A Statistical Approach to the Evaluation of Covariate Effects on the Receiver Operating Characteristic Curves of Diagnostic Tests in Glaucoma

Felipe A. Medeiros, Pamela A. Sample, Linda M. Zangwill, Jeffrey M. Liebmann, Christopher A. Girkin, and Robert N. Weinreb

PURPOSE. To describe an approach for the evaluation of covariate effects on receiver operating characteristic (ROC) curves and to apply this methodology to the investigation of the effects of disease severity and age on the diagnostic performance of frequency doubling technology (FDT) and standard automated perimetry (SAP) visual function tests for glaucoma detection.

METHODS. The study included 370 eyes of 211 participants, with 174 eyes of 110 patients having glaucomatous optic neuropathy and 196 eyes of 101 subjects being normal. All patients underwent visual function testing with FDT 24-2 Humphrey Matrix and SAP SITA (Carl Zeiss Meditec, Inc., Dublin, CA). Disease severity was evaluated by the amount of neuroretinal rim loss assessed by confocal scanning laser ophthalmoscopy. An ROC regression model was fitted to evaluate the influence of disease severity and age on the diagnostic performance of the pattern SD (PSD) index from FDT 24-2 and SAP SITA.

RESULTS. After adjustment for age, the areas under the ROC curves (AUCs) for SAP SITA PSD for 10%, 30%, 50%, and 70% loss of neuroretinal rim area were 0.638, 0.756, 0.852, and 0.920, respectively. Corresponding values for FDT 24-2 PSD were 0.766, 0.857, 0.922, and 0.962. For 10% and 30% rim loss, FDT 24-2 PSD had a significantly larger AUC than did SAP SITA PSD.

CONCLUSIONS. A regression methodology to evaluate covariate effects on ROC curves can be useful for assessment of diagnostic tests in glaucoma. Using the proposed methodology, a significantly better performance of FDT 24-2 compared to SAP SITA for diagnosis of early glaucoma was demonstrated. (Invest Ophthalmol Vis Sci. 2006;47:2520–2527) DOI:10.1167/iovs.05-1441

From the Hamilton Glaucoma Center and Department of Ophthalmology, University of California, San Diego, California; the New York Eye and Ear Infirmary, New York, New York; and the University of Alabama at Birmingham, Birmingham, Alabama.

Supported in part by the National Eye Institute Grants EY11008 (LMZ) and EY08208 (PAS). Participant retention incentive grants were provided in the form of glaucoma medications at no cost: Alcon Laboratories, Inc., Allergan, Inc., Pfizer, Inc., and Santen, Inc. (PAS).

Submitted for publication November 8, 2005; revised December 15, 2005, and January 17, 2006; accepted March 14, 2006.

Disclosure: F.A. Medeiros, Carl Zeiss Meditec, Inc. (F); P.A. Sample, Carl Zeiss Meditec, Inc., Alcon Laboratories, Inc., Allergan, Inc., Pfizer, Inc. (F); L.M. Zangwill, Heidelberg Engineering (F, R); J.M. Liebmann, None; C.A. Girkin, None; R.N. Weinreb, Carl Zeiss Meditec, Inc. and Heidelberg Engineering (F)

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Corresponding author: Felipe A. Medeiros, Hamilton Glaucoma Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92039-0946; fmedeiros@eyecenter.ucsd.edu.

Receiver operating characteristic (ROC) curves are a well-accepted measure of accuracy of diagnostic tests that yield continuous or ordinal results. Based on the notion of using a threshold to classify subjects as positive (diseased) or negative (nondiseased), an ROC curve is a plot of the true-positive rate (TPR) versus the false-positive rate (FPR) for all possible cutpoints. Thus, it describes the whole range of possible operating characteristics for the test and hence its inherent capacity for distinguishing between subjects with and those without glaucoma. In most studies, however, ROC curves for diagnostic tests have been reported without taking into account the possible effects of covariates on test results. For example, in most studies of diagnostic tests in glaucoma, a single ROC curve is reported to represent the performance of the test in all included patients. Patients in these studies, however, frequently have different degrees of disease severity or different values of other covariates, such as age. Although the single “pooled” ROC gives the average performance of the test in the population, we are frequently interested in knowing how the test performs in subgroups of patients—for example, patients with early disease or with a specific value on another covariate.

Diagnostic tests tend to be more sensitive in advanced stages of the disease, and measures of diagnostic accuracy obtained from studies that include only patients with moderate or severe disease may not be applicable to patients with early disease or to those suspected of having the disease. Further, it is possible that the comparison of the diagnostic abilities of different tests are influenced by the severity of glaucomatous damage. For example, it is possible that a particular test is more sensitive at early stages of the disease, whereas another test may be more sensitive at moderate or advanced stages. Therefore, it is important to characterize the relationship between the performance of the diagnostic test and the severity of disease and to evaluate how this relationship affects the comparison between different tests. Regression methods have been proposed to analyze covariate effects on the ROC curves. These methods allow the evaluation of the influence of covariates such as disease severity on the diagnostic performance of the test, so that ROC curves for specific values of the covariates can be obtained. Another advantage of this method is that it allows comparison of ROC curves for different tests after adjusting for the effects of covariates, so that the tests can be better compared.

In the present study, we used ROC regression methodology to analyze the effects of two covariates, disease severity and age, on the ROC curves representing the diagnostic performance of two visual function tests for glaucoma, standard automated perimetry using the Swedish Interactive Threshold Algorithm (SAP SITA) and frequency-doubling technology (FDT) perimetry. Based on its selective visual function testing properties, FDT has been proposed as a test for early detection of glaucomatous functional damage. This test has recently undergone major modifications in its procedures, with the development of the 24-2 test pattern, but the ability of the FDT 24-2 test to diagnose different levels of glaucomatous damage in patients has not yet been determined. The proposed
methodology allowed the comparison of the diagnostic performance of these two visual function tests adjusting for the severity of disease and age, so that their accuracies could be compared at specific levels of these covariates.

METHODS

This was an observational case-control study. Participants in the study were included in a prospective longitudinal study designed to evaluate optic nerve structure and visual function in glaucoma (Diagnostic Innovations in Glaucoma Study [DIGS]) conducted at the Hamilton Glaucoma Center (University of California, San Diego), with additional participants supplied by the University of Alabama at Birmingham and the New York Eye and Ear Infirmary. Participants at all sites were longitudinally evaluated according to a pre-established protocol that included regular follow-up visits, during which they underwent clinical examination and several imaging and function tests. For the present study, participants were observed for at least 6 months and performed at least two visual field tests of each type. All participants who met the inclusion criteria were enrolled in the present study. Informed consent was obtained from all participants. The protocols were approved by the designated Human Subjects Committee from each center, and the methods described adhered to the tenets of the Declaration of Helsinki. Data were entered into a computer database.

Each subject underwent a comprehensive ophthalmic examination, including review of medical history, best corrected visual acuity, slit lamp biomicroscopy, intraocular pressure (IOP) measurement with Goldmann applanation tonometry, gonioscopy, dilated funduscopy examination with a 78-D lens and stereoscopic optic disc photography. To be included, subjects had to have best corrected visual acuity of 20/40 or better, spherical refraction within ± 3.0 D, and cylinder correction < 3.0 D, and open angles on gonioscopy. Eyes with coexisting retinal disease, uveitis, or nonglaucomatous optic neuropathy were also excluded from the investigation.

Participants were classified according to the presence of structural damage to the optic disc as assessed by simultaneous stereoscopic optic disc photographs. Visual field results were not used to classify patients. The photographs were evaluated by two experienced graders, and each was masked to the subject’s identity and to the other test results. For inclusion, a photograph had to be deemed of adequate quality or better. Glaucomatous optic neuropathy (GON) was defined as the presence of neuroretinal rim thinning, excavation, notching, or characteristic retinal nerve fiber layer defects. Discrepancies between the two graders were either resolved by consensus or by adjudication of a third experienced grader.

Patients with glaucoma were required to have a diagnosis of GON in at least one eye, regardless of IOP. When both eyes of the same patient had GON, both eyes were included in the study; provided that they satisfied other inclusion criteria. When only one eye had a diagnosis of GON, that eye was included. Normal control subjects had IOPs of 22 mm Hg or less with no history of increased IOP and normal optic discs in both eyes. The photographs were evaluated by two experienced graders, and each was masked to the subject’s identity and to the other test results. For inclusion, a photograph had to be deemed of adequate quality or better. Glaucomatous optic neuropathy (GON) was defined as the presence of neuroretinal rim thinning, excavation, notching, or characteristic retinal nerve fiber layer defects. Discrepancies between the two graders were either resolved by consensus or by adjudication of a third experienced grader.

Patients with glaucoma were required to have a diagnosis of GON in at least one eye, regardless of IOP. When both eyes of the same patient had GON, both eyes were included in the study; provided that they satisfied other inclusion criteria. When only one eye had a diagnosis of GON, that eye was included. Normal control subjects had IOPs of 22 mm Hg or less with no history of increased IOP and normal optic discs in both eyes.

Two perimetric procedures were used to test visual function: SAP SITA and FDT, both using a 24-2 stimulus pattern. To minimize learning effects, the most recent field was used for all patients. All patients were required to have reliable results on visual field tests including fixation losses and false-positive and false-negatives rates not greater than 25%. For evaluation of disease severity, we used a visual-function-independent scale: the amount of neuroretinal rim loss in glaucomatous eyes as assessed by optic disc topographic measurements obtained by confocal scanning laser ophthalmoscopy (HRT II; Heidelberg Engineering, GmbH, Dossenheim, Germany). This parameter was chosen among other HRT parameters, as it showed the best correlation with histologic optic nerve fiber count in a previous experimental study in monkeys. Visual function tests and optic disc imaging with HRT II were all obtained within an interval not greater than 6 months.

Visual Function Testing Procedures

Standard achromatic automated perimetry was performed using a Humphrey Visual Field Analyzer II (Carl Zeiss Meditec, Inc., Dublin, CA), program 24-2 and SITA testing algorithm. It utilizes a small (0.47°) 200-ms flash of white light as the target presented on a dim background (10.5 cd/m²). FDT perimetry (FDT 24-2) was performed with the commercially available Humphrey Matrix perimeter (Carl-Zeiss Meditec, Inc.). The Humphrey Matrix presents 5° stimuli, with a spatial frequency of 0.5 cyc/deg and temporal frequency of 18 Hz, on a background with a luminance of 100 cd/m². Stimuli are presented for 500 ms, including ramped onsets and offsets of 100 ms. The principles and psychometric properties of the ZEST strategy used for threshold estimation have been described in detail elsewhere. The test locations of the FDT 24-2 program are similar to those of the SAP SITA 24-2 test. For both SAP SITA and FDT 24-2, 54 locations were tested within the central 24° of visual field. The two locations just above and below the blind spot were not included in the analysis.

The visual field global index PSD (pattern SD) was used for evaluation of diagnostic accuracy of the visual function tests. It was calculated for each test of each patient according to the following formula:

\[
PSD = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} S_i^2 - \frac{1}{n-1} \sum_{i=1}^{n} T_i - N_i - MD},
\]

where \(T_i\) is the measured age-adjusted threshold at point \(i\), \(N_i\) is the normal age-adjusted reference threshold (obtained from our control subjects) at point \(i\), \(S_i^2\) is the variance of normal field measurements at point \(i\), \(n\) is the number of points in the test \((n = 52\) for both FDT 24-2 and SITA standard), and MD is the mean deviation. PSD was selected based on its comparable or superior performance compared with other indexes in previous studies involving SAP and FDT visual function tests.

HRT II Confocal Scanning Laser Ophthalmoscope

The HRT II employs a diode laser (670-nm wavelength) to scan the retinal surface sequentially, in the x and y directions at multiple focal planes. According to confocal scanning principles, a three-dimensional topographic image is constructed from a series of optical image sections at consecutive focal planes. The topography image determined from the acquired three-dimensional image consists of 384 × 384 (147,456 total) pixels, each of which is a measurement of retinal height at its corresponding location. For each patient, three topographical images were obtained and were combined and automatically aligned to make a single mean topography used for analysis. Magnification errors were corrected using patients’ corneal curvature measurements. An experienced examiner outlined the optic disc margin on the mean topographic image while viewing stereoscopic photographs of the optic disc. Good-quality images required a focused reflectance image with a standard deviation not greater than 50 μm.

The HRT II software calculates several topographic parameters of the optic disc. In the present study, the percentage loss of neuroretinal rim area from the HRT II was used as a measure of disease severity in glaucomatous eyes. The percentage was calculated from the difference between the expected rim area adjusted for optic disc size and the observed rim area. For example, a 10% loss of neuroretinal rim area indicates that the difference between the expected and observed values of rim area was 10% of the expected neuroretinal rim area. For each eye, the expected rim area was obtained from the HRT II software based on a linear regression adjusting for the size of the optic disc. The parameters were obtained by a regression of the log rim area to the size of the optic disc in a group of normal subjects from the Moorfields Regression Analysis.

ROC Regression Methodology

The ROC regression methodology applied in the current study was originally proposed by Pepe et al. and previously used to evaluate
the influence of the degree of hearing loss on results of diagnostic tests in audiology, as well as in other applications. As this modeling approach has not been previously applied to evaluation of diagnostic tests in ophthalmology, we will describe it in some detail. Further detail can be found in several publications.

The ROC \( C_{X,X_2}(q) \) is the probability that a diseased individual with disease-specific covariates \( X_2 \) and common covariates \( X \) has test results exceeding the \( q \)th quantile of the distribution of test results from nondiseased individuals. That is, when the specificity of the test is \( 1 - q \), the sensitivity is \( ROC_{X,X}(q) \). An example of disease-specific covariate is severity of the disease, as this covariate is obviously not defined for healthy subjects. In contrast, age is an example of a common covariate, as it is defined for subjects without and those with disease. The effects of \( X \) and \( X_2 \) can be modeled on \( ROC_{X,X_2}(q) \) by a generalized linear regression model (ROC-GLM model).

The general ROC regression model can be represented by

\[
RO_{X,X_2}(q) = g \left( \sum_{i=1}^n \alpha_i h_i(q) + \beta X + \beta_0 X_0 \right)
\]

The ROC is a function of covariates common to diseased and subjects without disease, covariates specific to diseased subjects, and a function \( h_i(q) \) which defines the location and shape of the curve. This approach is referred to as parametric distribution-free, as it specifies a parametric model for the ROC curve but does not assume distributions for the test results, which makes it advantageous compared with other modeling procedures.\(^{26,27}\) The functions \( g() \) and \( h_i() \) are chosen so that the ROC curve is monotone, increasing on the unit square. In most applications, \( g() \) is \( \Phi \), the normal cumulative distribution function, \( h_1(q) = 1 \) (with coefficient \( \alpha_1 \)) and \( h_2(q) = \Phi^{-1}(q) \) (with coefficient \( \alpha_2 \)) are generally used, which results in the binormal ROC model

\[
RO_{C,X}(q) = \Phi(\alpha_1 + \alpha_2 \Phi^{-1}(q) + \beta X + \beta_0 X_0)
\]

where the coefficients \( \alpha_1 \) and \( \alpha_2 \) are the intercept and slope of the ROC curve, respectively. If the coefficient for a specific variable \( X_i(\beta) \) is greater than zero, then the discrimination between those with disease and those without increases with increasing values of this covariate. Similarly, if the coefficient for the disease-specific covariate \( X_{2i}(\beta_2) \) is greater than zero, then diseased subjects with larger values of this covariate are more distinct from nondiseased subjects than are diseased subjects with smaller values of \( X_2 \).

After the estimation of the parameters using generalized linear models, the area under the ROC curve can be obtained by:

\[
AUC = \int_0^1 RO_{C,X}(q)dq = \Phi \left( \frac{\alpha_1 + \beta X + \beta_0 X_0}{\sqrt{1 + \alpha_2^2}} \right)
\]

In the present study, an ROC-GLM model was fitted to assess the influence of the disease-specific covariate severity and the common covariate age on the diagnostic performance of FDT 24-2 and SAP SITA parameters as evaluated by ROC curves. The following ROC regression model was then fitted:

\[
RO_{C,X,X_2}(q) = \Phi(\alpha_1 + \alpha_2 \Phi^{-1}(q) + \beta_1 FDT + \beta_2 FDT \times \Phi^{-1}(q) + \beta_3 \text{severity} + \beta_4 \text{severity} \times \Phi^{-1}(q) + \beta_5 FDT \times \text{severity} + \beta_6 \text{age} + \beta_7 \text{age} \times \Phi^{-1}(q) + \beta_8 \text{FDT})
\]

where FDT is a binary variable indicating the type of test (SAP SITA was used as the reference test), severity is the variable indicating severity of glaucomatous damage as measured by percentage loss of rim area, and age is a variable indicating patient’s age. Interaction terms between the variables and \( \Phi^{-1}(q) \) were included to allow the effects of the covariates to differ by varying amounts depending on the FPR\(q\) (or specificity \( 1 - q \), that is, to influence the shape of the curve. Interaction terms between FDT and severity and between FDT and age were included to assess whether the influence of disease severity and age was similar or different between FDT 24-2 and SAP SITA tests.

Parameters were estimated using probit regression. To obtain confidence intervals for regression parameters, a bootstrap resampling procedure was used (\( n = 500 \) resamples).\(^{29}\) As measurements from both eyes of the same subject are likely to correlate, the use of standard statistical methods for parameter estimation can lead to underestimation of standard errors and to confidence intervals that are too narrow.\(^{30}\) Therefore, to account for the fact that both eyes of some subjects were used for analyses, the cluster of data for the study subject was considered as the unit of resampling when calculating standard errors. This procedure has been used in other studies to adjust for the presence of multiple correlated measurements from the same unit.\(^{27,28} \)

Statistical analyses were performed on computer (Stata ver. 9.0; StataCorp., College Station, TX; and SPSS ver. 13.0; SPSS Inc., Chicago, IL). The \( \alpha \) level (type I error) was set at 0.05.

**Results**

The study included 370 eyes of 211 participants. One hundred seventy-four eyes of 110 patients had a diagnosis of GON, whereas 196 eyes of 101 subjects were normal control eyes. Normal subjects were significantly younger than patients with glaucoma (mean ± SD; 50 ± 14 years vs. 67 ± 12 years, \( P < 0.001 \); Student’s \( t \)-test). The average neuroretinal rim loss in glaucomatous eyes was 22%. The distribution of severity of disease according to percentage loss of rim area in glaucomatous eyes is shown on Figure 1.

The estimates of the coefficients of the ROC regression models are shown in Table 1. To provide more meaningful values, the variables severity and age were centered on their approximate mean values in glaucomatous eyes (22% and 65 years, respectively) when estimating the coefficients. Therefore, for an average neuroretinal rim loss of 22% and 65 years of age, FDT 24-2 PSD performed better than SAP SITA PSD, as indicated by the significant coefficient attributed to the variable \( FDT \) in the regression model. The superior performance of FDT 24-2 was similar throughout the range of false-positive (i.e., 1 − specificity) values, as indicated by the nonsignificant coefficient associated with the interaction term \( FDT \times \Phi^{-1}(q) \) (\( P = 0.400 \)). That is, the ROC curves for FDT 24-2 and SAP SITA had a similar shape and did not cross.

The severity of disease, as measured by HRT II percentage loss of rim area, had a significant influence on the diagnostic performance of both SAP SITA and FDT 24-2, as indicated by the statistically significant value attributed to the coefficient representing severity (\( P < 0.001 \)). As expected, both tests performed better in patients with more severe disease. The influence of the severity of disease was not significantly different between the two tests, as indicated by the nonsignificant value of the coefficient representing the interaction between severity and \( FDT \) (\( \beta_5; P = 0.892 \)). There was a tendency for disease severity to exert a relatively greater, but not statistically significant, effect on lower FPRs (i.e., higher specificities), as indicated by the negative coefficient for the term \( \text{Severity} \times \Phi^{-1}(q) \) (\( P = 0.497 \)). Figure 2 shows ROC curves for SAP SITA and FDT 24-2 for arbitrarily chosen levels of percentage of neuroretinal rim loss and for age at 65 years, as calculated from the regression model. ROC curve areas and probabilities for the comparison between tests are shown on Table 2. For 10% and 30% neuroretinal rim loss, FDT 24-2 PSD had a significantly larger area under the ROC curve than did SAP SITA PSD. For 50% and 70% rim loss, although the area under the ROC curve
for FDT 24-2 was larger than for SAP SITA PSD, the difference was not statistically significant.

Age was also significantly associated with the test results. Both SAP SITA and FDT 24-2 performed better in older patients, as indicated by the positive and statistically significant coefficient associated with age ($P = 0.008$). The influence of age was similar between the two visual function tests ($P = 0.289$ for the interaction term age $\times$ FDT) and throughout the false positive (or 1 specificity) range ($P = 0.603$ for the interaction term age $\times$ $\Phi^{-1}(q)$).

Based on the results provided by the regression model, we calculated sensitivities at fixed specificities at 80% and 95% for FDT 24-2 and SAP SITA throughout the range of disease severity (Fig. 3). Sensitivities for FDT 24-2 were higher than those for SAP SITA.

**DISCUSSION**

In the present study, we demonstrated that a methodology of evaluating covariate effects on ROC curves can be useful to assess the performance of diagnostic tests for glaucoma detection. By incorporating disease severity and age as covariates on ROC regression models, we were able to demonstrate a significant effect of these two variables on the diagnostic performance of SAP SITA and FDT 24-2. More important, we were able to detect significant differences between these two tests under various matched levels of these covariates. The results can potentially have significant influence on the use of these tests in clinical practice.

Visual field testing with SAP is not selective for a particular ganglion cell type. Because there is a considerable overlap in the receptive fields of retinal ganglion cells and redundancy in the coverage of a given location in the retina, a nonselective test may not be sensitive for the earliest loss of retinal ganglion cells that occurs in glaucoma.$^{31}$ In fact, our findings demonstrated that SAP performed poorly for diagnosis of patients with early disease. For a 10% loss of rim area, the ROC curve area for SAP SITA was 0.638, with a sensitivity of only 21% for 95% specificity. With increasing disease severity, the performance of SAP SITA improved, with the area under the ROC

**Table 1. Results of the ROC Regression Model Incorporating Age, Disease Severity and Type of Visual Function Test as Covariates**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Coefficient</th>
<th>Estimate</th>
<th>95% CI</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>$\alpha_1$</td>
<td>0.691</td>
<td>(0.457–0.957)</td>
<td>$&lt;0.001$</td>
</tr>
<tr>
<td>$\Phi^{-1}(q)$</td>
<td>$\alpha_2$</td>
<td>0.759</td>
<td>(0.543–0.925)</td>
<td>$&lt;0.001$</td>
</tr>
<tr>
<td>FDT</td>
<td>$\beta_1$</td>
<td>0.531</td>
<td>(0.223–0.921)</td>
<td>0.002</td>
</tr>
<tr>
<td>FDT $\times$ $\Phi^{-1}(q)$</td>
<td>$\beta_2$</td>
<td>0.117</td>
<td>(–0.154–0.414)</td>
<td>0.400</td>
</tr>
<tr>
<td>Severity</td>
<td>$\beta_3$</td>
<td>2.070</td>
<td>(1.259–3.170)</td>
<td>$&lt;0.001$</td>
</tr>
<tr>
<td>Severity $\times$ $\Phi^{-1}(q)$</td>
<td>$\beta_4$</td>
<td>–0.168</td>
<td>(–0.649–0.325)</td>
<td>0.497</td>
</tr>
<tr>
<td>Severity $\times$ FDT</td>
<td>$\beta_5$</td>
<td>0.072</td>
<td>(–1.035–1.135)</td>
<td>0.892</td>
</tr>
<tr>
<td>Age</td>
<td>$\beta_6$</td>
<td>0.021</td>
<td>(0.008–0.041)</td>
<td>0.008</td>
</tr>
<tr>
<td>Age $\times$ $\Phi^{-1}(q)$</td>
<td>$\beta_7$</td>
<td>0.005</td>
<td>(–0.007–0.013)</td>
<td>0.603</td>
</tr>
<tr>
<td>Age $\times$ FDT</td>
<td>$\beta_8$</td>
<td>0.011</td>
<td>(0.002–0.020)</td>
<td>0.289</td>
</tr>
</tbody>
</table>

$q$, false-positive fraction (1 – specificity), $\Phi^{-1}$, inverse normal cumulative distribution function. The variables severity and age were centered on their mean values in glaucomatous subjects.
curve being as high as 0.920 for patients with more advanced damage (70% loss of neuroretinal rim area). In a histologic study in human eyes, Kerrigan-Baumrind et al. showed that an average loss of 27.3% of retinal ganglion cells is necessary for the corrected PSD index of standard achromatic perimetry to fall below the 95% normal confidence limits. Of interest, using SAP SITA PSD at 95% specificity in our study, the average percentage loss of rim area of the patients with glaucoma identified as abnormal was 30%, very close to the number provided by Kerrigan-Baumrind et al.

Our results demonstrated the superior performance of FDT perimetry for detection of early disease, as previously suggested by other investigators. Longitudinal studies have dem-

**FIGURE 2.** ROC curves for SAP SITA PSD and FDT 24-2 PSD for arbitrary values of percentage neuroretinal rim loss, according to the regression model.

**TABLE 2.** ROC AUCs According to the Regression Model for Arbitrary Values of Percentage Loss of Neuroretinal Rim Area

<table>
<thead>
<tr>
<th>Neuroretinal Rim Area Loss (%)</th>
<th>SAP SITA PSD</th>
<th>FDT 24-2 PSD</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.638</td>
<td>0.766</td>
<td>0.004</td>
</tr>
<tr>
<td>30</td>
<td>0.756</td>
<td>0.857</td>
<td>0.004</td>
</tr>
<tr>
<td>50</td>
<td>0.852</td>
<td>0.922</td>
<td>0.089</td>
</tr>
<tr>
<td>70</td>
<td>0.920</td>
<td>0.962</td>
<td>0.164</td>
</tr>
</tbody>
</table>

ROC AUC is for age of 65 years.
onstrated that abnormalities on FDT perimetry can precede detectable SAP damage by several years. In another study, we observed 105 patients with suspected glaucoma and demonstrated that functional abnormalities on FDT tests (N-30 strategy) were predictive of future onset and location of SAP visual field loss by as many as 4 years. In the present study, for earlier stages of damage (10% and 30% loss of neuroretinal rim), FDT had significantly higher areas under the ROC curves than did SAP SITA. It should be noted that, although FDT performed better than SAP SITA, most of the patients with early glaucomatous damage were still not detected by this test. However, for patients with more advanced damage and, therefore, more easily detectable disease, the diagnostic performances of the two tests were similar, with no statistically significant difference between the areas under the ROC curves.

FDT perimetry has recently undergone major modifications in its testing procedures. The use of smaller (5°) test stimuli in the 24-2 program incorporated in the commercially available Humphrey Matrix resulted in better characterization of the pattern of visual field defects in patients with glaucoma compared with the 10° stimuli pattern available in the first-generation version of the instrument, without adversely affecting within-test variability. The 24-2 pattern for FDT was developed based on the fact that the larger number of points would make this test more helpful for disease follow-up. Although our study suggests a benefit of FDT 24-2 in detecting early disease compared with SAP, its role for longitudinal assessment of visual field progression still has to be evaluated.

Other investigators have also evaluated the influence of covariates on the performance of diagnostic tests in glaucoma. Based on a logistic regression model proposed by Leisenring et al., Stroux et al. evaluated the influence of disease severity on the sensitivities of several different visual function and electrophysiological tests. The logistic model developed by Leisenring et al., however, was originally proposed for evaluation of tests with categorical results. Therefore, the evaluation of tests with continuous results using this approach requires that the tests results be dichotomized according to arbitrary cutoffs of specificity or sensitivity. The method used in the present study is advantageous, as the effects of covariates can be assessed on the whole ROC curve and therefore do not require dichotomization of test results.

In the present study, PSD was the only parameter used to evaluate the diagnostic performance of both FDT 24-2 and SAP SITA. This choice was based on our previous research and that of others showing that PSD is the best parameter for glaucoma diagnosis when several SAP SITA and FDT parameters are compared. The recent Ocular Hypertension Treatment Study also showed that PSD, but not MD, is a predictor of glaucoma development among ocular hypertensive subjects, suggesting that this parameter may be important for identification of early glaucoma cases. However, other studies have suggested the possibility that a generalized depression of sensitivity may be a prominent feature of early glaucoma cases and the visual field index MD would be more likely to capture this abnormality than PSD. To investigate this, we tested whether the use of the MD index instead of PSD in the ROC regression models would improve detection of glaucoma. Corresponding values of ROC curve areas for 10%, 30%, 50%, and 70% percentage of neuroretinal rim loss were 0.706, 0.806, 0.858, and 0.901, respectively, for SAP SITA MD and 0.727, 0.813, 0.881, and 0.931 for FDT 24-2 MD. It is interesting to note that, although ROC curve areas for SAP SITA MD and FDT 24-2 MD were lower than those for FDT 24-2 PSD, SAP SITA MD actually performed better than SAP SITA PSD for detection of early glaucoma, in agreement with the previous observations of diffuse sensitivity loss in early glaucoma, when evaluated by SAP.

As a global index, PSD may miss generalized loss or asymmetries on the visual field indicative of glaucoma. A more comprehensive comparison of SAP SITA and FDT 24-2 visual function tests would require evaluation of the number of abnormal points on total and pattern deviation plots, as well as of other indices, such as the glaucoma hemifield test. The performance of PSD seems to decrease in patients with very advanced disease due to a collapse of its mathematical calculation. This effect has been demonstrated recently on SAP visual fields of patients with values of MD worse than -24 dB (i.e., end-stage disease). Although this could have affected the evaluation of the influence of disease severity on our study,

Figure 3. Sensitivities at fixed specificities of 80% and 95% for SAP SITA PSD and FDT 24-2 PSD, according to the severity of disease.
the patients included in the analysis had a maximum percentage of neuroretinal rim loss of approximately 70%, and only four eyes had values of SAP SITA MD worse than −20 dB, indicating that patients with end-stage disease were not a major component in the study.

Our study has limitations. Disease severity was measured by the percentage of rim loss estimated from HRT II measurements. Although an experimental study in monkeys shows a good correlation between HRT topographic measures and histologic optic nerve fiber count, such evidence is not yet available for humans. Another limitation of our study is that the diagnosis of GON was based on cross-sectional assessment of stereophotographs. Ideally, for a more definitive diagnosis, progressive change of optic disc appearance would have to be demonstrated. Unfortunately, such longitudinal information is not yet available for all our patients. Future studies using progressive GON as the reference standard should be able to assess the performance of these tests under this circumstance.

Although ROC curves are a useful and important index for evaluation and comparison of the performance of diagnostic tests under certain circumstances, they are widely abused in other situations in which they are not particularly suited. For example, construction of ROC curves and application of the proposed regression methodology for tests with categorical results, particularly when there are only a few categories, requires careful attention to avoid bias on the calculation of the ROC curve areas. It should be noted that ROC curve areas also have limited intrinsic clinical meaning. Other indexes, such as likelihood ratios, may have more straightforward clinical interpretation and application. We have recently demonstrated the usefulness of likelihood ratio for interpretation of results of imaging tests in glaucoma. However, statistical methods for evaluation of covariate effects on likelihood ratios have not been well described in the literature and deserve further research.

In conclusion, we demonstrated that a regression methodology to evaluate covariate effects on ROC curves can be useful in the assessment of diagnostic tests for glaucoma. The proposed methodology allowed us to demonstrate a significantly better performance of FDT 24-2 compared to SAP SITA for diagnosis of glaucoma patients with early disease, whereas it maintained comparable performance in later-stage disease.

References

with threshold visual field tests in the same persons. Invest Ophthal Vis Sci. 2000;41:741–748.


