A Comparison of Different Prediction Models in Glaucoma Screening

First Year Analysis in the Salzburg-Moorfields-Collaborative-Glaucoma-Study using Visual Field, Optic Nerve Head Topographic and Nerve Fiber Layer Parameters

WOLFGANG HITZL*, ANDREA MISTLBERGER † and GUENTHER GRABNER‡

Landesklinik für Augenheilkunde und Optometrie, St. Johanns-Spital, Landeskliniken Salzburg, Müllner Hauptstraße 48, 5020 Salzburg, Austria

(Received 29 October 2002; In final form 4 August 2003)

Purpose: This study analyses the possibility of risk-assessment the risk of developing glaucoma over the period of 1 year based on the “Number” (provided by the Nerve Fiber Analyzer) by using a multitude of parameters that are gathered by several different eye examination techniques at the initial investigation.

Methods: Within the above-mentioned study, a total of 336 patients with an increased risk of having glaucoma were analyzed. The complete ophthalmological examination included biomicroscopy of the optic nerve head, achromatic automated perimetry (Humphrey Field Analyzer), quantitative disc (Topographic Scanning Systems, TopSS®) and nerve fiber layer measurements (Nerve Fiber Layer Analyzer, GDx®) at the beginning of the study and 1 year thereafter. Three visual field parameters (mean deviation, corrected pattern standard deviation, glaucoma hemifield test), 7 topographic and 19 polarimetric parameters were used for these statistical analyses.

The problem was considered as a regression problem (RP) as well as a classification problem (CP): the simplest predictor, that is the “Number” at the initial investigation (CP), linear discriminant analyses without and with a forward stepwise variable selection algorithm (CP), four different classification tree analyses (CP) and different types of neural networks: regression networks (RP), linear networks (CP) and three layer perceptron networks (CP) with various variable selection algorithms and network architectures were applied in order to build models with sufficient prediction power. All models, except the simple predictor were tested with independent test set data, to ensure first a generalization for new patients and secondly that the results are not artifacts of the training process. The performance of the models was measured by sensitivity and specificity rates for CPs, multiple correlation coefficients between predicted and actual outcome for regression networks in each of the samples. Due to the large amount of computations, the models were computed for right eyes only.

Results: The simple predictor showed a specificity rate of 73% (95% CI: 65–80%) based on all observations. The following specificity rates could be found in the test samples: the linear discriminant analysis (LDA) without variable selection algorithm could not be applied, LDA plus variable selection algorithm: 85% (95% CI: 75–93%), four different models based on classification tree analyses: 87% (95% CI: 70–96%), 90% (95% CI: 74–98%), 87% (95% CI: 70–96%) and 71% (95% CI: 52–86%); linear neural networks (not all eyes were classified due to the doubt option) 95% (95% CI: 75–100%) and three layer perceptron network (also with doubt option) 100% (95% CI: 81–100%). The simple predictor showed a sensitivity rate of 76% (95% CI: 69–83%) based on all observations. The following sensitivity rates were observed in the test samples: the LDA without variable selection algorithm could not be applied, LDA plus variable selection algorithm: 61% (95% CI: 49–72%), four different models based on classification tree analyses: 63% (95% CI: 47–77%), 58% (95% CI: 42–72%), 63% (95% CI: 47–77%) and 61% (95% CI: 44–75%); linear neural networks (not all eyes were classified due to the doubt option) 78% (95% CI: 56–100%) and three layer perceptron network (also with doubt option): 88% (95% CI: 47–100%). The regression network showed a correlation of 0.63 (95% CI: 0.49–0.76).

*Corresponding author. Tel.: +43-662-4482-58446. Fax: +43-662-4482-3703. E-mail: w.hitzl@lks.at
†Formerly at the County Clinic for Ophthalmology and Optometry, St. Johanns-Spital, Landeskliniken Salzburg, now in private practice in Salzburg, Austria.
‡Director of the Eye Clinic of the County Hospital Salzburg, St. Johanns-Spital, Landeskliniken Salzburg, Salzburg, Austria.
CONCLUSION

This study yielded a negative result as to the initial exam, since in spite of different approaches none of the eight considered, quite elaborate models showed a considerably better performance than the simple predictor. Since the mean of the “Number” did not change considerably and the correlation of the “Number” (initial exam vs findings at 1 year) was moderate at best, we suggest to extend the prediction periods to 2 or even 5 years. During a longer prediction period, more changes of the “Number” may occur and further attempts can be made to find prediction models that can serve as an early warning system for the clinician.

KEYWORDS: Glaucoma damage prediction; Number; Neural network; Perceptron; Nerve fiber layer analysis

INTRODUCTION

Glaucoma is a leading cause of blindness in the world (Quigley, 1996). The disease is characterized by optic nerve head and nerve fiber layer alterations combined with the corresponding visual field defects. The rate of progression to blindness from treated glaucoma is generally not rapid (Sommer et al., 1991), therefore effective glaucoma screening could reduce blindness by identifying cases of undiagnosed glaucoma and eyes at risk for developing significant glaucoma and allowing earlier treatment.

In the past several years, non-invasive computerized analysis of the optic nerve head topography and optic nerve fiber layer measurements became more and more popular. Those instruments provide more objective assessments of the optic nerve head or the nerve fiber layers and were shown to be useful for glaucoma screening (Caprioli and Miller, 1989; Weinreb et al., 1995; 1998; Anton et al., 1997; Yamada et al., 2000). In order to analyze the optic nerve head topography, the Topographic Scanning System (TopSS®, Laser Diagnostic Technologies, Inc., San Diego, CA, USA) is currently in use in a large number of glaucoma clinics and provides more objective, quantitative and three-dimensional analysis of the optic nerve head (Ahn and Kee, 2000). This confocal imaging technique uses the principle of pinhole focusing for both the incoming laser beam and the beam that returns to the imaging detector. Then, only light reflected at or near the adjusted focal plane is detected and contributes to the image. The reproducibility of this method was found to be quite high, at least for some of the parameters studied (Geyer et al., 1998).

In order to assess the characteristics of the optic nerve fiber layer in more detail, the Glaucoma Detection System (GDx®, Laser Diagnostic Technologies, Inc., San Diego, CA, USA) is used. This is a scanning laser polarimeter for quantitative measurements of the peripapillary retinal nerve fiber layer thickness, which allows more objective, and reproducible measurements. The details of this instrument and imaging technique have been described elsewhere (Colen et al., 2000). In brief, the GDx® is a confocal scanning laser ophthalmoscope that illuminates the peripapillary retina with a 780-nm laser. An ellipsometer detects the reflected light from the retina and calculates the change in polarization (known as “retardation”), resulting from the birefringence of the retinal nerve fiber layer. The reproducibility was found—apart from some parameters—to be quite high (Colen et al., 2000).

Although the GDx® Nerve Fiber Analyzer provides a large number of parameters, it is still unclear at present, if there exists a subset of parameters and if so, which subset is of most importance for the initial diagnosis and assessment of the prognosis of glaucomatous damage. Currently the “Number”—a number ranging between 0 and 100—generated by a neural network that is implemented in the GDx® instrument — is postulated to be most relevant and useful in a clinical setting. In the daily practice of screening for glaucoma, the “Number” is considered as “normal” with values below 31, “border-line” between 31 and 70, and “outside normal limits” above 70. Two typical printouts are depicted in Fig. 1 to illustrate the “Number” of 2 patients. A patient with no risk: female, aged 70 years, the “Number” is found to be “7” in the right eye and “8” in the left eye (Fig. 1a). A patient at high risk: male, aged 80 years, the “Number” is “97” in the right eye and “99” in the left eye (Fig. 1b).

Whereas these instruments and the calculated parameters are considered to be useful in diagnosis and especially in follow-up of glaucoma, few data however, have been published about the potential “predictive power” of these instruments regarding future changes in eyes at risk. In daily clinical practice the glaucoma specialist is routinely confronted with the need to judge the probability of significant progression in individual patients. This study addresses the question, “To what extent can data from both instruments (TopSS® and GDx®), in combination with data gathered from standard ophthalmological examinations and thorough statistical analysis assist in predicting a critical change of the ‘Number’ at an early point in time?” so as to alarm the glaucoma specialist to potential damage with clinical significance.

METHODS

In a group of subjects sampled in the Salzburg-Moorfields-Collaborative-Glaucoma-Study (SMCGS), a total of 336 patients with an increased risk of having glaucoma were screened both at the beginning of the study and 1 year thereafter (for details of the study see Mistlberger et al., 1998). An eye was regarded as having an increased risk of
glaucoma, if the intraocular pressure (either in miosis or mydriasis) was greater than 20 mm Hg or the subjective cup to disc (C/D) ratio was greater than or equal to 0.45.

In accordance with the design of the SMCGS, the following inclusion criteria had to be fulfilled: age $\geq 40$ years, best corrected visual acuity $>20/30$, refraction within $\pm 6.00$ to $\pm 4.00$ diopters, difference of refraction between both eyes $<3.00$ diopters.

The exclusion criteria comprised the following conditions: pseudophakia, current antiglaucomatous therapy, other eye diseases potentially leading to visual field defects or secondary IOP elevation, contraindication against beta blockers, systemic corticosteroid therapy or pregnancy.

The parameters assessed at the initial investigation and after an 1 year follow-up period included a complete eye exam (family history (FH), biomicroscopy of the anterior segment, refraction, measurement of the best corrected visual acuity, applanation tonometry, fundus examination and assessment of the subjective C/D ratio). Those parameters were used to establish inclusion and exclusion criteria, but data of biomicroscopy, refraction, slit-lamp and fundus examinations were not used for the following statistical evaluations. Furthermore visual field data using the Humphrey Visual Field Test Program (Humphrey Instruments, Inc., San Leandro, CA, USA), such as the mean deviation (MD), the corrected pattern standard deviation (CPSD) and the glaucoma hemifield test were gathered (3 parameters for each eye). A description of the parameters of the optic nerve head parameters (TopSS®, Laser Diagnostic Technologies, Inc., San Diego, CA, USA) and those provided by the Glaucoma Detection System (GDX®), Laser Diagnostic Technologies, Inc., San Diego, CA, USA) is given in the appendix (7 optic nerve head parameters for each eye of the TopSS® instrument and 19 parameters for each eye of the GDX® instrument).

STATISTICAL METHODS

Description of the Sample

Casewise Deletion of Missing Data

All patients with missing data in at least one of the variables were excluded from the analysis. Univariate outliers: The range of each of the variables were checked and univariate outliers were removed from the data matrix.

Definition of Further Input Variables and the Output Variable

Definition of Categorical Parameters

An existing eye disease was defined as follows: (0) no eye disease, (1) existing glaucoma and (2) other eye disease. Family history was coded as follows: (0) there is no relative with an eye disease, glaucoma is present only in the patient’s (1) 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. Glaucoma hemifield test: visual field is normal (0), borderline (1) and abnormal (2).

Inclusion of quadratic terms: Quadratic terms were defined of all continuous variables.

A summary of all variables used for prediction is given the following list. The number in brackets specify the number of defined parameters: gender (1), age (2), eye disease (1), family history (1), visual acuity (2×2, i.e. right and left eye and quadratic terms), intraocular eye pressure in miosis (2×2), subjective C/D ratio (2×2), mean deviation (2×2), corrected standard deviation (2×2), glaucoma hemifield test (2), TopSS® parameters (2×7×2) and parameters of the glaucoma hemifield test (2×19×2). In total, 131 variables were defined.

Definition of the output variable: Different attempts were made to define a risk for glaucoma based on the “Number”. The prediction problem was considered as a regression problem (RP) that understood the “Number” as a continuous variable, but also as a classification problem (CP) that categorized the “Number” into two classes (no risk/patient at risk). In accordance with a suggestion of Laser Diagnostic Technologies, Inc., given in the manual of the GDx® instrument, values ranging from 0 to 30 were defined as normal and values ranging from 31 to 100 were defined as patients at risk.

Description of the Modeling Process

(A) Approach with the Simplest Predictor (CP)

Only the “Number” as the simplest predictor at the initial investigation was used to learn more about the corresponding prediction power.

(B) Linear Discriminant Analysis, LDA (CP)

(B.1) A linear discriminant model (LDM) with all parameters as input and (B.2) a LDM with a forward stepwise variable selection algorithm was computed. The stepwise procedure was guided by the respective F to enter values (5% quantile of the corresponding F-distribution) and F to remove values (10% quantile) (CP). Attention was given to whether the ratio of the number of variables to the number of patients was adequate (about 20 or more).

(C) Classification Tree Analyses (CP)

In order to use a model with a hierarchical nature, classification tree analyses were done. Prior probabilities estimated from data were used, i.e. these probabilities specify how likely it is, without using any prior knowledge of the parameters in the model, that a subject will fall into one of two classes. Misclassification (Breiman et al., 1984) costs were taken to be two times higher for false negative than for false positive subjects in order to increase the sensitivity. 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. In order to compare these models with all other models, additional analyses were done with misclassification costs that were set equal to 1.

(D) Neural Networks

In a fourth step, (D.1) regression networks (RP), (D.2) linear networks (CP) and (D.3) three layer perceptron networks (CP) with different network architectures were applied and tested (Bishop, 2000). Pre-processing involved conversion of nominal values and scaling of numeric values. This produced a numeric vector with one entry for each network input. The input values were then fed into the neural network. Feature selection was used to identify input variables that do not contribute significantly to the performance of the network by applying STATISTICA’s internal Intelligent Problem Solver (StatSoft Inc., 1999) using a variety of search, regularization and sensitivity based techniques (forward selection method and genetic algorithm). Training algorithm: Conjugate gradient descent was used for regression, linear and multilayer perceptron networks. The sum-squared error function was used in training the networks and in reporting the errors. Doubt option: In a first step, the accept and reject thresholds were set automatically in the approaches (D.2) and (D.3) such that the misclassification rate was minimized. If the activation was above the accept threshold, the case was deemed to belong to the risk class; if it was below the reject threshold, the case was deemed to belong to the class without risk, and if it was in between, the prediction is deemed to be “unknown”. In a 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 of the classified patients.

Test of Performances of the Approaches

Data Splitting into Training, Verification and Test Sample

In general, it was attempted to find models with the ability to make accurate predictions when faced with data not drawn from the original training set (generalization). To do so, the original data set was split into three subsets in the ratio of 2:1:1 for the training, verification and test sample based on random numbers generated by the computer. Different approaches used different samples to check that a model is generalizing properly by observing whether the error is reasonably low. (A) Used the total sample, since
the simple predictor can be regarded as defined a priori, the approaches (B) and (C) merged the training and verification sample as learning sample and were tested in the independent test sample. The approaches (D) used all three samples. The verification set was used to track the neural network’s error performance, to identify the best network, and to stop training if over-learning occurs. 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. The verification subset was used to keep an independent check on the performance of the networks during training, with deterioration in the verification errors indicating over-learning. A more detailed description is given in StatSoft Inc. (1999) and Bishop (2000).

In general, Pearson–Clopper confidence intervals (Hartung et al., 1993) were computed for the sensitivity and specificity rates based on the test sample. The performance of approach (A) was assessed based on the whole sample. For the multiple correlation coefficient of the RP, the corresponding confidence interval based on the test sample was computed.

**Further Statistical Analyses**

A dependent two-sided t-test together with the corresponding a posteriori power analysis was applied for the comparison of the “Number” at the initial and 1-year follow-up. A p-value smaller than 5% was considered as statistically significant. The corresponding empirical correlation and contingency coefficients were computed and the comparison of the categorized “Number” is illustrated in a cross tabulation table. The discriminant analyses were done with SPSS 10.0 (SPSS Inc., 1999), the test of the discriminant functions in the test sample was done with MATHEMATICA 3.0.1 (Wolfram Research, 1999). The classification tree analyses and power analysis were done with STATISTICA 5.5 (Statsoft Inc., 1999) and STATISTICA Neural Networks module 4.0 (Statsoft, Inc., 1999) was used to build the neural networks. Due to the large amount of evaluations, it was attempted to find models for the right eyes only.

**RESULTS**

After casewise data deletion, 42 patients had to be removed due to missing data, false data entries and univariate outliers. Therefore, a sample of n = 294 patients could be enrolled into the study. The sample was split into a training sample: 147 patients, a verification sample: 73 patients and a test sample: 74 patients. Demographic data are shown in Table I, the comparison of the “Number” at the initial investigation and at the 1-year follow-up did not show significant differences of the means ($p = 0.86$). The empirical distribution of risk at the initial and 1-year follow-up is given in Table II. The empirical contingency coefficient is medium/low with 0.45, the corresponding correlation coefficient is 0.70 (95% confidence interval: 0.63–0.76). A scatter plot illustrates the relation of the “Number” between the initial and 1-year investigation (Fig. 2). An overview of the performances of the models is given in Table III. A summary of the settings in the perceptron model is given in Table IV.
estimated sensitivity and specificity rates, percentages of unclassified patients and patients without risk and with risk of the three layer perceptron are given in Table V.

**DISCUSSION**

**Relationships between the “Number” at the Initial Investigation and the 1-year Follow-up**

The comparison could not detect a statistically significant difference of the means of the “Number”. An *a posteriori* power analysis for the dependent *t*-test showed—assuming a given mean of the “Number” at the initial investigation of 36, a standard deviation of 21, a correlation of 0.7, a sample size of *n* = 294, a type I error of 5% and a power of 90%—that an assumed true difference of 3 points or more is likely to be proven as statistically significant with the dependent *t*-test. So, even if the sample size were increased, it is very unlikely that the true, but unknown difference would be more than 3 points. Such a small difference is regarded as clinically irrelevant. However, there are some patients, however, with a marked deterioration of the “Number”, e.g. at the initial investigation with a value of “13” and with “71” at the 1-year follow-up. In order to exclude any mistakes in the recording of data, the case histories of those patients were checked individually. No mistakes during medical examination, no violation of the inclusion and exclusion criteria or wrong data entry could be found, however. Therefore, it can be concluded that such a severe deterioration can indeed occur within the time frame of 1 year and that it would be important to detect and/or predict these cases at an early time point.

**Performance and Applicability of the Different Approaches**

**(A) The Approach with the Simplest Predictor**

Since no model had to be constructed, the total sample could be used to compute the confidence intervals. Hence, the intervals were smaller than those based on the test samples of the other models. This approach that is used intuitively in daily practice can be regarded as giving good to moderate results (Table IV).

**(B.1) LDA**

The ratio of the number of eyes to the number of variables is 294/128 = 2.3. However, subsequent Monte Carlo research (Barcikowski and Stevens, 1975; Huberty, 1975) has shown that the discriminant function coefficients run in danger to be unstable, unless the sample size is fairly large (e.g. if there are 20 times more cases than there are variables). Considering the results of these earlier Monte-Carlo studies, we decided not to present results for this approach. *(B.2) A stepwise forward variable selection:* The confidence intervals in the test sample showed similar results to (A). Since no considerably better performance could be detected, we do not recommend to use such a model either.

**(C) Classification Tree Analyses**

These models showed good to moderate performances in the training sample. However, to adequately judge the performance of the classification tree models, it has to be stated that—due to software limitations—only 29 predictors could be used for the analyses and this could possibly reduce the performance of the classification trees. The additional analyses with equal misclassification costs
<table>
<thead>
<tr>
<th>Type of model</th>
<th>Approach with simple predictor</th>
<th>LDA</th>
<th>Classification tree analyses</th>
<th>Neural networks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(A) Pure “Number”</td>
<td>(B.1) Standard LDA</td>
<td>(B.2) LDA + stepwise forward algorithm</td>
<td>(C.1)</td>
</tr>
<tr>
<td>Variable selection</td>
<td>No</td>
<td>No</td>
<td>Stepwise forward algorithm</td>
<td>SL1 and SP1*</td>
</tr>
<tr>
<td>Number of variables</td>
<td>1</td>
<td>128</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>Test sample</td>
<td>Yes 69–83%†</td>
<td>No</td>
<td>61–79%§</td>
<td>68–84%§</td>
</tr>
<tr>
<td>Specificity</td>
<td>–</td>
<td>–</td>
<td>73–88%†</td>
<td>75–90%§</td>
</tr>
<tr>
<td>Interpretation possible</td>
<td>Yes</td>
<td>–</td>
<td>Yes, but more complicated than simple predictor</td>
<td>Yes, comes closest to “natural way of thinking”</td>
</tr>
<tr>
<td>Performance</td>
<td>Good/moderate</td>
<td>–</td>
<td>Good/moderate</td>
<td>Good/moderate</td>
</tr>
<tr>
<td>Applicable</td>
<td>Intuitive approach in daily practice</td>
<td>Not to be recommended for theoretical reasons</td>
<td>Not to be recommended, since not considerably better than (A)</td>
<td>Not to be recommended, since not considerably better than (A)</td>
</tr>
</tbody>
</table>

* SL1, Discriminant based univariate splits and SL2, CART style for exhaustive search for univariate splits (Gini measure for goodness of fit). Different stopping rules were applied. SP1, Pruning on misclassification error (1 SE rule) and SP2, FACT-style direct stopping (fraction of objects 10%).
† Total sample.
§ Learning sample.
‡ Independent test sample.
§ 95% confidence for the models with equal misclassification rates (set to 1 for both cases) are: sensitivity (C.1) 46–78%, (C.2) 45–76%, (C.3) 42–77% and (C.4) 29–62%, specificity (C.1) 57–83%, (C.2) 48–82%, (C.3) 58–89% and (C.4) 46–78%.
showed similar results than those with unequal misclassification costs. In summary, the classification rates were similar to (A) and again, we do not recommend to use these models.

**D) Neural Networks**

**(D.1) The regression network** (RP) showed moderate performance, since the observed multiple correlation coefficient is too low in the training sample. An attempt was therefore made to increase the sensitivity and specificity rates by transforming the RP into a CP and by using the doubt option of classification networks (D.2) The linear network showed also a similar performance to the previous models, so an implementation of such a model does not seem to be useful. (D.3) Three layer perceptron network: The sensitivity of this model in the test sample ranges from 47 to 100%. Based on the confidence intervals, the true, but unknown sensitivity rates in the approach (A) and (D.3) can be regarded as being equal. The confidence intervals suggests that the corresponding specificity rate is greater in (D.3) than in (A), since both confidence intervals do not overlap. The confidence interval ranges from 81 to 100%. However, we do not recommend to use the neural network model for two reasons: First, because the sensitivity rates can be assumed to be equal to the ones of (A) and secondly, because the upper limit of the confidence interval of (A) is not considerably greater than the lower limit of the confidence interval in (D.3).

The use of a doubt option makes it difficult to compare results. A dependent test for probabilities, e.g. the McNemar test can not be applied, since the sample size in the test sample of (D.3) is a subset of test sample of (A).

In order to cope with this problem, 95% confidence intervals were used.

**General Remarks to the Process of Modeling**

It is of major importance to emphasize that the models were proven in independent test samples. This is a particularly important point to consider and unfortunately, this crucial point is often overlooked in this and related fields of glaucoma research (Tomlinson and French, 1975; Tomlinson et al., 1975; Hart et al., 1979; Garway-Heath and Hitchings, 1998; Garway-Heath et al., 1998; Wollstein et al., 1998; Haper and Reeves, 2000; Bowd et al., 2002). As theoretically well understood, pointed out by mathematical statisticians and also observed in some of our own analyses (classification tree analyses, linear neural networks, regression networks), it is comparatively easy to find a model that performs well in the training sample and ends up with high classification rates between predicted and observed variables. However, good performance of a model in the learning sample does not necessarily allow a generalization to new patients, unless the performance of a model is tested independently. Otherwise, such a model runs in danger loosing its predictive power for new and independent observations. This could mean that a promising model with sufficiently high classification rates in the learning sample—that is used by the glaucoma expert in clinical routine—fails to classify patients, i.e. observations that are not in the learning sample. The presented modeling process should add weight to our

<table>
<thead>
<tr>
<th>TABLE IV</th>
<th>Overview of the three layer perceptron network model (D.3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event to be predicted</td>
<td>“Minimal” risk for glaucoma</td>
</tr>
<tr>
<td>Missing data (removed)</td>
<td>42</td>
</tr>
<tr>
<td>Sample sizes of subsamples</td>
<td>(147, 73, 74)</td>
</tr>
<tr>
<td>Number networks tested</td>
<td>52</td>
</tr>
<tr>
<td>Type of network</td>
<td>Three layer perceptron</td>
</tr>
<tr>
<td>Feature selection</td>
<td>Genetic algorithm</td>
</tr>
<tr>
<td>Type of training algorithm used</td>
<td>Conjugate gradient descent</td>
</tr>
<tr>
<td>Maximal number of epochs</td>
<td>500</td>
</tr>
<tr>
<td>Cross verification in verification sample</td>
<td>Yes</td>
</tr>
<tr>
<td>Test in independent test sample</td>
<td>Yes</td>
</tr>
<tr>
<td>Number of input neurons</td>
<td>116</td>
</tr>
<tr>
<td>Number of hidden neurons</td>
<td>26</td>
</tr>
<tr>
<td>Error function</td>
<td>Sum-squared</td>
</tr>
<tr>
<td>Layer 1: PSP-function*</td>
<td>Linear</td>
</tr>
<tr>
<td>Activation function</td>
<td>Linear</td>
</tr>
<tr>
<td>Layer 2: PSP-function</td>
<td>Linear</td>
</tr>
<tr>
<td>Activation function</td>
<td>Logistic</td>
</tr>
<tr>
<td>Layer 3: PSP-function</td>
<td>Linear</td>
</tr>
<tr>
<td>Activation function</td>
<td>Logistic</td>
</tr>
<tr>
<td>Threshold for acceptance</td>
<td>0.75</td>
</tr>
<tr>
<td>Threshold for rejectance</td>
<td>0.23</td>
</tr>
</tbody>
</table>

* Post synaptic potential function.

<table>
<thead>
<tr>
<th>TABLE V</th>
<th>Classification statistics in the training, verification and test sample of the three-layer perceptron network (D.3). Patients classified at the initial investigation (rows) vs. patients observed at 1-year follow-up (columns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right eyes</td>
<td>Training sample</td>
</tr>
<tr>
<td></td>
<td>No risk</td>
</tr>
<tr>
<td>Total</td>
<td>73</td>
</tr>
<tr>
<td>Correct</td>
<td>26</td>
</tr>
<tr>
<td>Wrong</td>
<td>1</td>
</tr>
<tr>
<td>“unknown”</td>
<td>46</td>
</tr>
<tr>
<td>“without risk”</td>
<td>26</td>
</tr>
<tr>
<td>“at risk”</td>
<td>1</td>
</tr>
<tr>
<td>Observed sensitivity rate</td>
<td>–</td>
</tr>
<tr>
<td>Observed specificity rate</td>
<td>96%</td>
</tr>
<tr>
<td>% of unknown</td>
<td>63%</td>
</tr>
</tbody>
</table>

Note: Model has doubt option, subjects with “unknown” classification were not predicted, but are clearly specified by the perceptron model.
One reason for this could be the fact that not all output variables of the GDx® scan were recorded in the data bank and therefore, important GDx® input variables could be missing. Therefore, even with a perfect correlation of one of the “Number” at the initial and 1-year follow-up, a prediction of the “Number” would not be possible. However, this can be ruled out, since a successful computation of the unknown “Number” at the initial investigation based on the GDx® parameters recorded in our data bank (with exclusion of the “Number”) at the initial investigation could be done. We found a three-layer perceptron network with excellent performance for this RP (multiple correlation coefficients of (0.99, 0.99, 0.98)). This suggests that some of the important information is lost during the prediction period of 1 year.

- It may be partially attributed to the intra- and inter-observer reliability of the measurement, but also
- that the change of “Number” is caused by factors that are clinically not observed or observable and therefore, are not recorded.

**CONCLUSION AND FURTHER DEVELOPMENTS**

This study suggests a negative result. None of the eight models showed a considerably better performance than the simple predictor. Therefore, no prediction model is currently available for an accurate prediction of the risk based on the “Number”, although the early recognition of a deterioration of the “Number” is one of the major challenges in the struggle for the prevention of glaucomatous damage.

What other options can be tested to achieve a better prediction? A model with additional information might show better performances, e.g. information about the initial and the 1-year examinations for the prediction of the “Number” at the 2-year or later examinations or the inclusion of relevant ocular parameters, such as data about the intraocular blood flow. Another possibility is to extend the prediction period to 2 or even 5 years, since the results show that—on average—not more than 3 points in the “Number” had changed within 1 year. A follow-up of, e.g. 5 years would possibly reveal significant changes of the “Number” and the corresponding deterioration of the glaucoma status of a patient. A further challenge is the quest for models for an accurate prediction of the loss of other nerve fiber layer parameters, such as the Ellipse Modulation, one of the polarimetric data or mean sector values. We also suggest in future to consider survival analysis particularly over the a period of 5 years or longer to study the questions at which time and how many eyes develop a risk for glaucoma (based on the “Number”) in our screening program. Such models could offer additional information for the glaucoma specialist for identifying subjects with high risks of significant damage and alert him to take adequate measures against progression of this disease in a timely manner.

**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 like to thank Dr David Mathews. He kindly gave us editorial assistance with this manuscript. 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**


APPENDIX

TopSS® Parameters (7 Variables)
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 Volume Above is the volume of all tissue or structures within the Neuroretinal Rim 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 for all parameters are mm, mm², mm³ or unit less for distances, areas, volumes or ratios, respectively.

G Dx® Parameters (19 Variables)
Symmetry: This is the ratio of the average of the 1500 thickest pixels in the superior quadrant over the average of the 1500 thickest pixels in the inferior quadrant. The closer the ratio is to 1.0, the more symmetric the nerve fiber layer of the superior and inferior quadrants; Superior Ratio: This is the ratio of the average of the 1500 thickest pixels in the superior quadrant over the average of the 1500 median pixels in the temporal quadrant. Inferior Ratio: This is the ratio of the average of the 1500 thickest pixels in the inferior quadrant over the average of the 1500 median pixels in the temporal quadrant. Superior/Nasal: This is the ratio of the average of the 1500 thickest pixels in the superior quadrant over the average of the 1500 median pixels in the nasal quadrant. Max Modulation: Provides an indication of the difference between the thickest parts of the nerve fiber layer and the thinnest parts. In a normal eye, where the superior and inferior nerve fiber layer is much thicker than the nerve fiber layer of the nasal or temporal quadrants, the modulation number will usually be greater than 1.0. Ellipse Modulation: is an indication of the difference between the thickest parts of the nerve fiber layer and the thinnest parts, similar to the Max Modulation. The difference is that, rather than using all of the points in the image, Ellipse Modulation uses the pixels covered by the ellipse surrounding the optic nerve. The “Number”: an experimental number based on a neural network between 0 and 100, estimating the potential risk of having glaucoma, with the computation automatically performed by the GDx® instrument. The manufacturer states in the manual, “The higher the number, the higher the probability that the patient has glaucoma”. In daily practice of glaucoma screening, the “Number” is considered as normal with values ranging from 0 to 30, borderline with values between 31 and 70 and outside normal limits with values above 70. Average Thickness: The average thickness of all pixels in the image. Ellipse Average or Total Polar Average: The average thickness of the nerve fiber layer beneath the ellipse surrounding the optic nerve. Superior Average: The average thickness of the nerve fiber layer beneath the portion of the ellipse surrounding the optic nerve in the superior quadrant. Temporal Average: The average thickness of the nerve fiber layer beneath the portion of the ellipse surrounding the optic nerve in the temporal quadrant. Inferior Average: The average thickness of the nerve fiber layer beneath the portion of the ellipse surrounding the optic nerve in the inferior quadrant. Nasal Average: The average thickness of the nerve fiber layer beneath the portion of the ellipse surrounding the optic nerve in the nasal quadrant. Total Polar Integral: The total area under the curve of the nerve fiber layer beneath the entire ellipse surrounding the optic nerve. Temporal Integral: The total area under the curve of the nerve fiber layer beneath the temporal portion of the ellipse surrounding the optic nerve. Inferior Integral: The total area under the curve of the nerve fiber layer beneath the inferior portion of the ellipse surrounding the optic nerve. Nasal Integral: The total area under the curve of the nerve fiber layer beneath the nasal portion of the ellipse surrounding the optic nerve.
Superior Integral: The total area (in mm²) under the curve (or total volume) of the nerve fiber layer along the superior portion of the ellipse surrounding the optical nerve head.