeth GL – Knowledge of chronology of optic disc stereophotographs influences the determination of glaucomatous change. Ophthalmology 2005; 112: 40-3.
(11) Zeyen T, Miglior S, Pfeiffer N, Cunha-Vaz J, Adamsons I – European Glaucoma Prevention Study Group. Reproducibility of evaluation of optic disc change for glaucoma with stereo optic disc photographs. Ophthalmology 2003; 110: 340-4.

.....

Adress for correspondence:
Department of Ophthalmology, University Hospitals
Kapucijnenvoer, 33, 3000 Leuven, BELGIUM
Fax: +32 16 332367
Tel: +32 16 332385
Email: thierry.zeyen@uzleuven.be