Computers in Biology and Medicine 41 (2011) 857–864

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Event-based progression detection strategies using scanning laser polarimetry images of the human retina

K.A. Vermeer [a,c,*,1], B. Lo [d,2], Q. Zhou [d,2], F.M. Vos [c,e], A.M. Vossepoel [c,f], H.G. Lemij [b]

[a] Rotterdam Ophthalmic Institute, Rotterdam Eye Hospital, Schiedamse Vest 160, NL-3011 BH Rotterdam, The Netherlands
[b] Glaucoma Service, Rotterdam Eye Hospital, Schiedamse Vest 180, NL-3011 BH Rotterdam, The Netherlands
[c] Quantitative Imaging Group, Delft University of Technology, Lorentzweg 1, NL-2628 CJ Delft, The Netherlands
[d] Carl Zeiss Meditec, Inc., 5160 Hacienda Drive, Dublin, CA 94568, USA
[e] Department of Radiology, Academic Medical Center, P.O. Box 22660, NL-1100 DD Amsterdam, The Netherlands
[f] Biomedical Imaging Group Rotterdam, Erasmus MC – University Medical Center Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands

ARTICLE INFO

Article history:
Received 6 April 2010
Accepted 22 June 2011

Keywords:
Progression detection
Simulation
Glaucoma
Polarimetry
Optimization
Image processing

ABSTRACT

Monitoring glaucoma patients and ensuring optimal treatment requires accurate and precise detection of progression. Many glaucomatous progression detection strategies may be formulated for Scanning Laser Polarimetry (SLP) data of the local nerve fiber thickness. In this paper, several strategies, all based on repeated GDx VCC SLP measurements, are tested to identify the optimal one for clinical use. The parameters of the methods were adapted to yield a set specificity of 97.5% on real image series. For a fixed sensitivity of 90%, the minimally detectable loss was subsequently determined for both localized and diffuse loss. Due to the large size of the required data set, a previously described simulation method was used for assessing the minimally detectable loss. The optimal strategy was identified and was based on two baseline visits and two follow-up visits, requiring two-out-of-four positive tests. Its associated minimally detectable loss was 5−12 μm, depending on the reproducibility of the measurements.

1. Introduction

Automated detection of glaucoma, one of the world's most common causes of blindness [1], has shown promising performance over the past years, based on various imaging technologies [2,3]. Much less progress, however, has been made in the clinically equally important problem of detecting progression in glaucoma. An automated and objective detection method would help the clinician in monitoring patients and assessing the effectiveness of the current treatment. Additionally, individuals with a high risk of contracting glaucoma might be monitored as well, since such a progression analysis is likely to be more sensitive to detect conversion to glaucoma than a diagnosis based on a single exam. Large inter-patient variability, due to biological differences, limit statistical analyses based on population-based normative values. Therefore, assessing intra-patient changes can be a much more sensitive way of detecting the onset of glaucoma.

One of the available imaging modalities for the detection of glaucoma is scanning laser polarimetry (SLP). Its working principle is as follows. The structure of the axons of the ganglion cells in the retinal nerve fiber layer (NFL) gives rise to birefringence. Due to their bundled ordering in the retina, polarized light that passes the NFL shows retardation. After reflection by the retinal pigment epithelium, the amount of retardation, assumed to be a measure for the thickness of the NFL, is analyzed by a crossed analyzer. SLP is commercialized as the GDx VCC (Carl Zeiss Meditec, Inc., Dublin, CA), which contains both the scanner itself and a software program that assists in the acquisition procedure. It also analyzes the scan, derives various parameters and translates these into an overall score, the Nerve Fiber Indicator [3], which may be interpreted as a soft classification of glaucoma likelihood. A significantly higher amount of NFL loss over time, as measured with the GDx VCC, has been found in eyes showing progression by standard methods, compared to stable eyes [4].

Perimetry, a method to assess the local functional sensitivity of the patient's eye to a light stimulus (visual field), is often regarded as the golden standard in glaucoma diagnosis, despite its relatively poor sensitivity and specificity [5]. While the assessed diagnostic accuracy depends on the reference standard, the low diagnostic accuracy is illustrated by the perimetry's poor reproducibility [6]. The appearance of the papilla is another clinically important feature, but this is hard to define in an objective, quantitative way. Previous studies indicate that, at least in

* Corresponding author at: Rotterdam Ophthalmic Institute, Rotterdam Eye Hospital, Schiedamse Vest 160, NL-3011 BH Rotterdam, The Netherlands. Tel.: +31 104023433.
E-mail addresses: k.vermeer@eyehospital.nl, koen@vermeer.tv (K.A. Vermeer), barrick@gmail.com (B. Lo), f.m.vos@tudelft.nl (F.M. Vos), a.m.vossepoel@tudelft.nl (A.M. Vossepoel), hlemij@me.com (H.G. Lemij).
[1] This author is no longer with the Quantitative Imaging Group.
[2] These authors are no longer with Carl Zeiss Meditec, Inc.

0010-4825/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2011.06.022

a subset of eyes, significant structural loss may be required before abnormal visual fields occur [7–9]. This suggests that in those eyes SLP, rather than perimetry, is a good modality for early detection of glaucoma and possibly also for early progression detection. Indeed, the GDx VCC's software enables a serial analysis by showing the differences between measurements at different times. It does not, however, provide a progression analysis, as the results of the serial analysis do not attribute the detected change to either real change or random measurement errors.

If we knew how progression took place over time, e.g., linearly, curvilinearly, stepwise, etc., an appropriate regression analysis (or trend-based analysis [10]) could be performed. The progression would then be modeled as a certain amount of loss over time and any change within an eye could be compared to the loss predicted by the model [11]. Provided that such a model resembled real progression faithfully, this approach might serve as a sensitive progression detection method. However, this type of knowledge about the process of glaucomatous nerve fiber loss is not readily available.

In contrast to regression analysis, a different approach is to look only at the mere amount of resulting loss and to weigh this against measurement variability. Obviously, no knowledge about the process itself is required, because only the net effect after a certain amount of time is considered. This is called an event-based analysis [10], as only the occurrence of the event is analyzed, not its specific development over time [12]. Such an event-based approach will always perform worse than the optimal regression analysis, given its obliviousness to any knowledge of the loss over time. On the other hand, it may be better than a regression analysis based on a wrong model (e.g., an exponential model whereas the true loss is a linear process) [13]. Unless a good model is available, we argue that the event-based approach is therefore the safer one and may eventually be used to assess the validity of a given model.

Obviously, both stable measurement series and progressing series are required to assess the performance of any progression detection method. Due to the generally slow progression of glaucoma, acquiring such a data set may be a lengthy operation spanning many years and will probably result in just a few cases of confirmed progression. Therefore, simulation has been suggested and used to study progression detection methods in visual fields [14–16] and in other imaging modalities such as confocal scanning laser topographs [17,18]. Previously, we have described a method to simulate progression in SLP images [19] and briefly discussed progression detection methods. These simulations may be used for initial progression detection algorithm development. Ultimately, these algorithms will have to be integrated into clinically relevant progression detection software and validated in clinical trials.

For progression in SLP images, no validated progression models are available. Therefore, we choose an event-based analysis as the initial method to detect progression. In this paper, we introduce and test several variations of possible event-based methods for progression detection in SLP images. Real images of stable eyes will be used to assess the specificity of each variant, whereas the sensitivity will be determined by simulated images, with predetermined amounts and different kinds (i.e., diffuse or localized) of progression. Note that we only consider the net effect of the progression; we do not model progression stages and time is therefore not used in the analyses. See Fig. 1 for an overview of the full approach, which is described in more detail in the next section. In this paper, we describe various progression detection strategies (extending our previous work [19]) and present a method to evaluate their detection performance for various types and amounts of glaucomatous progression. The resulting optimal strategy is selected based on its *minimally detectable loss* and used to build a clinically useful progression detection method. This resulting method will eventually have to be further evaluated in a clinical setting.

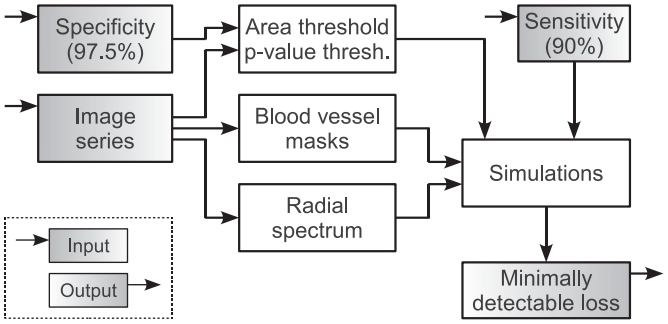

Fig. 1. Overview of the procedure to derive the minimally detectable loss from a number of image series and set sensitivities and specificities. See the main text for a full description of each step.

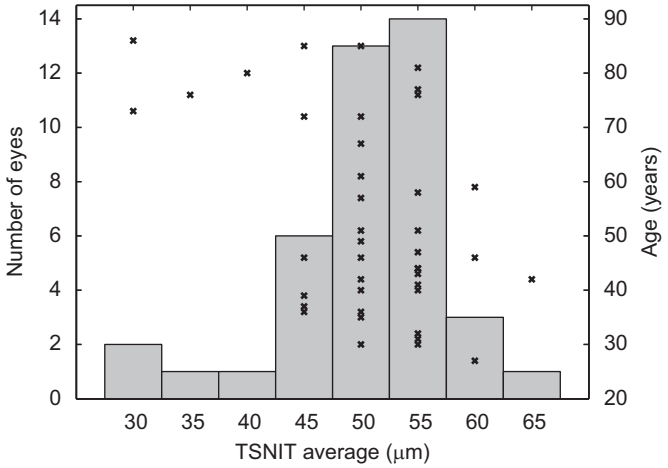

Fig. 2. Histogram of the TSNIT averages of the included eyes. The crosses denote the ages (in years) of the people in each bin.

2. Materials and methods

2.1. Data

The data set contained an image series of 41 stable eyes (i.e., no pathological changes were observed during the time of this study). One eye was randomly selected from each of 41 subjects. On four different days within a period of one month, three images were acquired with a GDx VCC (resulting in four sets of three images each). As the period between successive visits was short compared to the rate of change normally encountered in glaucoma, the assumption was made that no real structural change in the NFL had taken place between visits. The Enhanced Corneal Compensation (ECC) method was used for anterior segment birefringence compensation [20].

The mean age of the subjects was 53 years (SD 18 years, range 27–86), 63% of them were men, and 19 left eyes and 22 right eyes were selected. The mean of the TSNIT average (which is the mean of the NFL thickness on a circle around the optic nerve head, running through the **t**emporal, **s**uperior, **n**asal, **i**nferior and again **t**emporal quadrants) was 50 μm (SD 7 μm, range 29–66); see also Fig. 2. All eyes with a TSNIT average in the lower three bins (30, 35 and 40 μm) corresponded to people over 70 years. The

other bins covered the full range of ages, except for the two last bins (60 and 65 μm), where the age was below 60 years for all four cases. This is related to the thinning of the NFL that is associated with normal aging.

The range of eyes that were included in this study is not typical for the general glaucoma population. Based on a typical set of glaucomatous eyes, no simulations are possible for healthy eyes and therefore no specificity data can be derived. Therefore, a wider range of eyes was included. The collection of the data followed the tenets of the Declaration of Helsinki. All subjects gave their informed consent after explanation of the nature of the study.

2.2. Detection method

Any progression detection method should relate the reproducibility of the measured entity to the observed change, either implicitly or explicitly. Briefly, the proposed method first estimated the local change and variance by analyzing all available images of an eye. Student's *t*-test was then applied to each pixel individually to estimate the probability of change, under the null hypothesis of no change. If more than two visits were available, a *t*-test was performed on each pair of two out of all available visits and the resulting probabilities were combined according to the number of required positive tests. Pixels exceeding a pre-set combined probability (of θ_p) were flagged and subsequently groups of connected flagged pixels (with a minimal size of θ_A) were considered as areas of change. The selection of both threshold values is explained below. In this approach, the estimation of both the change and the reproducibility was determined from the observed eye only. No population-derived data are used for the statistical tests; only the thresholds that are placed on those tests are derived from a population.

All steps of the procedure will be described in more detail below. To fully specify the algorithm, two thresholds (θ_p and θ_A) were defined, together with the number of baseline visits (B), the number of follow-up visits (F) and the number of tests that indicate progression (n) out of all possible tests (m). Note that we allow more than one baseline visit to reduce the dependency of the algorithm on a single baseline visit. Because each baseline visit may be combined with each follow-up visit, $m = B \cdot F$. These four numbers thus specify a strategy, which will be denoted by B, F, n/m.

2.2.1. Preprocessing

Because the scan area of both the optic nerve head and the blood vessels do not provide information about the NFL thickness, these were excluded from the analysis. The location and size of the optic nerve head were manually determined by the operator for each image after acquisition. The blood vessels were automatically detected [21] and were also used as landmarks for automatic alignment of the images of each eye, based on a multi-scale Levenberg–Marquardt minimization [22] of the number of mismatching pixels in both blood vessel masks. Mean images were calculated by pixel-wise averaging all aligned images of a visit. Subsequently, difference images were computed by subtracting the corresponding mean image from an image.

2.2.2. Student's t-test

An independent two-sample *t*-test, with pooled variance, was done separately for each combination of baseline visits and follow-up visits. For instance, with one baseline visit (V_B) and two follow-up visits ($V_{F,1}$ and $V_{F,2}$), two comparisons were performed: V_B against $V_{F,1}$ and V_B against $V_{F,2}$. Alternatively, two baseline visits ($V_{B,1}$ and $V_{B,2}$) and two follow-up visits result in four possible tests: $V_{B,1}$ against $V_{F,1}$, $V_{B,1}$ against $V_{F,2}$, $V_{B,2}$ against $V_{F,1}$ and $V_{B,2}$ against $V_{F,2}$.

First, the variance of each pixel was estimated based on all available difference images, thereby assuming that the reproducibility depends on the specific eye and not on the time of the exam. Then, Student's *t*-test was applied to each set of corresponding pixels from two mean images obtained from a specific baseline and follow-up visit. The use of more than just these two visits for variance estimation aims to give a better estimate of the variance and results in a larger number of degrees of freedom. Note that per added visit, the number of degrees of freedom increases by two (for three images per visit). In this way, the *t*-test, applied to each pixel, resulted in a *p*-value map (see Fig. 3a). Note that these *p*-value maps are two-sided: A high *p*-value (near 1) corresponds to likely increase of NFL thickness, whereas a low *p*-value (near 0) corresponds to likely decrease of NFL thickness. *p*-Values around 0.5 indicate that the difference between the mean measurements are not significant with respect to the observer variance.

In the case of multiple baseline or follow-up visits, the resulting *p*-value maps were combined, depending on the number of required positive tests (n). This is done by choosing the n-th smallest *p*-value for each pixel. For example, if the *p*-values of the three tests based on one baseline and three follow-up visits are 2%, 1% and 3.5%, the resulting *p*-value for the 1,3,2/3 test is 2% (2 tests agree on the 2% level), or 3.5% for the 1, 3, 3/3 test (3 tests agree on the 3.5% level).

2.2.3. Thresholding and spatiality

The combined *p*-value map was first thresholded at θ_p, resulting in a binary map (see Fig. 3b). For the remainder, the superior and inferior hemispheres were treated separately on anatomical grounds. As the result of the thresholding may not be coherent, small holes were filled in and areas connected by small strings were split up. (For details on the morphological operators used in

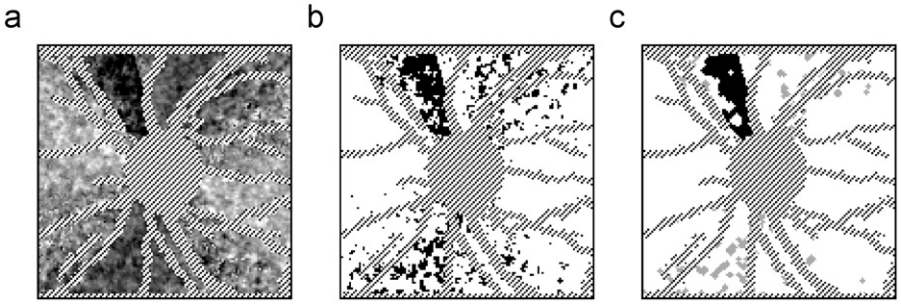

Fig. 3. Example of (a) a *p*-value map (black pixels correspond to a high chance of NFL decrease, white pixels correspond to a high chance of NFL increase), (b) after thresholding at θ_A (resulting objects shown in black) and (c) after morphological filtering (gray objects) and thresholding at θ_A (large black object).

this procedure, we refer to [19].) Then, the size of each area was calculated. Because areas of loss may be partly covered by blood vessels, areas on opposing sides of a blood vessel were treated as one area. Finally, all areas smaller than the area threshold θ_A, defining the smallest clinically relevant size of an area showing loss, were removed and the remaining areas were considered to show change (see Fig. 3c).

2.3. Parameters

Because the progression detection algorithm was designed to be used at every follow-up exam, the specificity of the algorithm was set to a rather high value of 97.5% to prevent false positives. The area size threshold (θ_A) was 100 pixels, corresponding to a retinal area of approximately 0.21 mm^2. For comparison, thresholds of 50 pixels (≈ 0.10 mm^2) and 200 pixels (≈ 0.42 mm^2) were tested as well. Based on the available data of stable eyes and these two fixed parameters, the correct p-value threshold (θ_p) was determined for each progression test as follows. Per decade, 50 log-evenly spaced p-values were tested and the resulting specificities were calculated. θ_p was then set to the largest p-value with a specificity closest to the required 97.5%. The visits, which were assumed to be interchangeable as the eyes were stable, were permuted to (artificially) increase the number of data points.

2.4. Simulations

With two parameters θ_A and θ_p, each progression detection strategy was fully specified. However, to pick the optimal strategy, the performance of each strategy had to be assessed. To this end, images with progression were simulated. The simulation was based on the radial spectrum derived from Fourier analysis of the images of the 41 stable eyes in the data set. This spectrum defines the various frequency components that together constitute the correlated noise in the difference images. The resulting correlated noise was assumed to be equal for all eyes and set the minimum variability encountered in the data. Additional eye-specific variability was incorporated by including the effects of incomplete cornea compensation, where increasing incomplete cornea compensation also produced increasing variability in the simulated measurements. A blood vessel mask of one of the real images was randomly picked, and a specified amount and type of loss was added to the simulated image series. More details on the simulation procedure may be found in [19].

Two types of loss that are also observed clinically were tested: localized and diffuse loss [23,24]. In the former case, nerve fiber loss is observed in a cluster. Given the approximately radial distribution of nerve fiber bundles, centered at the optic nerve head, this results in clearly defined loss in a sector. In the latter case, a general loss of nerve fiber is found in the whole peripapillary area. For diffuse loss, the specified amount of loss was therefore subtracted from every pixel of all mean images excluding those corresponding to the baseline visits. For localized loss, only a part of the image was changed. We chose to induce loss in a sector of varying widths (10°, 20°, 30°, 40° and 60°), with its center at an angle of approximately 60° from the horizontal meridian. See Fig. 4 for an illustration.

The eye-based sensitivity was fixed at 90%, meaning that a certain amount of loss had to be detected in at least 90% of the cases with (simulated) progression. Any loss detected outside of the area of induced loss was ignored, meaning that the overlap of the areas of detected and induced loss had to exceed θ_A. The minimally detectable loss was defined as the minimum amount of loss that could be detected in 90% of the eyes. This minimally detectable loss was determined for all strategies, different areas of loss, and various levels of reproducibility.

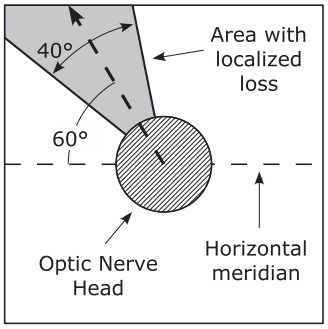

Fig. 4. Illustration of localized loss in a 40° wide sector.

2.5. Cross-validation

Cross-validation is commonly used to estimate the optimal value of certain model parameters and statistics to assess the performance, such as specificity, sensitivity or overall accuracy. In this paper, it was used to assess the specificity of the selected p-value, which is likely to be slightly different from the specified value of 97.5%. We applied n-fold cross-validation, meaning that the data are divided into n (the folds) more or less evenly sized sets of patients. All but one of these sets was used to select the p-value corresponding to a specificity of 97.5%. The resulting specificity was estimated by calculating the specificity of the unused set with this p-value. This was then repeated n times, such that each set was used $n-1$ times to determine the p-value and once to estimate the specificity. Finally, the specificities of each fold were averaged to get the cross-validation estimation of the specificity. By repeating the cross-validation, the standard deviation of this estimate was also computed.

In our application of cross-validation, the data set contains multiple images of each eye. Considering all images independently would result in an unrealistically small error. Instead, the cross-validation was performed between patients, which meant that each fold contained images of a disjoint set of eyes. In this way, all images of an eye are either used for training or testing.

3. Results

For all visits in the data set, the mean standard deviation across all pixels was calculated and used as a measure of reproducibility. The distribution of these reproducibilities is shown in Fig. 5. The distribution is clearly not normal (a Kolmogorov–Smirnov, KS, test, used to compare an observed distribution to a given one, indicates that $P<0.01$), but more closely follows a log-normal distribution ($P=0.09$ according to the KS test). Note that more than 96% of all reproducibilities are between 1.5 and 5.5 μm and more than 98% fall within the range of 1.5−6.5 μm.

The resulting p-values for an area threshold of 100 pixels are shown in Table 1. Also shown is the resulting specificity, estimated by repeated 10-fold cross-validation. Note that the θ_p listed for each test strategy is determined on the full data set; the actual θ_p for each of the folds of the cross-validation may be different. The relations between the p-values are also illustrated in Fig. 6, clustered per combination of available exams. Table 1 and Fig. 6 show that

- If more positive tests are required, θ_p for each test was higher (i.e., less strict). Example: θ_p for 1,3,2/3 is larger than for 1,3,1/3.
- If the same number of positive tests are required and the number of available exams (and therefore the number of possible tests) increases, θ_p for each test was lower (i.e., more strict). Example: θ_p for 1,3,2/3 is smaller than for 1,2,2/2.

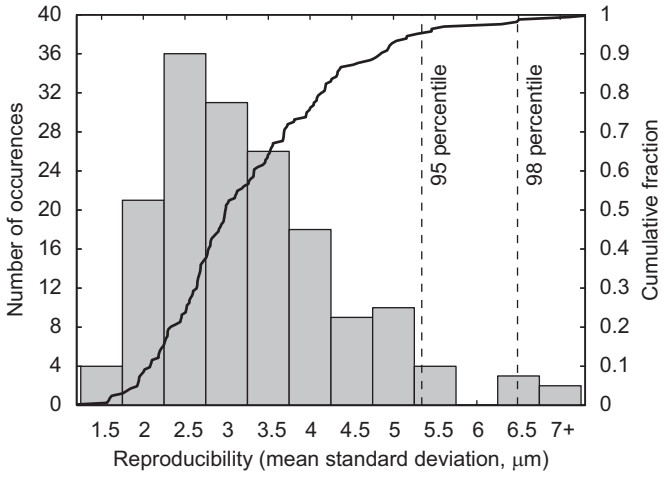

Fig. 5. Histogram of the reproducibilities derived from all visits. The thick solid line shows the cumulative fraction. The 95 percentile (5.34 μm) and 98 percentile (6.49 μm) are indicated by the dashed vertical lines.

Table 1
Resulting p-value thresholds for $\theta_A = 100$ pixels and a specificity of 97.5%. The first column lists the test used, the second column the resulting θ_p and the last column the estimated resultant specificity and standard deviation.

Test	θ_p (%)	Spec. % (SD)
1, 1, 1/1	1.15	97.26 (0.2)
1, 2, 1/2	0.79	97.59 (0.2)
1, 2, 2/2	3.8	98.02 (0.15)
1, 3, 1/3	0.66	96.86 (0.2)
1, 3, 2/3	2.4	97.49 (0.2)
1, 3, 3/3	5.0	97.33 (0.3)
2, 1, 1/2	0.79	97.35 (0.3)
2, 1, 2/2	4.0	97.49 (0.14)
2, 2, 1/4	0.55	97.87 (0.7)
2, 2, 2/4	2.4	97.15 (0.2)
2, 2, 3/4	6.3	96.50 (0.2)
2, 2, 4/4	11.0	97.18 (0.2)

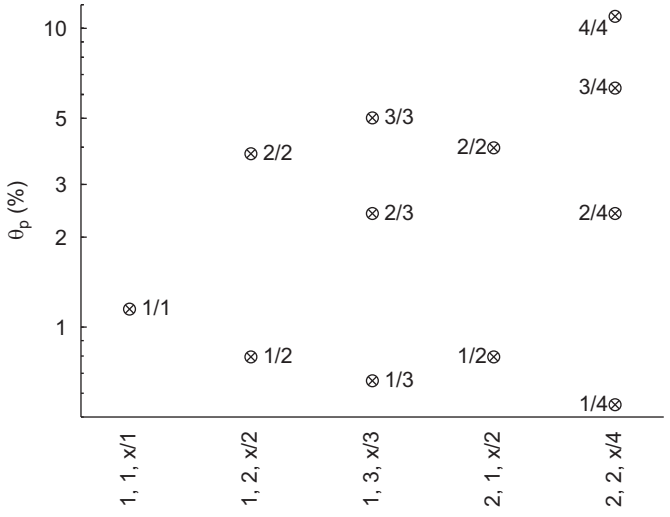

Fig. 6. Resulting p-value thresholds for all strategies (area threshold is 100 pixels), corresponding to Table 1. The horizontal axis denotes the family of test strategies (i.e., with a fixed number of baseline and follow-up visits). The points in the graph correspond to the number of positive tests ($x \in \{1, \ldots, n\}$) out of all possible tests (n).

- If the number of baseline and follow-up exams are swapped, θ_p is approximately the same. Example: θ_p is similar for 1,2,2/2 and 2,1,2/2.

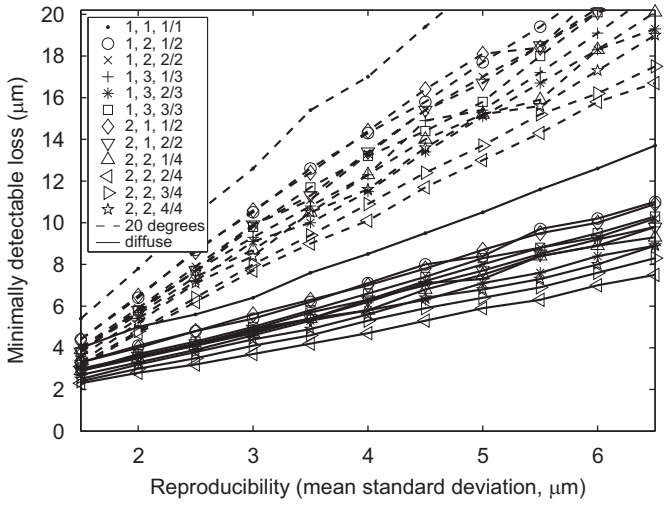

Fig. 7. Results of simulation study for all strategies (see Table 1; different markers indicate different strategies) with a localized loss of 20° (dashed line) or diffuse loss (solid line) and an area threshold of 100 pixels.

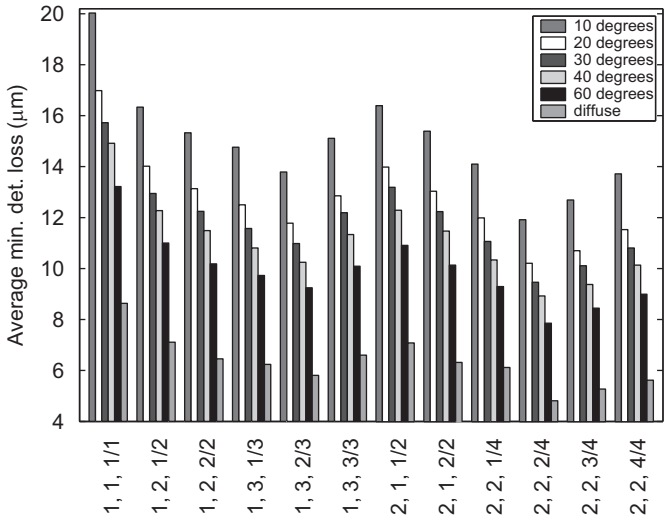

Fig. 8. Mean minimally detectable loss for all strategies and all regions of loss for $\theta_A = 100$ pixels.

Because the encountered reproducibilities in the data set were in the range of 1.5−6.5 μm for almost all cases (see Fig. 5), the simulation study was restricted to these reproducibilities as well. In Fig. 7, the results for this range of reproducibilities is shown, for localized loss of 20° and diffuse loss, for all strategies. The loss was simulated by subtracting a certain amount of NFL signal from all pixels in a specified region in the mean images. This simulated loss was increased until it was detected in 90% of the simulated cases, thus defining the minimally detectable loss.

Since the measurements indicate a linear relationship between reproducibility and minimally detectable loss, a straight line is fitted through them for further analysis. For better comparison between the strategies, the mean minimally detectable loss for the range of reproducibilities is calculated from these fitted lines and shown in Fig. 8 for all strategies and all regions of loss.

To compare the chosen area threshold of 100 pixels to other area thresholds, the analysis was also done for an area threshold of 200 pixels. First, the p-values were recalculated to get a specificity of 97.5% with the new area threshold (see Table 2). Then, the minimally detectable loss was calculated based on the

Table 2
Resulting p-value thresholds for $\theta_A = 200$ pixels and $\theta_A = 50$ pixels and a specificity of 97.5%. The estimated specificity is also shown.

Test	$\theta_A = 50$		$\theta_A = 200$	
	θ_p (%)	Spec. % (SD)	θ_p (%)	Spec. % (SD)
1, 1, 1/1	0.66	97.3 (0.2)	1.66	97.5 (0.2)
1, 2, 1/2	0.53	97.3 (0.2)	1.10	97.1 (0.2)
1, 2, 2/2	2.8	97.3 (0.15)	5.0	97.4 (0.14)
1, 3, 1/3	0.40	96.9 (0.2)	0.79	97.0 (0.6)
1, 3, 2/3	1.66	97.4 (0.2)	2.9	97.3(0.3)
1, 3, 3/3	3.8	97.3 (0.3)	6.9	97.4 (0.2)
2, 1, 1/2	0.48	97.2 (0.3)	1.00	97.0 (0.16)
2, 1, 2/2	3.0	97.4 (0.14)	5.5	97.3 (0.16)
2, 2, 1/4	0.35	96.9 (0.4)	0.69	96.9 (0.8)
2, 2, 2/4	1.66	97.2 (0.2)	2.8	97.1 (0.3)
2, 2, 3/4	5.2	97.3 (0.3)	7.9	96.9 (0.2)
2, 2, 4/4	7.9	97.1 (0.3)	13.8	97.0 (0.4)

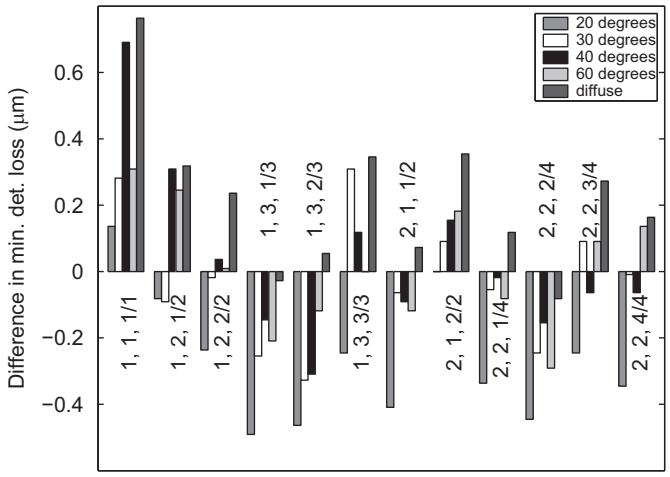

Fig. 9. Difference between $\theta_A = 100$ pixels and $\theta_A = 200$ pixels. Positive numbers indicate a smaller minimally detectable loss for $\theta_A = 200$ pixels.

simulations. Again, straight lines were fitted to the data and the mean minimally detectable loss was calculated. Fig. 9 shows the differences between the mean minimally detectable loss for both area thresholds. Note that the absolute differences are rather small, but that in general the minimally detectable loss is slightly better for larger areas of loss, but decreases for smaller areas. For most test strategies, it results in a generally smaller minimally detectable loss.

Likewise, a smaller area threshold of 50 pixels was tested. The p-values are listed in Table 2 and the resulting differences with the mean minimally detectably loss for $\theta_A = 100$ pixels were calculated. The smaller area threshold only resulted in slightly lower minimally detectable losses for loss in a sector of 10°; for all other types of loss, the minimally detectable loss was larger.

4. Discussion

For all regions of loss, either localized or diffuse, the 2, 2, 2/4 strategy (meaning two baseline visits, two follow-up visits and two positive tests out of the out possible comparisons) showed the lowest minimally detectable loss and is therefore the preferred method for progression detection out of the simulated progression detection strategies. Depending on the region size of simulated loss, the mean minimally detectable loss was about 5−12 μm, for a sensitivity of 90% and a specificity of 97.5%. A balanced strategy (i.e., a strategy based on an equal number of baseline visits and follow-up visits) such as the 2,2,3/4 strategy performs better than an unbalanced strategy based on the same total number of visits, such as the 1,3,2/3 strategy. The best strategy for a total number of visits of three is the 2,1,2/2 strategy.

The resulting p-values could be explained qualitatively as follows. The 1,2,1/2 may be interpreted as two 1,1,1/1 tests, for which only one has to succeed. The test is therefore less strict, and consequently, a stricter, therefore smaller p-value, is required to achieve the same specificity. Likewise, the 1,2,2/2 strategy is a combination of two 1,1,1/1 tests, where both tests have to agree. The test is therefore stricter, and a less strict p-value results. Writing $p_{B,F,n/m}$ for the p-value threshold for strategy $B,F,n/m$, this may be generalized as $p_{B,F,n/m} \leq p_{B,F-1,n/m-B} \leq p_{B,F,n+1/m}$. In a similar way, the relations between strategies based on a different number of baseline visits can be explained and written down as $p_{B,F,n/m} \leq p_{B-1,F,n/m-F} \leq p_{B,F,n+1/m}$. Finally, there should be no significant difference between the thresholds of the $B,F,n/m$ and $F,B,n/m$ strategies, which can be denoted as $p_{B,F,n/m} \approx p_{F,B,n/m}$. The experimental p-values, as shown in Fig. 6, closely follow the relations predicted by the theory.

Increasing the area threshold to 200 pixels did not reduce the mean minimally detectable loss for the best performing strategies, although some less optimal strategies showed a reduction of the mean minimally detectable loss of about 0.5 μm. A smaller area threshold of 50 pixels did not improve the mean minimally detectable loss for any strategy or region of loss. Therefore, the optimal area threshold for all strategies seems to be roughly 100 pixels, whereas only some of the strategies profit from a larger area threshold. Further optimizing the area threshold was considered to be of little clinical value.

One might expect increased sensitivity when the area size threshold is lowered. Although this is true if all other parameters were fixed, this is not the case for our study. A lower area size threshold would not only result in more detected cases for the simulated data, but also for the stable eyes, causing a drop of the specificity. As the specificity was set to 97.5%, this would have to be compensated by decreasing the p-value thresholds, which results in increasing the specificity. Thus, in the end, both thresholds change and the net effect on the sensitivity is hard to predict. Only by running new simulations with these thresholds, the sensitivity can be determined.

The data used for the simulations was derived from a population with a mean age that is lower than that of the typical glaucoma population. Although the simulations were performed for various individual eye related reproducibilities, thereby extending its domain to poorly reproducible eyes, there may be more subtle differences between the tested glaucoma populations. One possibility is the shape of the image power spectrum, which defines the way the noise is modeled. However, we expect that only large changes in this spectrum would significantly affect the reported minimally detectable losses. One may also expect the severity of glaucoma to influence the measurements. However, the variability of the NFL measurements is stable across disease severity [25,26].

Birefringence due to the anterior segment was compensated by application of ECC instead of the more conventional Variable Corneal Compensation (VCC). With the latter method, estimating anterior birefringence is time-consuming when repeated for each measurements. ECC does not suffer from this problem, and has shown to decrease the occurrence of 'atypical' scans [27,28] without adversely affecting the reproducibility [29]. In addition, ECC largely reduces the number of atypical retardation patterns, making it much more suitable for the detection of progression [30].

Irrespective of whether VCC or ECC is used, one problem for applying these methods in clinical settings is the increased number of images that have to be acquired. Currently, the device requires the operator to take one image of each eye. Moving the

scanner head, aligning and focusing requires much more time than the acquisition itself. Modifying the machine, allowing the operator to take multiple images of one eye, would therefore greatly reduce the extra time. Additionally, alternative strategies that reduce the number of required images at follow-up visits may be developed. For example, after the acquisition of three images at two baseline visits, only one image is required at each follow-up visit. Only if this single image suggests progression, a full set of three images would have to be acquired.

In addition to these variations, which focus on optimizing the trade-off between ease-of-use and performance in terms of sensitivity, specificity and minimally detectable loss, other detection algorithms may be applied as well, such as statistical image mapping [18]. Therefore, a side-by-side comparison of these methods may be the subject of future research. In those studies, the simulated images may again prove to be beneficial in the absence of a large-enough set of real images. Such a comparison may be extended to also include other imaging modalities commonly used for RNFL assessment in glaucoma, such as optical coherence tomography [31,32] after segmentation of the RNFL [33,34].

Recent software versions of the GDx include progression detection analysis based on the described detection method. The settings of this commercial implementation were optimized according to the procedure described in this paper. However, slightly different parameters were used than presented here (i.e., specificity was set to 98% and area threshold to 150 pixels). The progression detection algorithm in the commercially available software includes both suspected and confirmed progression, allows both repeated and single measurements per exam (the latter by including population-derived variability maps) and also applies progression detection to RNFL summary parameters (optimized with a set specificity of 99%) and RNFL circumferential profiles (optimized with a set specificity of 98% and a size threshold of 4 points) derived from the thickness images. The analysis always requires two baseline exams, but allows one (with the 2,1,2/2 strategy) or more follow-up visits (with the 2,2,3/4 strategy).

The most important limitation of this study is that the sensitivity of the method is only determined by simulated data. Also, the minimally detectable loss was assessed on simulated image series, while ideally these numbers should be derived from real data of eyes showing glaucomatous progression. If such data sets became available, the presented analyses could be applied to real images instead of simulated ones.

Our research enables an evaluation of progression detection strategies without running large-scale clinical trials. All these results were produced by simulated images instead of images of real eyes showing progression. The optimal settings for application on real data are likely to differ somewhat from the ones derived from these simulation studies. This research does show, however, that the 2,2,2/4 strategy outperforms all other tested strategies and that the expected minimally detectable loss for this strategy is in the range of $5-12$ μm. This warrants the application of the best strategies in a clinical setting to further optimize the parameters to detect glaucomatous progression.

Conflict of interest statement

None declared.

Acknowledgments

The authors would like to acknowledge the Eye Care of San Diego for their assistance in collecting the data. The authors also thank William Simons and Kate Zhou for collecting the data.

This research was funded by Carl Zeiss Meditec, Inc. and employees and consultants of the company were involved in all parts of this study.

References

[1] H.A. Quigley, Number of people with glaucoma worldwide, Br. J. Ophthalmol. 80 (1996) 389–393.
[2] L.M. Zangwill, C. Bowd, C.C. Berry, J. Williams, E.Z. Blumenthal, C.A. Sanchez-Galeana, C. Vasile, R.N. Weinreb, Discriminating between normal and glaucomatous eyes using the Heidelberg retina tomograph, GDx nerve fiber analyzer, and optical coherence tomograph and optical coherence tomograph, Arch. Ophthalmol. 119 (2001) 1069–1070.
[3] F.A. Medeiros, L.M. Zangwill, C. Bowd, R.N. Weinreb, 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. 122 (2004) 827–837.
[4] F.A. Medeiros, L.M. Alencar, L.M. Zangwill, C. Bowd, G. Vizzeri, P.A. Sample, R.N. Weinreb, Detection of progressive retinal nerve fiber layer loss in glaucoma using scanning laser polarimetry with variable corneal compensation, Invest. Ophthalmol. Vis. Sci. 50 (2009) 1675–1681.
[5] A. Tafreshi, P.A. Sample, J.M. Liebmann, C.A. Girkin, L.M. Zangwill, R.N. Weinreb, M. Lalezary, L. Racette, Visual function-specific perimetry to identify glaucomatous visual loss using three different definitions of visual field abnormality, Invest. Ophthalmol. Vis. Sci. 50 (2009) 1234–1240.
[6] J.L. Keltner, C.A. Johnson, J.M. Quigg, K.E. Cello, M.A. Kass, M.O. Gordon, Confirmation of visual field abnormalities in the ocular hypertension treatment study, Arch. Ophthalmol. 118 (2000) 1187–1194.
[7] H.A. Quigley, G.R. Dunkelberger, W.R. Green, Retinal ganglion cell atrophy correlated with automated perimetry in human eyes with glaucoma, Am. J. Ophthalmol. 107 (1989) 453–464.
[8] R.S. Harwerth, L. Carter-Dawson, E.L. Smith III, G. Barnes, W.F. Holt, M.L.J. Crawford, Neural losses correlated with visual losses in clinical perimetry, Invest. Ophthalmol. Vis. Sci 45 (2004) 3152–3160.
[9] N.J. Reus, H.G. Lemij, The relationship between standard automated perimetry and GDx VCC measurements, Invest. Ophthalmol. Vis. Sci. 45 (2004) 840–845.
[10] P.H. Artes, B.C. Chauhan, Longitudinal changes in the visual field and optic disc in glaucoma, Prog. Retinal Eye Res. 24 (2005) 333–354.
[11] A.I. McNaught, D.P. Crabb, F.W. Fitzke, R.A. Hitchings, Modelling series of visual fields to detect progression in normal–tension glaucoma, Graefes Arch. Clin. Exp. Ophthalmol. 233 (1995) 750–755.
[12] T. Fayers, N.G. Strouthidis, D.F. Garway-Heath, Monitoring glaucomatous progression using a novel Heidelberg retina tomograph event analysis, Ophthalmology 114 (2007) 1973–1980.
[13] W.L. Hays, R.L. Winkler, Statistics: Probability, Inference, and Decision, Holt, Rinehart and Winston, New York, NY, pp. 778–779.
[14] P.G.D. Spry, A.B. Bates, C.A. Johnson, B.C. Chauhan, Simulation of longitudinal threshold visual field data, Invest. Ophthalmol. Vis. Sci. 41 (2000) 2192–2200.
[15] S.K. Gardiner, D.P. Crabb, Examination of different pointwise linear regression methods for determining visual field progression, Invest. Ophthalmol. Vis. Sci. 43 (2002) 1400–1407.
[16] N.M. Jansonius, Progression detection in glaucoma can be made more efficient by using a variable interval between successive visual field tests, Graefes Arch. Clin. Exp. Ophthalmol. 245 (2007) 1647–1651.
[17] W. Adler, T. Hothorn, B. Lausen, Simulation based analysis of automated classification of medical images, Methods Inf. Med. 43 (2004) 150–155.
[18] A.J. Patterson, D.F. Garway-Heath, N.G. Strouthidis, D.P. Crabb, A new statistical approach for quantifying change in series of retinal and optic nerve head tomography images, Invest. Ophthalmol. Vis. Sci. 46 (2005) 1659–1667.
[19] K.A. Vermeer, F.M. Vos, B. Lo, Q. Zhou, H.G. Lemij, A.M. Vossepoel, L.J. van Vliet, Modeling of scanning laser polarimetry images of the human retina for progression detection of glaucoma, IEEE Trans. Med. Imaging 25 (2006) 517–528.
[20] R.W. Knighton, Q. Zhou, New techniques, in: M. Iester, D. Garway-Heath, H.G. Lemij (Eds.), Optic Nerve Head and Retinal Nerve Fibre Analysis, Dogma, Savona, Italy, 2005, pp. 117–119.
[21] K. Vermeer, F.M. Vos, H.G. Lemij, A.M. Vossepoel, A model based method for retinal blood vessel detection, Comput. Biol. Med. 34 (2004) 209–219.
[22] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical recipes, in: The Art of Scientific Computing, third ed., Cambridge University Press, New York, NY, USA, pp. 801–805.
[23] P.J. Airaksinen, S.M. Drance, G.R. Douglas, D.K. Mawson, H. Nieminen, Diffuse and localized nerve fiber loss in glaucoma, Am. J. Ophthalmol. 98 (1984) 566–571.
[24] A. Tuulonen, P.J. Airaksinen, Initial glaucomatous optic disk and retinal nerve fiber layer abnormalities and their progression, Am. J. Ophthalmol. 111 (1991) 485–490.
[25] J.E. DeLeón Ortega, L.M. Sakata, B. Kakati, G. McGwin Jr, B.E. Monheit, S.N. Arthur, C.A. Girkin, Effect of glaucomatous damage on repeatability of confocal scanning laser ophthalmoscope, scanning laser polarimetry, and optical coherence tomography, Invest. Ophthalmol. Vis. Sci. 48 (2007) 1156–1163.

[26] C.K. Leung, C.Y. Cheung, D. Lin, C.P. Pang, D.S. Lam, R.N. Weinreb, Longitudinal variability of optic disc and retinal nerve fiber layer measurements, Invest. Ophthalmol. Vis. Sci. 49 (2008) 4886–4892.
[27] M. Tóth, G. Holló, Enhanced corneal compensation for scanning laser polarimetry on eyes with atypical polarisation pattern, Br. J. Ophthalmol. 89 (2005) 1139–1142.
[28] M. Tóth, G. Holló, Evaluation of enhanced corneal compensation in scanning laser polarimetry, J. Glaucoma 15 (2006) 53–59.
[29] M. Sehi, D.C. Guaqueta, D.S. Greenfield, An enhancement module to improve the atypical birefringence pattern using scanning laser polarimetry with variable corneal compensation, Br. J. Ophthalmol. 90 (2006) 749–753.
[30] F.A. Medeiros, L.M. Alencar, L.M. Zangwill, P.A. Sample, R. Susanna Jr, R.N. Weinreb, Impact of atypical retardation patterns on detection of glaucoma progression using the GDx with variable corneal compensation, Am. J. Ophthalmol. 148 (2009) 155–163.
[31] D.C. Hood, R.H. Kardon, A framework for comparing structural and functional measures of glaucomatous damage, Prog. Retinal Eye Res. 26 (2007) 688–710.
[32] R.S. Harwerth, J.L. Wheat, N.V. Rangaswamy, Age-related losses of retinal ganglion cells and axons, Invest. Ophthalmol. Vis. Sci. 49 (2008) 4437–4443.
[33] M. Mujat, R. Chan, B. Cense, B. Park, C. Joo, T. Akkin, T. Chen, J. de Boer, Retinal nerve fiber layer thickness map determined from optical coherence tomography images, Opt. Express 13 (2005) 9480–9491.
[34] K.A. Vermeer, J. van der Schoot, H.G. Lemij, J.F. de Boer, Automated segmentation by pixel classification of retinal layers in ophthalmic OCT images, Biomed. Opt. Express 2 (2011) 1743–1756.