Journal of Theoretical Medicine, September–December 2003 VOL. 5 (3–4), pp. 161–170

Taylor & Francis
healthsciences

Application of Discriminant, Classification Tree and Neural Network Analysis to Differentiate between Potential Glaucoma Suspects With and Without Visual Field Defects*

W. HITZL[a,b,†], H.A. REITSAMER[a,c], K. HORNYKEWYCZ[a], A. MISTLBERGER[d,‡] and G. GRABNER[a]

[a]*Department of Ophthalmology and Optometry, Paracelsus University Salzburg, Müllner Hauptstraße 48, 5020 Salzburg, Austria;* [b]*Institute of Mathematics, University of Salzburg, Hellbrunnerstrasse 34, 5020 Salzburg, Austria;* [c]*Department of Physiology, University of Vienna, Vienna, Austria;* [d]*County Clinic for Ophthalmology and Optometry, St. Johanns-Spital, Salzburg, Austria*

(Received 31 May 2002; Revised 1 April 2004; In final form 13 May 2004)

Purpose: This study has two objectives. The first one is to investigate the question whether it is possible to discriminate between eyes with and without a glaucomateous visual field defect based on standard ophthalmologic examinations as well as optic nerve head topographic parameters. The second objective raises the question of the ability of several suggested statistical models to generalize their results to new, previously unseen patients.

Methods: To investigate the above addressed question: (a) independent, two-sided *t*-tests, (b) a linear discriminant analysis with a forward stepwise variable selection algorithm, (c) four classification tree analyses and (d) three different neural network models with a forward, backward and a genetic variable selection algorithm were applied to 1020 subjects with a normal visual field and 110 subjects with a glaucomateous visual field defect. The Humphrey Visual Field Analyzer was used to test the visual fields and the TopSS® Scanning Laser Tomograph measured the optic nerve topography. A 10-fold cross-validation method was used for the models (b), (c) and (d) to compute approximative 95% confidence intervals for the specificity and sensitivity rates.

A literature study of 18 studies dealt with the question whether/how the generalization error was controlled (control of sample bias, cross-validation procedures, training net size for valid generalization). It was also looked up whether point estimators or 95% confidence intervals were used to report specificity and sensitivity rates.

Results: (a) Only few differences of the means could be found between both groups, including age, existing eye diseases, best corrected visual acuity and visual field parameters. The linear discriminant analysis (b), the classification tree analyses (c) and the neural networks (d) ended up with high specificity rates, but low sensitivity rates.

The literature study showed that three models did not apply a cross-validation procedure to report their results. Two models used a test sample cross-validation and thirteen used a v-fold cross-validation method. Although most authors reported confidence intervals for the area under the ROC, no author reported confidence intervals for the true, but unknown sensitivity and specificity rates.

Conclusions: (a) The results of this study suggest that the combination of standard ophthalmologic eye parameters and optic nerve head topographic parameters of the TopSS® instrument are not sufficient to discriminate properly among normal eyes and eyes with a glaucomateous visual field defect. (b) We follow important suggestions given in statistical learning theory for proper generalization and suggest to apply these methods to the given models or to models in future. At least three conditions should be met: (1) confidence intervals instead of point estimators to assess the classification performance of a model (control of test sample bias); (2) sensitivity and specificity rates should be estimated instead of reporting a point estimator for the area under the ROC and (3) the generalization error should be reported both in a training and a test sample and methods should be applied to select an appropriate training sample size for valid generalization.

Keywords: Glaucoma; Visual field defect; TopSS®; Discriminant analysis; Classification tree analysis; Neural network

*An analysis in the Salzburg-Moorfields-Collaborative-Glaucoma-Study using visual field and optic nerve head topographic parameters.

†Corresponding author. Address: Department of Ophthalmology and Optometry, Paracelsus Private Medical University Salzburg, Müllner Hauptstraße 48, 5020 Salzburg, Austria. Tel.: + 43-448-258446. Fax: + 43-662-8044-137. E-mail: w.hitzl@lks.at

‡Private practice in Salzburg, Austria.

ISSN 1027-3662 print/ISSN 1607-8578 online © 2003 Taylor & Francis Ltd
DOI: 10.1080/10273360410001728011

INTRODUCTION

The original intention of the scanning laser tomograph (e.g. TopSS®, HRT®) and related instruments that investigate the structure of the optical nerve head (e.g. GDx®, OCT®) was to detect topographic/morphological changes of the optic nerve head while observing the course of the glaucomateous eye. Typical changes that are regarded as suspicious indicators are, e.g. a deeper excavation of the optic nerve head and/or a reduction of the neuroretinal rim area of the optic nerve head.

On one hand, several studies found moderate relations of topographic changes (TopSS®) and changes in the visual field. Ahn and Kee (2000) investigated 110 eyes at one point in time and suggested a model with high sensitivity and specificity in the diagnosis of glaucoma and the TopSS® can be useful in the early detection of changes in the glaucomatous optic disc. Cullinane *et al.* (2002) found the average slope of the TopSS® instrument as capable of discriminating OHT and POAG patients from normal subjects (investigated at a fixed time). This topographic parameter was also well correlated with the visual field mean defect. Chauhan *et al.* (2001) followed up 77 patients with glaucomatous visual field damages at the initial investigation over a period of about 5.5 years and could demonstrate that glaucomatous disc changes determined with scanning laser tomography occur more frequently than field changes. Most of these patients with field changes also had disc changes. However, less than half of those with disc changes had field changes. In greater detail, 27% showed no progression with either technique, 40% progressed with scanning laser tomography only, while 4% progressed with conventional perimetry only. There were 29% who progressed with both techniques, 45% progressed with scanning laser tomography first and 41% with conventional perimetry first, while 14% progressed at the same time. Lan *et al.* (2003) found relations between some topographic parameters, RNFL parameters and visual field indices (based on 62 patients investigated at a fixed time). However, great interindividual variation limited the prediction of one parameter from the other. They suggested the evaluation of both structural and functional aspects in order to obtain full characterization of the glaucomatous damage for clinical judgment and treatment. The results of theses studies are encouraging to find a statistical model for predicting a glaucomateous visual field defect.

However, on the other hand, there is a bulk of data which suggests that functional changes may not necessarily be linked with structural changes: there is also evidence to support the idea that the loss of ganglion cell axons can occur without the structural cupping of the optic nerve head, e.g. optic atrophy. Harwerth *et al.* (2002) studied the structure–function relationships from 12 monkeys with unilateral experimenal glaucoma and compared the ganglion cell loss (%) via the loss of sensitivity (dB). This experiment revealed that visual sensitivity losses were not correlated with retinal ganglion cell losses until a substantial number of neurons (about 50%) have been lost. Thus, ganglion cell losses lower than 50% cannot be detected with standard clinical perimetry.

In addition, statistical models were suggested to indicate functional changes, i.e. a deterioration of the visual field or a deterioration of status of the glaucomateous eye (Brigatti *et al.*, 1996; Uchida *et al.*, 1996; Weinreb *et al.*, 1998; Mardin *et al.*, 1999; Iester *et al.*, 2000; Nicolela *et al.*, 2001; Zangwill *et al.*, 2001; Bowd *et al.*, 2002; Greaney *et al.*, 2002).

In order to make a contribution to resolve these paradoxical findings, this study has two objectives. The first one is the attempt to answer the question: "Can a classifier based on a standard ophthalmologic eye parameters in combination with topographic parameters of the TopSS® scanning laser tomograph be found to detect eyes with a glaucomateous visual field defect?" The findings of Harwerth *et al.* (2002) suggest that the relation between structural and functional losses have to be considered with more care. However, it cannot be concluded that such a classifier does not exist. It is still possible that TopSS® and standard ophthalmologic parameters can discriminate between eyes with and without a glaucomateous visual field defect, independently of a beginning structural loss of ganglion cells.

The second objective raises the question whether/how the generalization error of the proposed statistical models was controlled (control of sample bias, cross-validation procedures, training net size for valid generalization). It was also looked up whether point estimators or 95% confidence intervals were used to report specificity and sensitivity rates. We refer to important insights and results suggested by authors working in the field of statistical learning theory. We follow these suggestions in order to make results more reliable and comparable.

PATIENTS AND METHODS

A cohort of 4629 subjects were enrolled in the Salzburg-Moorfields-Collaborative-Glaucoma Study (SMCGS) between December 1996 and July 2003. For details of the study, see Mistlberger *et al.* (1998). The following inclusion criteria had to be fulfilled: age ≥ 40 years, best spectacle corrected visual acuity $> 6/9$, refraction ranging from -6.00 to $+4.00$ dpts, difference of refraction < 3.00 dpts.

The exclusion criteria comprised the following conditions: pseudophakia, current glaucoma therapy, eye diseases with a potential for visual field defects (except glaucoma) or secondary increase of intraocular eye pressure, contraindications against beta blockers, systemic corticosteroid therapy or pregnancy. Eyes of patients which did not fulfill the first or the second condition, were removed from this study. All subjects underwent extensive ophthalmologic examinations including history of pre-existing eye diseases and eye therapy, family history (FH), refraction, visual acuity, intraocular eye pressure, slit-lamp and fundus examinations and assessment of subjective

C/D-ratios. All in all, 3228 patients fulfilled the inclusion and exclusion criteria.

Furthermore, the Visual Field Test Analyzer (Humphrey Visual Field Test Program, Humphrey Instruments, Inc., San Leandro, CA, USA) calculated the mean deviation (MD), corrected standard pattern deviation (CPSD) and the glaucoma hemifield test (GHT). The GHT assesses a visual field as normal, borderline or abnormal. A reliable test of the visual field was defined as having fewer than 33% fixation losses. Each eye indicating a glaucomateous visual field defect was retested after 3 and 6 months, to confirm the glaucomateous visual field defect. Eyes with a moderate visual field defect (glaucoma hemifield test was 'borderline' of Humphrey Visual Field Analyzer) were removed.

A short description of the optic nerve head parameters of the scanning laser tomograph (TopSS®, Laser Diagnostic Technologies, Inc., San Diego, CA, USA) is given in the appendix (7 optic nerve head parameters). Due to the amount of computations, only right eyes were analyzed. All eyes with a normal GHT were allocated to the first group (NGHT), eyes with an abnormal GHT were allocated to the second group (AGHT).

STATISTICAL METHODS

Description of the Sample

Casewise deletion of missing data: all patients with missing data in at least one variable were excluded from the analysis. After data cleaning, 1130 eyes could be included (NGHT: 1020, AGHT: 110).

Definition of Input Variables

All variables of Table I were used—except from MD and CPSD. If these have been measured then the GHT is also available providing a 100% correct classifier by definition, so clinically a classifier including these two measures is not useful. An existing eye disease was defined as follows: (0) no eye disease and (1) an eye disease is present. Family history was coded as follows: (0) there is no relative with an eye disease, (1) glaucoma is present only in the patient's siblings, (2) parents, (3) other relatives, (4) parents and siblings, (5) siblings and other relatives, (6) parents and other relatives and (7) an eye disease other than glaucoma is present.

The decision whether or not a proper identification of glaucomateous visual field defects based on all specified parameters is possible, is in general a question whether or not the posterior distribution functions differ sufficiently (Johnson and Wichern, 1999). To tackle this problem, four different statistical approaches were done in order to cover different aspects of the problem above.

Four Approaches were Carried out for this Classification Problem

(A) Univariate, independent, two-sided *t*-tests were applied to test the means in both groups. In order to

TABLE I Variables submitted to the analysis, except from MD and CPSD

		Glaucoma hemifield test				
		Normal (NGHT)		Abnormal (AGHT)		
		Mean	Std	Mean	Std	*p*-values
Standard parameters	Gender	(38,62)%*		(35,65)%		0.51
	Existing eye disease	(92,8)%†		(85,15)%		0.0011**
	Family history	(68,2,8,3,1,0,1,17)%‡		(64,1,11,2,1,0,1,20)%		0.87
	Age	59.8	8.9	67.3	8.4	< 0.001**
	Best corrected visual acuity	0.96	0.11	0.89	0.14	< 0.001**
	IOP (mmHg)	15.3	3.1	15.5	3.4	0.47
	Subj. C/D ratio	0.27	0.16	0.33	0.19	0.004
Visual field	Mean defect (MD)	−0.51	1.4	−3.6	4.4	< 0.001**
	Corrected pattern standard deviation (CPSD)	0.94	0.75	4.22	2.5	< 0.001**
Optic nerve head topographic analysis: TOPSS®	Total contour area (mm^2)	2.0	0.45	2.0	0.37	0.15
	Effective area (mm^2)	0.9	0.41	0.9	0.37	0.51
	Neuroretinal rim area (mm^2)	1.2	0.32	1.1	0.34	0.03
	Volume below (mm^3)	−0.3	0.19	−0.3	0.22	0.34
	1/2 Depth volume (mm^3)	0.3	0.21	0.4	0.19	0.25
	1/2 Depth area (mm^2)	−0.1	0.05	−0.1	0.06	0.38
	C/D ratio (TopSS®)	0.4	0.15	0.4	0.15	0.06

Means, standard deviations and corresponding *p*-values of all eye parameters in patient groups with normal (NGHT) and abnormal glaucoma hemifield test (AGHT).
* Male and female.
† Eye disease (not present and present).
**After Bonferroni adjustment, a difference is considered as statistically significant, if the corresponding *p*-value is smaller than $0.05/16 \cong 0.003$.
‡ See, "Statistical Methods" section.

estimate the magnitude of the effect under the alternative hypothesis, the standardized effect size is estimated (Cohen, 1988). This dimensionless number divides the difference of the expectation values by the common standard deviation and is independent of the sample size. If the effect size is zero, the corresponding null hypothesis cannot be rejected with any sample sizes. For our sample sizes ($n_1 = 1020$ and $n_2 = 110$), that effect size is estimated at which the independent *t*-test is statistically significant (ES_{crit}). The empirical effect size is the point estimator obtained by the concerning plug-in estimator, i.e. the observed mean estimates the expectation value, the empirical common standard deviation estimates the common standard deviation. Differences among means with empirical effect sizes larger than ES_{crit} are detected by the *t*-test, differences smaller than those are not. Comparisons of categorical variables were done with the Maximum Likelihood Chi-square statistic. It is well known that multiple comparisons demand the use of a type I error rate adjustment in order to protect against an increase of the overall type I error rate (Miller, 1981). The unadjusted type I error was set to 5%. The *p*-value adjustment was done with the Bonferroni-method.

(B) A linear discriminant analysis with a forward stepwise algorithm was applied to compare this frequently used method with other models (as defined below). Attention was given to the question whether the ratio of the number of variables to the number of eyes was adequate (about 20 or more, Huberty, 1975; Barcikowski and Stevens, 1975). Prior probabilities were set to 85% for the NGHT group and 15% for the AGHT group. These estimations are based on accumulate observations made in our glaucoma screening program.

(C) In order to use models with a hierarchical nature, four classification tree analyses (Breiman *et al.*, 1984) were done to compare the results with those of the discriminant analysis. The same prior probabilities as in approach B were used. Different split selection methods were used. *SL1*: discriminant based univariate splits and *SL2*: CART style for exhaustive search for univariate splits (Gini measure for goodness of fit). Also, different stopping rules were applied. *SP1*: pruning on misclassification error (1 SE rule) and *SP2*: FACT-style direct stopping (fraction of objects 10%). The following models were tested: (C.1) SL1 and SP1, (C.2) SL1 and SP2, (C.3) SL2 and SP1 and (C.4) SL2 and SP2. For details of these methods, see Breiman *et al.* (1984) and StatSoft, Inc. (1999).

(D) Finally, the following types of neural networks were tested: linear networks, radial basis function networks and three-layer perceptron networks (Bishop, 1995). Radial basis function networks with a hidden layer of radial units, each actually modeling a Gaussian response surface (center assignment was done by the k-means algorithm and each unit's deviation is individually set to the mean distance to its k nearest neighbors). Pre-processing involved conversion of nominal values and scaling of numeric values. The minimum and maximum of each input variable was found and scaling factors were selected so that these were mapped to 0 and 1. The normalized input values were then fed into the neural network. A forward, backward and genetic variable selection algorithm (StatSoft, Inc., 1999) was applied to determine an "optimal" set of input eye parameters. Binary masks were constructed which indicate which inputs to retain and which to discard. The network complexity (number of hidden units) was determined automatically by the software used (StatSoft, Inc., 1999). Levenberg-Marquardt (Bishop, 1995) was used for training of the three-layer perceptrons. After testing linear, radial basis function and three-layer perceptron networks with different architectures, the model with the best performance was selected. Doubt option: In first step, the accept and reject thresholds were set automatically such that the misclassification rate was minimized. If the activation was above the accept threshold, the eye was deemed to belong to the risk class (AGHT); if it was below the reject threshold, the eye was deemed to belong to the class without risk (NGHT), and if it was in between, the prediction is deemed to be 'unknown'. In second step, the accept thresholds were increased and the reject thresholds decreased step-by-step to avoid dubious classification, perhaps reflecting a point in areas of overlap between the two classes and to increase the classification rates. In order to learn more about the performance of the best model, the thresholds for accept and reject were readjusted to find models with high specificity and as high sensitivity as possible (model D.1) and high sensitivity rates and as high specificity as possible (model D.2).

Test of Performances of the Approaches

(A) *Data splitting into training, verification and test sample*: different approaches used different sample splits to check that a model was generalizing properly by observing whether the error in the test sample was reasonably low. Approach A did not use a sample split. A 10-fold cross-validation (Stone, 1974) procedure was applied to the approaches B and C. Data were split into training and test samples in the ratio of 9:1. Approach D used a training sample, a verification sample and a test sample in the ratio 6:3:1, respectively. The verification set was used to track the neural network's error performance during training, to identify the best network and to stop training, if over-learning occurred (early stopping method of training, Morgan and Bourlard, 1990; Amari *et al.*, 1996). The test set was not used in training at all and was designed to give an independent assessment of the network's performance when an entire network design procedure was completed. (B) *Sufficient large training sample size for valid generalization*: different methods to estimate the training sample size were applied to obtain models with valid generalization. Firstly, a method suggested by Vapnik and Chervonenkis (1971) was applied that estimates the worst-case measure of generalization performance. This measure estimates the maximum discrepancy which can occur between

generalization performance estimated from the sample and the true generalization. Secondly, a method suggested by Baum and Haussler (1989) was applied for single-hidden-layer feedforward neural networks with k units and d weights. As suggested by Baum and Haussler (1989), such a network that has been trained on m examples so that at least a fraction of $1 - (\varepsilon/2)$, where $0 < \varepsilon < 1/8$ of the examples were correctly classified, the network will—with a probability approaching 1—correctly classify the fraction $1 - \varepsilon$ of future random test samples drawn from the same distribution as long as $m \geq O(d/\varepsilon \log_2(k/\varepsilon))$. Thirdly, a method suggested by Haykin (1998) was applied that is similar to Widrow's rule of thumb. This method suggests 'good generalization', if the condition $m = O((d + k)/\varepsilon)$ is satisfied. Further details of the described methods are given in Vapnik and Chervonenkis (1971), Baum and Haussler (1989), Bishop (1995) and Haykin (1998). Based on the results of the 10-fold cross-validation, 95% confidence intervals for the true, but unknown sensitivity and specificity rates of the models were computed (Pearson-Clopper values; Hartung, 1993). All computations and illustrations were done with STATISTICA 5.5 (StatSoft, Inc., 1999) and MATHEMATICA 3.0.1 (Wolfram, 1996).

RESULTS

Results of the Study with the Scanning Laser Tomograph (TopSS®)

(A) The percentage of subjects with an eye disease was slightly higher in AGHT than in NGHT (95% CI: 5.5–8.5%). Age was higher in AGHT than in NGHT (95% CI: 5.9–9.3 years). The best corrected visual acuity was slightly better in NGHT (95% CI: 0.04–0.09). The MD was considerably higher in NGHT (95% CI: 2.3–3.9 db). The CPSD was considerably higher in AGHT (95% CI: 2.8–3.8 db). The observed means and standard deviations are given in Table I. No further statistically significant differences could be detected (Table I). Variables with an effect size larger than 0.38 are statistically significantly different (Fig. 1). An overview of the performance of the models with approximative 95% confidence intervals of the sensitivity, specificity, overall correct classification rates and percentages of unclassified eyes is given in Table II for the approaches B, C.1, C.2, C.3, C.4, D.1 and D.2. A table that characterizes the best three layer perceptron network is given in Table III. A table with the corresponding weights and thresholds are given in Table IV.

The method suggested by Vapnik and Chervonenkis (1971) for the worst-case generalization suggested an inappropriate large sample size. The method of Baum and Haussler (1989) suggested a training sample size of $m = 1580$ for $\varepsilon = 1/8$, $k = 12$ and $d = 30$. The method of Haykin (1998) suggested for the above values a training sample size of $m = 240$.

Results of the Literature Study

A table with an overview of important contributions in this field is given in Table V. This table lists the applied statistical models, the instruments to detect structural changes of the eye, whether eyes with a moderate risk for glaucoma/moderate glaucomateous visual field defect were excluded and the number of the used variables. In addition, it indicates the observed sensitivity and

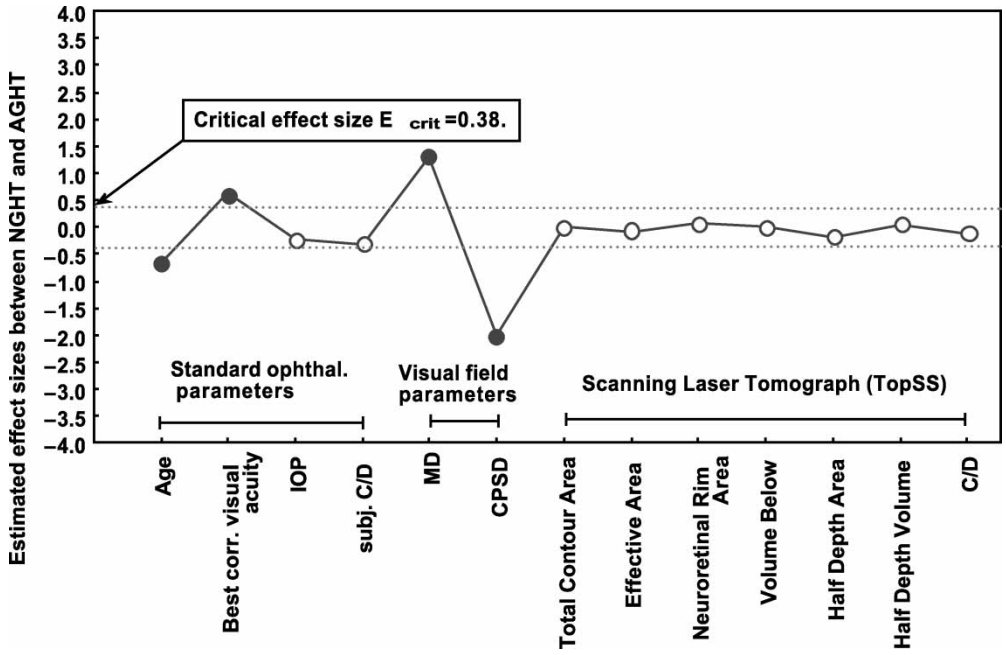

FIGURE 1 Line plot of the effect sizes of 13 eye parameters of eyes with normal and glaucomateous visual field. Labeled eye parameters are statistically significantly different and have high univariate separation power.

TABLE II Number of eyes in the training sample, number of variables used and approximative 95% confidence intervals with lower and upper limits for different performance parameters for the corresponding approach

Approach	Number of eyes in the training sample	Number of variables included in the final model	Specificity		Sensitivity		Overall correct classification rate		Unclassified (%)	
			Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Linear discriminant analysis B	1017*	13	98	99	4	16	88	92	0	0
Classification tree model C.1	1017*	0	99.6	100	0	3	88	92	0	0
Classification tree model C.2	1017*	0	99.5	100	0	3	88	92	0	0
Classification tree model C.3	1017*	0	99.6	100	0	3	88	92	0	0
Classification tree model C.4	1017*	3	99.3	100	0.6	7.8	88	92	0	0
Three layer percpetron model D.1	678	6	98.1	99.4	6.9	19.9	91.1	94.2	50.1	53.8
Three layer perceptron model D.2	678	6	14.9	18.4	93.5	97.8	27.5	31.4	21.1	24.2

* No verification sample.

specificity rates and the type of cross-validation that was used. Based on the given results, lower 95% confidence limits for specificity and sensitivity are given by the authors of this study. The literature study turned out that the models in Table V were tested with different extent of accuracy. There were three out of 18 models that did not use any cross-validation procedure. There were two models that were cross-validated in a test sample: Zangwill *et al.* (2001) applied the model of Weinreb *et al.* (1998) to 50 healthy subjects and 41 patients with glaucoma to estimate the classification rates in their sample. While Weinreb *et al.* (1998) suggested an observed sensitivity of 74%, Zangwill *et al.* (2001) suggested an observed sensitivity rate of 54% for the same statistical model, i.e. the rate was 20% lower than those given by Weinreb *et al.* (1998). An approximative 95% confidence interval for the true sensitivity rate based on the 41 patients in the test sample ranges from 38 to 69% and shows that a considerable loss of accuracy (about 31%) is entailed. Iester *et al.* (2000) applied the model of Mikelberg *et al.* (1995) and Bathija *et al.* (1998a,b) (Table V, 8a and 8b). A 95% confidence interval for the true sensitivity rate based on the 61 patients ranges from 74 to 93% (for Mikelberg *et al.*, 1995). The same confidence interval holds for the model of Bathija *et al.* (1998a,b). All other models listed in Table V used a k-fold cross-validation method to compute the standard error/95% confidence interval of the area under the ROC. No author reported 95% confidence intervals for the true but unknown specificity and sensitivity rates for a fixed cut-off.

TABLE III Details of the three-layer perceptron network model

Event to be predicted	Abnormal glaucoma hemifield test
Sample sizes of subsamples in training, verification and test sample	(678,339,113)
Number of networks tested to find a model with "good" network architecture	35
Type of network	Three-layer perceptron model
Feature selection	Forward, backward and genetic algorithm
Type of training algorithm used	Levenberg-Marquardt
Maximal number of epochs	100
Cross verification in verification sample to identify best network and stop training, if over-learing occurred	Yes
k-fold cross-validation	Yes, 10-fold
Number of input neurons	6
Number of hidden neurons	5
Error function	Sum-squared
Layer 1: PSP-function*	Linear
Layer 1: Activation function	Linear
Layer 2: PSP-function	Linear
Layer 2: Activation function	Logistic
Layer 3: PSP-function	Linear
Layer 3: Activation function	Logistic
Observed correct classification rate in the training sample	97.2%
Observed correct classification rate in the test sample	92.6%
Threshold for acceptance	0.9 for D.1 and 0.05 for D.2
Threshold for rejectance	0.03 for D.1 and 0.02 for D.2

* Post-synaptic potential function.

DISCUSSION

Study with Scanning Laser Tomograph (TopSS®)

Discussion of the Results

Approach A showed that—besides the MD and CPSD of the Humphrey Visual Field Test Analyzer—only age, existing eye diseases and the best corrected visual acuity were statistically significantly different. However, the confidence intervals showed that the means of those variables are close together. No parameter of the TopSS® scanning laser tomograph could be shown as statistically significant among both the groups (Table I). The critical effect size ES_{crit} was 0.38. An effect size of 0 indicates that the distribution of the first group overlaps completely the distribution of the other group, i.e. there is 0% of nonoverlap. An effect size of 0.5 indicates a nonoverlap of 33% in the two distributions. So, the effect size suggests a considerable overlap of most marginal distributions. This result can be considered as a first indication that a proper discriminiation of both distributions is difficult—if at all—to achieve and that the parameters of the scanning laser tomograph (TopSS®) may have relative small univariate separation power. Although a high overall

TABLE IV Input variables used, weights and thresholds for the input and hidden layer of the best three-layer perceptron model

		Hidden node 1	Hidden node 2	Hidden node 3	Hidden node 4	Hidden node 5
Input layer	Threshold	0.2219908	0.4324216	0.7126304	−0.646	−0.7253
	Age (years)	−1.301592	0.7465843	−0.02862	−3.167576	−3.068769
	Eye disease	−0.6893	−0.4334	−0.346	1.050109	−0.2273
	Family history	0.4558598	−0.1097	−0.6847	−1.524361	−0.009787
	IOP (mmHg)	−0.9422	0.9893949	0.3516825	−1.486126	−0.1271
	Total Contour Area (mm^2)	−0.901	−0.6168	−0.08261	0.04717	0.9559903
	Neuroretinal Rim Area (mm^2)	0.4983845	−0.6972	0.4459717	1.39762	1.822461
Hidden layer	Threshold					
	0.002056	−1.326357	1.236196	0.05219	−3.620329	−3.64186

correct classification rate (95% CI: 88–92%) could be achieved in B, the sensitivity rate of this approach was small (95% CI: 4–16%). The approaches C.1, C.2, C.3 selected trivial classifiers with 0 input variables with the given settings, the approach C.4 selected three input variables. The results of these four approaches showed very similar classification rates as approach B (Table II). The comparison of linear, radial basis function and three-layer perceptron networks suggested a three-layer perceptron network with 6 input units and 5 hidden units as to be useful. The results of the genetic algorithm was compared with those of the forward and backward stepwise variable selection algorithms. Age, eye disease and best corrected visual acuity were selected by the stepwise algorithms as well as the genetic algorithm. However, the algorithms differed in the remaining variable combinations. This suggests that there is too less structure in the data such that both groups can be properly discriminated. After adjusting the accept and reject thresholds to achieve a model with a high specificity rate and an as high sensitivity rate as possible, the model D.1 could be found (accept threshold: 0.9, reject threshold: 0.03). The specificity rate was very high (95% CI: 98.1–99.4%). The sensitivity rate, however, was low (95% CI: 6.9–19.9%). About 52% of all eyes were unclassified (95% CI: 50.1–53.8%). The model D.2 (accept threshold: 0.05, reject threshold: 0.02) could be found with very good sensitivity rate (95% CI: 93.5–97.8%), but the specificity rate considerably decreased (95% CI: 27.5–31.4%). The observed overall classification rate in the training sample was 23% and remained stable in the test sample with 29.5%. About 22.6% of all eyes were unclassified (95% CI: 21.4–24.2%). No model could be found with both, a high specificity and a high sensitivity rate (e.g. lower confidence limits larger than 90% for both rates).

The results of the four approaches suggest that knowledge of the standard ophthalmologic and TopSS® parameters is not sufficient to classify an eye as having a glaucomateous visual field defect. In general, we suspect that functional losses (e.g. glaucomateous visual field defects) will be difficult—if at all—to be detected based on structural changes measured with the current instrument.

Discussion of the Methods for Generalization

The observed overall classification rate in the training set of D.1 was 97.2%. Although the method of Baum and Haussler (1989) could not be applied, this method suggested a correct classification rate of 94.6%. In fact, 92.6% were correctly classified in the test sample. This is in accordance to the method as described by Haykin (1998).

One might speculate whether other classifiers (e.g Gaussian Kernel support vector machines, comittee machines) could show better performance. Although this is theoretically possible, we suppose—based on the current findings—that these models will not perform significantly better.

Further Remarks Concerning the Study Design

We would like to emphasize that eyes with a suspect glaucoma hemifield test were not included in this study. We expect that the inclusion of these eyes will make it much more difficult to discriminate properly among these three patient groups (normal, borderline, abnormal glaucoma hemifield test).

Literature Study

Remarks to the Results

Most authors (beside Brigatti *et al.*, 1996) reported only point estimators for the specificity and sensitivity rates. The *k*-fold cross-validation procedure was used by almost all authors for computation of standard errors or confidence intervals for the area under the ROC. No author reported 95% confidence intervals for the true but unknown specificity and sensitivity for a fixed cut-off. However, more efforts should be made to control the generalization abilities of the classifiers used, because the specificity and sensitivity rates of the studies in Table V are not reported in the training and test samples separately. Haykin (1998) and many other authors emphasize that there are various sources to bias a point estimator in connection with neural networks. If sensitivity or specificity rates are estimated based on training data alone, nothing can be said of how the model performs when faced with new data, previously unseen (variance/bias dilemma). If sensitivity or specificity

TABLE V An overview of 18 studies in this field together with some characteristics of the corresponding models

No.	Study	Model	Instrument used	Subset excluded*	Number of variables	Total sample size	Observed specificity rates (%)	Observed sensitivity rates (%)	Cross-validation	Estimated lower 95% confidence limit for specificity (%)	Estimated lower 95% confidence limit for sensitivity (%)
1	Uchida *et al.* (1996)	Three layer nn	*HRT*	Yes	9	43† + 53‡	92	91	3-fold	80	80
2	Brigatti *et al.* (1996)	Four layer nn	ONHP	Yes	13	54 + 185	87 (SE: 6%)	56 (SE: 4%)	4-fold	75	49
3	Weinreb *et al.* (1998)	LDA	GDx	Yes	3	84 + 83	92	74	12-fold	84	63
4	Mardin *et al.* (1999)	LDA	HRT	Yes	6	50 + 61	95	42	No	–	–
5	Zangwill *et al.* (2001)	LDA	GDx	Yes	3	50 + 41	90	54	Tested the model of Weinreb *et al.* (1998)	78	38
6	Nicolela *et al.* (2001)	Logistic regression	GDx	No	11	32 + 60	69	94	No	–	–
7	Bowd *et al.* (2002)	LDF Bathja *et al.*	HRT	Yes	4	189 + 108	90	67	10-fold	85	57
7b	Bowd *et al.* (2002)	LDF Mardin *et al.*	HRT	Yes	6	189 + 108	90	70	10-fold	85	60
7c	Bowd *et al.* (2002)	LDF Iester *et al.*	HRT	Yes	8	189 + 108	90	69	10-fold	85	59
7d	Bowd *et al.* (2002)	LDF Mikelberg *et al.*	HRT	Yes	4	189 + 108	90	64	10-fold	85	54
7e	Bowd *et al.* (2002)	SVM Gaussian	HRT	Yes	83	189 + 108	90	83	10-fold	85	75
7f	Bowd *et al.* (2002)	MLP	HRT	Yes	83	189 + 108	90	78	8-fold	85	69
7g	Bowd *et al.* (2002)	SVM Gaussian + forward selection	HRT	Yes	31	189 + 108	90	91	10-fold	85	84
8a	Iester *et al.* (2000)	LDA	HRT	Yes	3	194 + 61	65	85	Tested the model of Mikelberg *et al.* (1995)	58	74
8b	Iester *et al.* (2000)	LDA	HRT	Yes	4	194 + 61	75	85	Tested the model of Bathija *et al.* (1998a,b)	68	74
8c	Iester *et al.* (2000)	LDA	HRT	Yes	8	194 + 61	92	70	No	–	–
9	Greaney *et al.* (2002)	LDA	ONHP, SLP, OCT	Yes	37	63 + 63	96	97	Jackknife	88	89
10	Lauande-Pimentel *et al.* (2001)	LDA	GDx	yes	4	91 + 94	82	94	*k*-fold	73	87

Lower 95% confidence limits for specificity and sensitivity are estimated by the authors of this analysis.
* Subset of eyes with moderate glaucomateous visual field was removed.
† Healthy eyes.
‡ Glaucoma (suspect) eyes.

rates are estimated by a point estimator obtained by cross-validating the sample, one gets a (approximative) unbiased estimation of the sensitivity and specificity rates. However, in order to be able to draw statistically convincing conclusions, it is important to estimate the uncertainty around the error estimate. Such an estimation should be done by estimating the lower limit of an (at least approximative) 95% confidence intervals. Such estimations usually reveal that the model falls short of the expectations, especially, in case of small sample sizes. This situation can be observed in Table V. The differences between the observed sensitivity rate and the lower limit of the approximative confidence interval range between 5 and 12%, the corresponding difference for the specificity rate ranges between 7 and 16%.

Suggestions to Improve the Generalization Abilities and for Reporting Classification Results

We refer to important insights and results concerning learning and generalization as suggested by Vapnik and Chervonenkis (1971), Stone (1974), Morgan and Bourlard (1990) and Amari *et al.* (1996). These methods are described in a broader context by Bishop (1995) and Haykin (1998). Important methods to build models of the underlying process which generates the data are, e.g. early stopping method of training, training with noise, weight elimination, regularization methods, comittees of networks, test sample or *k*-fold cross-validation.

One of the most important issues—the computation of (approximative) 95% confidence intervals for the true, but unknown specificity and sensitivity rates—is still not considered to its full necessity for glaucomateous visual field defects/glaucomateous eyes. We suggest interval estimators for the true, but unknown sensitivity and specificity rates for a fixed cut-off rather than for the area under the ROC. Knowledge of such an area (e.g. 0.8 and 0.9) does not allow drawing convincing conclusions about the number of correct classified healthy and abnormal eyes.

General Remarks

More efforts should be made to improve the generalization ability of possible models and interval estimators should be reported to give a more realistic picture of the true performance of the classifiers. The results based on the TopSS® instrument are discouraging, if they are applied to one point in time. In order to learn more about structural and functional losses, we suggest the application of an instrument as sensitive as possible for identification of structural losses over time. This might include the use of objective electrophysiological methods (ERG) (Harwerth *et al.*, 2002) or ultrahigh-resolution optical coherence tomography (Drexler *et al.*, 2003).

A study design that applies such an instrument at the initial investigation, measures the changes within a short period of time (e.g. one or two years) and tries to predict visual field defects at, e.g. the five year follow-up, might have better chances to find a statistical classifier with sufficient high sensitivity and specificity rates. In any cases, the generalization ability of such a model should be tested as carefully as possible.

CONCLUSIONS

(A) The results of this study indicate that the posterior distribution functions of eyes with/without a glaucomateous visual field defect are very similar. An identification of eyes with a glaucomateous visual field defect based on TopSS® is hard or even impossible. The inclusion of eyes with a suspect glaucomateous visual field will even make it more difficult to find a classifier with sufficient discrimination power in this situation. The results suggest that the moderate relationships between TopSS® and visual field parameters (Cullinane *et al.*, 2002; Lan *et al.*, 2003) are not sufficient for a prediction of a glaucomateous visual field defect of individual eyes.

(B) We follow important suggestions given in statistical learning theory for proper generalization and suggest the application of these methods to the given models or to models in future. At least three conditions should be met: (1) confidence intervals instead of point estimators to assess the classification performance of a model (control of sample bias); (2) sensitivity and specificity rates should be estimated instead of reporting a point estimator for the area under the ROC. If a model is applied to a patient's eye in daily clinical practice, it has to prove its performance with a fixed cut-off value or threshold to accept and threshold to reject; (3) control of the size of the training sample size for valid generalization: the results of the literature study suggest that the generalization error should be reported both in a training and a test sample and methods should be applied to select a appropriate training sample size for valid generalization.

Acknowledgements

The authors wish to cordially thank the permanently and extremely ambitious co-workers at the "Glaukom-Vorsorgeambulanz" of the Landesaugenklinik Salzburg, Mrs Anna Konitsch and Mrs Anneliese Prieler and the many other clinical coworkers over the time of this study. We would also like to thank the Health Department of the Government of the County of Salzburg (Director: LH-Stv. Mag. Gabi Burgstaller), the 'Fond Gesundes Österreich' and the "Hauptverband der Sozialversicherungsträger" of Austria for their generous support.

References

Ahn, B.S. and Kee, C. (2000) "Ability of a confocal scanning laser ophthalmoscope (TopSS®) to detect early glaucomatous visual field defect", *British Journal of Ophthalmology* **84**(8), 852–855.

Amari, S., Murata, N., Müller, K.R., Finke, M. and Yang, H. (1996) *Statistical learning theory of overtraining. Is cross-validation*

asymptotically effective? Advances in Neural Information Processing Systems (MIT Press, Cambridge, MA), Vol. 8, pp 176–182.
Barcikowski, R. and Stevens, J.P. (1975) "A Monte Carlo study of the stability of canonical correlations, canonical weights and canonical variate-variable correlations", *Multivariate Behavioral Research* **10**, 353–364.
Bathija, R., Gupta, N., Zangwill, L. and Weinreb, R.N. (1998a) "Changing definition of glaucoma", *Journal of Glaucoma* **7**, 165–177.
Bathija, R., Zangwill, L., Berry, C.C., Sample, P.A. and Weinreb, R.N. (1998b) "Detection of early glaucomateous structural damage with confocal scanning laser tomography", *Journal of Glaucoma* **7**, 121–127.
Baum, E. and Haussler, D. (1989) "What size net gives valid generalization?", *Neural Computation* **1**(1), 151–160.
Bishop, C. (1995) *Neural Networks for Pattern Recognition* (Oxford University Press, New York), pp 333–380.
Bowd, C., Chan, K., Zangwill, L.M., Goldbaum, M.H., Lee, T.W., Sejnowski, T.J. and Weinreb, R.N. (2002) "Comparing neural networks and linear discriminant functions for glaucoma detection using confocal scanning laser ophthalmoscopy of the optic disc", *Investigative Ophthalmolology Visual Science* **43**(11), 3444–3454.
Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984) *Classification and Regression Trees* (CRC Press, New York).
Brigatti, L., Hoffman, D. and Caprioli, J. (1996) "Neural networks to identify glaucoma with structural and functional measurements", *American Journal of Ophthalmology* **121**(5), 511–521.
Chauhan, B.C., McCormick, T., Nicoleta, M. and LeBlanc, R.P. (2001) "Optic disc and visual field changes in a prospective longitudinal study in patients with glaucoma", *Archives of Ophthalmology* **119**, 1492–1499.
Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*, 2nd Ed. (Lawrence Earlbaum Associates, Hillsdale, NJ).
Cullinane, A.B., Waldock, A., Diamond, J.P. and Sparrow, J.M. (2002) "Optic disc cup slope and visual field indices in normal, ocular hypertensive and early glaucomatous eyes", *British Journal of Ophthalmology* **86**, 555–559.
Drexler, W., Sattmann, H., Hermann, B., Ko, T.H., Stur, M., Unterhuber, A., Scholda, C., Findl, O., Wirtitsch, M., Fujimoto, J.G. and Fercher, A.F. (2003) "Enhanced visualization of macular pathology with the use of ultrahigh-resolution optical coherence tomography", *Archives of Ophthalmology* **121**(5), 695–706.
Greaney, M.J., Hoffman, D.C., Garway-Heath, D.F., Nakla, M., Coleman, A.L. and Caprioli, J. (2002) "Comparison of optic nerve imaging methods to distinguish normal eyes from those with glaucoma", *Investigative Ophthalmology of Visual Science* **43**(1), 140–145.
Harwerth, R.S., Crawford, M.L.J., Frishman, L.J., Viswanathan, S., Smith, III, E.L. and Carter-Dawson, L. (2002) "Visual field defects and neural losses from experimental glaucoma", *Progress in Retinal and Eye Research* **21**, 91–125.
Haykin, S. (1998) *Neural Networks: A Comprehensive Foundation*, 2nd Ed. (Prentice Hall), pp 205–226.
Huberty, C.J. (1975) "Discriminant analysis", *Review of Educational Research* **45**, 543–598.
Iester, M., Jonas, J.B., Mardin, C.Y. and Budde, W.M. (2000) "Discriminant analysis models for early detection of glaucomatous optic disc changes", *British Journal of Ophthalmology* **84**, 464–468.
Johnson, R. and Wichern, D. (1999) *Applied Multivariate Statistical Analysis*, 4th Ed. (Prentice Hall, New Jersey), pp 629–665.
Lan, Y.W., Henson, D.B. and Kwartz, A.J. (2003) "The correlation between optic nerve head topographic measurements, peripapillary nerve fibre layer thickness and visual field indices in glaucoma", *British Journal of Ophthalmology* **87**, 1135–1141.
Lauande-Pimentel, R., Carvalho, R.A., Oliveira, H.C., Goncalves, D.C., Silva, L.M. and Costa, V.P. (2001) "Discrimination between normal and glaucomatous eyes with visual field and scanning laser polarimetry measurements", *British Journal of Ophthalmology* **85**(5), 586–591.
Mardin, C.Y., Horn, F.K., Jonas, J.B. and Budde, W.M. (1999) "Preperimetric glaucoma diagnosis by confocal scanning laser tomography of the optic disc", *British Journal of Ophthalmology* **83**(3), 299–304.
Mikelberg, F.S., Parfitt, C.M., Swindale, N.V. and Graham, S.L. (1995) "Ability of the Heidelberg Retina Tomograph to detect early glaucomatous visual field loss", *Journal of Glaucoma* **4**, 242–247.
Miller, J. (1981) *Simultaneous Statistical Inference*, 2nd Ed. (Springer Verlag, New York), pp 5–10.
Mistlberger, A., Sitte, S., Ruckhofer, J., Raithel, E., Alzner, E., Grabner, G. and Wormald, R. (1998) "The Salzburg-Moorfields-Collaborative Glaucoma Study (SMCGS)—Das Konzept und erste Erfahrungen bei der Umsetzung [The concept and first findings]", *Spektrum der Augenheilkunde* **12**(3), 93–96.
Morgan, N. and Bourlard, H. (1990) "Continuous speech recognition using multilayer perceptrons with hidden Markov models", *IEEE International Conference on Acousitcs, Speech and Signal Processing* **1**, 413–416.
Nicolela, M.T., Martinez-Bello, C., Morrison, C.A., LeBlanc, R.P., Lemij, H.G., Colen, T.P. and Chauhan, B.C. (2001) "Scanning laser polarimetry in a selected group of patients with glaucoma and normal controls", *American Journal of Ophthalmology* **132**(6), 845–854.
StatSoft, Inc. (1999) *STATISTICA for Windows [Computer program manual]* (StatSoft, Inc., Tulsa, OK).
Stone, M. (1974) "Cross-validatory choice and assessment of statistical predictions", *Royal Statistical Society* **B36**, 111–147.
Uchida, H., Brigatti, L. and Caprioli, J. (1996) "Detection of structural damage from glaucoma with confocal laser image analysis", *Investigative Ophthalmology Visual Science* **37**(12), 2393–2401.
Vapnik, V.N. and Chervonenkis, A.Y. (1971) "On the uniform convergence of relative frequencies of events to their probabilities", *Theory of Probability and its Applications* **16**(2), 264–280.
Weinreb, R.N., Zangwill, L., Berry, C.C., Bathija, R. and Sample, P.A. (1998) "Detection of glaucoma with scanning laser polarimetry", *Archives of Ophthalmology* **116**(12), 1583–1589.
Wolfram, S. (1996) *The Mathematica Book*, 3rd Ed. (Cambridge University Press, Wolfram Media, New York).
Zangwill, L.M., Bowd, C., Berry, C.C., Williams, J., Blumenthal, E.Z., Sanchez-Galeana, C.A., Vasile, C. and Weinreb, R.N. (2001) "Discriminating between normal and glaucomatous eyes using the Heidelberg retina tomograph, GDx nerve fiber analyzer and optical coherence tomograph", *Archives of Ophthalmology* **119**(7), 985–993.

APPENDIX

Scanning laser tomograph parameters (TopSS®): The total area is the area within the user-drawn contour area, the effective area is the cup area located 100 microns below the total area, the neuroretinal rim area is the difference between the total area and the effective area, the volume below is the volume of the cup below the effective area, the half depth area is the area at a height located halfway between the average height along the perimeter of the user-drawn contour area and the deepest points of the cup, the half depth volume is the volume of the cup below the half depth area, the cup to disc ratio is the ratio between the effective area and the total area. The units of measure of all parameters are mm, mm^2, mm^3 or for distances, areas, volumes or unitless for ratios, respectively.