Automated Analysis of Heidelberg Retina Tomograph Optic Disc Images by Glaucoma Probability Score

Annemiek Coops,[1] *David Barry Henson,*[1,2] *Anna J. Kwartz,*[1] *and Paul Habib Artes*[1,2]

Purpose. To compare the diagnostic performance of the Heidelberg Retinal Tomograph's (HRT; Heidelberg Engineering GmbH, Dossenheim, Germany) glaucoma probability score (GPS), an automated, contour line–independent method of optic disc analysis with that of the Moorfields regression analysis (MRA).

Methods. HRT images were obtained from one eye of 121 patients with glaucoma (median age, 70.2 years; median mean deviation [MD], −3.6 dB, range, +2.0 to −9.9 dB) and 95 healthy control subjects (median age, 59.7 years; median MD −0.1 dB, range +2.5 to −3.7). The diagnostic performances of GPS and MRA were evaluated by including borderline classifications, either as test negatives (most specific criteria) or as test positives (least specific criteria). Agreement between global and sectoral data of both analyses was established. Logistic regression analyses were performed to evaluate the effect of covariates such as optic disc size and age on the classification outcomes of both the GPS and the MRA.

Results. In 8 (7%) patients with glaucoma and 10 (11%) control subjects, the GPS failed to provide a complete global and sectoral optic disc classification. Although we could not identify a single distinct cause of this failure in the glaucoma group, failures in the control subjects occurred most often (7/10) with small and crowded optic discs. In subjects who were successfully classified at least globally by the GPS (117 patients with glaucoma, 88 control subjects), the diagnostic performances of GPS and MRA were similar (areas under the receiver operating characteristic [ROC] curve of 0.78 and 0.77, respectively; $P > 0.1$). With the GPS, sensitivity and specificity were 59% and 91% (most specific criteria) and 78% and 63% (least specific criteria), respectively. Combining GPS and MRA did not increase diagnostic performance significantly (ROC area of combined classifiers, 0.81). Both GPS and MRA were affected by disc size. In patients with glaucoma as well as healthy control subjects, the odds of a positive GPS classification (borderline or outside normal limits) increased by 21% (95% confidence interval [CI], 12%–30%) for each 0.1 mm^2 increase in optic disc area. With the MRA, the corresponding increase was 15% (95% CI, 7%–23%). Optic disc area alone accounted for approximately 30% and 22% of the explained variance with the GPS and MRA, respectively ($P < 0.001$). The proportional-odds logistic regression confirmed that optic disc size affected mainly the tradeoff between true- and false-positive classifications (criterion) rather than the absolute performance of the analyses (area under the ROC curve). There was some evidence of an age effect with the MRA, which showed a 53% (95% CI, 16%–102%) increase in the odds of a positive test (borderline or outside normal limits) associated with each decade of age ($P = 0.002$), but no age effects were observed with the GPS ($P > 0.1$).

Conclusions. The diagnostic performance of the contour line–independent GPS analysis is similar to that of the MRA. However, clinicians should be aware of the strong size dependence of both GPS and MRA. In large optic discs, both GPS and MRA are likely to produce many false-positive classifications. Correspondingly, the sensitivity to early damage is likely to be low in small optic discs. There is a need for automated classification systems that explicitly address the size dependence of current analyses. (*Invest Ophthalmol Vis Sci.* 2006;47:5348–5355) DOI:10.1167/iovs.06-0579

Glaucoma is a degenerative disease of the optic nerve, characterized by morphologic changes in the optic disc and the retinal nerve fiber layer and corresponding losses in the visual fields. Signs associated with glaucomatous optic nerve damage include progressive enlargement of the optic cup, focal notches in the neuroretinal rim, optic disc hemorrhages, nerve fiber layer defects, and parapapillary atrophy.[1] It is often difficult to distinguish early signs of glaucomatous optic disc damage from physiological variations of optic nerve appearance. Subjective evaluation of the optic disc (either by ophthalmoscopy or by inspection of stereophotographs) is highly dependent on the skill and the experience of the examiner, and even expert examiners do not always agree with each other.[2–6] Objective imaging technologies therefore carry some promise in helping clinicians to assess and document the status of the optic disc and are increasingly used to complement traditional methods of optic disc assessment.

The Heidelberg Retina Tomograph (HRT; Heidelberg Engineering GmbH, Dossenheim, Germany) is a confocal scanning laser ophthalmoscope that acquires three-dimensional topography images of the optic disc and the surrounding retina. Until recently, all diagnostic analyses of the HRT depended on the position of a manually placed contour line to outline the area of the optic disc. However, contour lines drawn by different observers vary, sometimes considerably.

A fully automated diagnostic decision-support system that does not rely on a manually drawn contour line has recently been incorporated into the software of the HRT. This analysis, referred to as the glaucoma probability score (GPS), is based on an innovative technique proposed by Swindale et al.[7] who fit a surface over the area of the optic disc and parapapillary retina. In the implementation incorporated in the commercial HRT software, a Bayesian machine-learning classifier then compares the parameters of the fitted surface to those obtained in healthy and glaucomatous optic discs and derives a numerical index for the likelihood of damage.

From the [1]Manchester Royal Eye Hospital, Manchester, United Kingdom; and the [2]Research Group for Eye and Vision Sciences, University of Manchester, Manchester, United Kingdom.

Presented at the annual meeting of the Association for Research in Vision and Ophthalmology, Fort Lauderdale, Florida, May 2006.

Supported by Nova Scotia Health Research Foundation Project Grant Med-727 (PHA) and National Health Service Research and Development Grant 95/18/04 (DBH).

Submitted for publication May 29, 2006; revised July 28 and August 16, 2006; accepted October 4, 2006.

Disclosure: **A. Coops**, None; **D.B. Henson**, None; **A.J. Kwartz**, None; **P.H. Artes**, None

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: Paul Habib Artes, Faculty of Life Sciences, University of Manchester, Office C32, Moffat Building, North Campus, PO Box 188, M60 1QD, UK; paul_h_artes@yahoo.co.uk.

In this study, we applied the GPS analysis to an independent sample of healthy and glaucomatous eyes from the Manchester Imaging Study. We estimated the diagnostic accuracy of the technique and compared it with that of the MRA. To establish how these analyses are affected by optic disc size, age, and image quality, we performed logistic regression analyses to determine the effect of these covariates on the classifications.

METHODS

Subjects

The data were obtained from the Manchester Imaging Study, a prospective study on the role of imaging technologies in the diagnosis and management of glaucoma. The study was performed at the Manchester Royal Eye Hospital (MREH) from 1998 to 2004 and included both cross-sectional and longitudinal arms.[8] It adhered to the tenets of the Declaration of Helsinki and was approved by the Central Manchester Research Ethics Committee. Informed written consent was obtained from all participants.

The patients with glaucoma were consecutively recruited from the clinics at the MREH. Inclusion criteria were a clinical diagnosis of open-angle glaucoma, age >40 years, refractive error within ±5.00 D equivalent sphere and ±3.00 D astigmatism, best corrected visual acuity (VA) of 6/18 (+0.5 logMAR; logarithm of the minimum angle of resolution), and a repeatably detectable visual field defect. Visual fields were examined using program 24-2 of the Humphrey Field Analyzer (HFA; Carl Zeiss Meditec, Dublin, CA) with the full-threshold strategy, and a defect was defined as a glaucoma hemifield test (GHT) result outside normal limits and/or a corrected pattern standard deviation (CPSD) significantly elevated beyond the 5% level. For the analyses reported herein, eyes with mean deviation [MD] worse than −10 dB were excluded. If both eyes were eligible, one eye was randomly selected as the study eye. Of those patients who participated in the longitudinal arm of the study, we selected the image in which the mean pixel height standard deviation (MPHSD), a measure of image quality, was closest to the median value observed during the entire follow-up; thus, we analyzed the most representative image of the available series. Three patients and three healthy control subjects were removed from the analysis because their median MPHSD exceeded 50 μm, the recommended cutoff for image quality (HRT user manual; Heidelberg Engineering GmbH). Normal control subjects were recruited from patients' spouses and by advertising through leaflets and posters distributed in local medical centers, universities, and other communal areas. Inclusion criteria were age >40 years, intraocular pressure below 22 mm Hg, refractive error within ±5.00 D equivalent sphere and ± 3.00 D astigmatism, best corrected VA of 6/18 (+0.5 logMAR), and normal findings in a visual field examination (HFA 24-2 full-threshold test, both CPSD >10% and GHT results within normal limits).

Data Collection

All participants were imaged with the HRT1 (Heidelberg Engineering GmbH). In each eye, five individual 10° scans were obtained. If the pupil diameter was smaller than 3 mm or if image quality was thought to be affected by media opacity, the pupil was dilated with 0.5% tropicamide. The three best images, judged by the examiner, were used to generate a mean topographic image that was subsequently recalculated and converted to HRT3 format by version 3.0.2 of the HRT viewer module. Contour lines were placed on the margin of the optic disc by experienced users, according to the instructions provided on the HRT Web site (www.heidelbergengineering.com). All contour lines were reviewed by the authors.

Glaucoma Probability Score

The principles of the analyses underlying the GPS have been described by Swindale et al.[7] Briefly, a geometric model is used to approximate the optic disc topography with a three-dimensional surface, described by five parameters of optic disc and peripapillary retinal shape. With a standard non–linear least-squares fitting technique, these parameters are adapted to the individual topography globally as well as in six separate sectors of the optic disc (Fig. 1). The obtained parameters are then interpreted by a relevance vector machine, a state-of-the-art machine learning classifier operating on Bayesian principles,[9] which provides a numerical index ranging from 0 (low probability of disease) to 1 (high probability of disease), to describe the estimated probability of finding similar data in the glaucoma group of the training data. Although details of the training data are not in the public domain, the score can be interpreted as an ordinal index of abnormality. Accordingly, sectors with a value of >0.28 or >0.64 (Volz D, Heidelberg Engineering GmbH, personal communication, November 2005) are classified as borderline or outside normal limits and flagged with yellow exclamation marks or red crosses. The overall outcome of the GPS analysis is determined by the sector with the highest probability score (worst result of global and sectoral analysis).

Moorfields Regression Analysis

The MRA has been described previously.[10] In brief, a reference plane is derived from the height of the contour line at the papillomacular bundle (350°–356°) and placed, by default, 50 μm posterior.[11] The software then segments the area inside the contour line into the optic cup (below reference plane) and the neuroretinal rim (above reference plane). The MRA compares the neuroretinal rim area globally and individually in six sectors with values predicted for a healthy subject with the same disc size and age. If the observed rim area is smaller than the 95% or 99.9% prediction limits (derived by linear regression of log rim area and disc area), it is classified as borderline or outside normal limits, respectively. Similar to the GPS analysis, the overall outcome of the MRA is determined by the most abnormal sector.

Analysis

Both GPS and MRA have a borderline category for optic discs not clearly identified as either within or outside normal limits. Clinically, this graded approach may be more useful than any somewhat arbitrary division into healthy and diseased categories necessary for the calculation of sensitivity and specificity. We therefore evaluated diagnostic

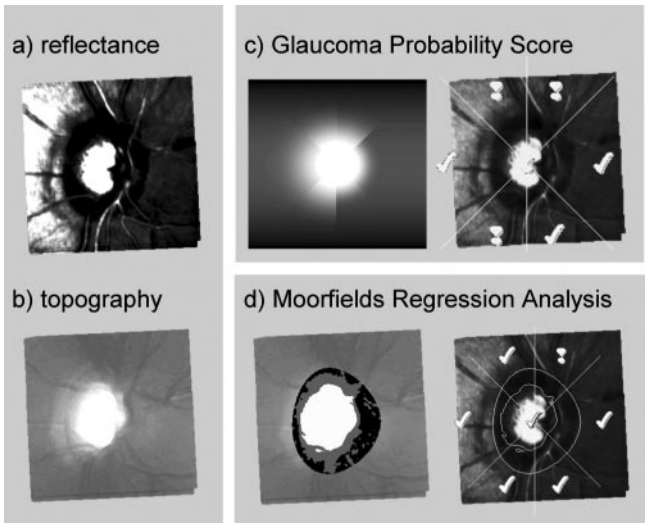

FIGURE 1. Example of optic disc classification with GPS and MRA. Reflectance (**a**) and topography (**b**) images. The surface fit of the GPS analysis (**c**, *left*) gives a borderline classification globally and in three optic disc sectors (**c**, *right*). With the MRA, the neuroretinal rim area is estimated to be within normal limits globally and in all but the nasal–superior optic disc sectors. The overall classification is borderline with both GPS and MRA.

TABLE 1. Details of the Glaucoma and Normal Control Groups

	Glaucoma (n = 121)	Controls (n = 95)
Age (y)	70.2 (38.2 to 89.2)	59.7 (39.6 to 87.9)
MD (dB)	−3.6 (+2.0 to −9.9)	−0.1 (+2.5 to −3.7)
PSD (dB)	4.1 (1.2 to 10.6)	1.7 (1.0 to 3.5)
Disc size (mm^2)	2.0 (1.3 to 3.4)	1.9 (1.1 to 3.2)
Image quality (MPHSD)	24 (11 to 48)	23 (10 to 48)

Data are medians and ranges.

performance in two separate ways, first by considering borderline cases as test negatives (most specific but least sensitive criterion) and second by considering borderline cases as test positives (least specific but most sensitive criterion). Similarly, the agreement between the overall classification of the GPS analysis and the MRA was analyzed by a contingency table with three categories (within normal limits, borderline, and outside normal limits).

To evaluate differences in the topographical information provided by the GPS and MRA, we compared the classifications in each of the six optic disc sectors. This analysis was performed separately in patients with glaucoma and healthy control subjects. A combined classification of GPS and MRA was derived to assess whether such a combination improves diagnostic performance over that achieved with either GPS or MRA alone.

Several groups have reported a difference in sensitivity and specificity of optic disc assessment in small versus large optic discs.[12,13] Because we wanted to evaluate such effects quantitatively, while accounting for other covariates, we performed proportional odds logistic regression (POLR).[14,15] Similar to the more familiar binary logistic regression, POLR models the effect of several covariates, providing odds ratios and measures of relative importance. Whereas standard logistic regression models binary outcomes, POLR models ordinal outcomes (such as the classification as within normal limits, borderline, and outside normal limits in the present analysis). In preliminary analyses, separate logistic regression models were fitted to establish the effect of visual field loss, measured by the MD and PSD global indices, in the patients with glaucoma and the healthy control group, and to confirm that the assumptions of a combined POLR model were met. To obtain the final POLR model, disease status (glaucoma or control) was entered as a factor alongside the continuous explanatory variables (optic disc size, age, and image quality). Stepwise backward analyses were performed manually to remove those covariates that did not appear to contribute significantly to the model (likelihood ratio χ^2-test, $P > 0.1$). Nagelkerke R^2 values were used to express the proportion of variance explained by the model. The predicted probabilities of a borderline and outside normal limits classification in patients with glaucoma and healthy control subjects with various optic disc sizes were combined into a "pseudoROC" curve for comparison with the empiric data. All analyses were performed in the open-source statistical environment R.[15–17] Anonymized data and code for this analysis are available from the corresponding author on request.

RESULTS

Demographics

Details of the glaucoma and control groups are given in Table 1. The patients with glaucoma were older (median age difference, 10.5 years; $P < 0.001$) and had somewhat larger optic discs (median difference, 0.1 mm^2; $P = 0.003$) compared with the control subjects.

GPS Unable to Classify

In four (3%) patients with glaucoma, and seven (7%) normal control subjects, the GPS was unable to match the surface of the optic disc and therefore failed to provide a classification. In four other cases (one patient with glaucoma, three control subjects) the GPS provided a global but not a sectoral classification. Although we could not identify a single distinct cause of failure in the glaucoma group, failures in the control eyes occurred most often in small, crowded discs (Table 2; see Fig. 2, for an example). The subsequent analyses refer only to the data classified at least globally by the GPS.

Diagnostic Accuracy

With the GPS, sensitivity and specificity were 59% and 91% (most specific criteria) and 78% and 63% (least specific criteria). Similar results were found with MRA, with sensitivity and specificity of 56% and 87% (most specific criteria) and 78% and 66% (least-specific criteria; Fig. 3). The overall diagnostic performances of GPS and MRA, measured by ROC analysis, were similar, with areas under the ROC curve of 0.78 and 0.77, respectively (pair-wise comparison of ROC curves; $P > 0.1$).[18]

Agreement between GPS and MRA

Based on the subjects who were successfully classified by the GPS (117 patients with glaucoma, 88 control subjects), close agreement was found between the overall classifications of GPS and MRA. Complete agreement was observed in 71% and 68% of glaucomatous and healthy control cases, respectively. Partial agreement (classified by one analysis as borderline and by the other as either within normal limits or outside normal limits) was observed in 24% and 26% of glaucoma and control cases, respectively (Table 3).

TABLE 2. Optic Disc Area in Cases in Which the GPS was Unable to Provide a Classification

Group	Classified	Not Classified	P
Glaucoma (n = 121)	2.01 (1.30–3.37) (n = 117)	1.80 (1.65–2.18) (n = 4)	0.37
Controls (n = 95)	1.90 (1.07–3.24) (n = 88)	1.47 (1.05–1.88) (n = 7)	0.007

Data are disc area (mm^2; median; range). Healthy control discs not classified by the GPS were significantly smaller than those classified. Because of the small sample size, statistical significance was derived from 10.000 bootstrap replications of the Mann-Whitney test.

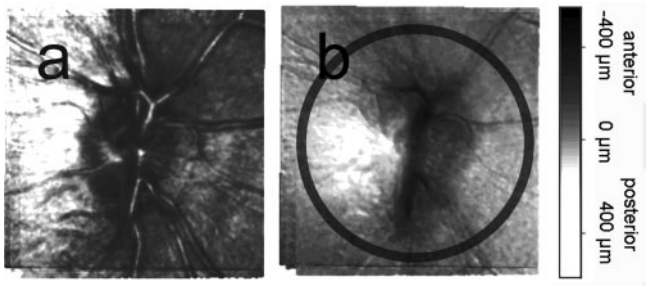

FIGURE 2. Example of optic disc that could not be classified with the GPS. The reflectance image (**a**) shows a small optic disc (area, 1.15 mm^2), whereas the topography image (**b**) reveals a prominent nasal surface. The legend refers to surface height relative to the reference ring (**b**, *circular gray area*).

To investigate whether it would be useful to combine GPS and MRA, the performance of a combined classifier was compared to that of either classification system alone. Owing to the close agreement of both classifications, the improvement in performance with the combined classifier was only small (ROC area of 0.81 with combined classifier, compared with 0.78 and 0.77, respectively, with GPS or MRA alone) and not statistically significant ($P > 0.10$).[18]

Topographical Differences

Whereas, with the MRA, the inferior sectors of the optic disc were more often classified as borderline and outside normal limits compared with superior sectors, topographical differences were much less apparent with the GPS (Fig. 4). To compare the association between each of the six sectors, we established the Cramer V statistic[19] with the GPS and the MRA in both groups of subjects (Fig 5). The Cramer V statistic expresses the association between two categorical variables as a proportion relative to their largest possible variation; a value of 0 means no association, and 1 stands for a perfect association. In both groups of subjects, GPS classifications of individual disc sectors were closely associated with each other (Fig. 5). For example, the association between the temporal and nasal sectors of glaucomatous optic discs was 0.95 with the GPS, whereas the corresponding value with the MRA was 0.47. In the glaucomatous eyes, the associations between sectors classified with the GPS ranged from 0.86 to 0.99, whereas those of the MRA ranged from 0.26 to 0.57. A high between-sector association with the GPS was also observed in the healthy control eyes.

Effects of Disc Size, Age, Image Quality, and Visual Field Damage

The results of the POLR analyses are shown in Table 4. With the GPS, only optic disc area was found to have a significant effect on the classification, with an estimated 21% increase in the odds of a positive classification (either borderline or outside normal limits) for each 0.1 mm^2 increase in area (95% CI, 12%–30%). Neither age nor image quality appeared to have an effect on the GPS classifications.

With the MRA, the effect of disc size was similar (approximately 15% increase in the odds of a classification as borderline or outside normal limits). In addition, there appeared to be a significant age effect. With each decade of age, the odds of a positive MRA increased by approximately 50% (95% CI, 16%–102%). This age effect accounted for roughly 10% of the explained variance with the MRA. Image quality, measured by global MPHSD, did not contribute to the model with either GPS or MRA. Of interest, the amount of visual field damage, measured by the global indices MD and PSD, did not appear to affect the classification outcome within each group ($P > 0.1$, likelihood ratio χ^2).

To illustrate the effects of disc size on the discrimination between patients with glaucoma and healthy control subjects and to assess graphically the fit of the models to the empiric data, we derived the estimated probabilities for positive classifications (borderline, outside normal limits) in patients with glaucoma and healthy control subjects, with optic disc sizes ranging from 1.0 to 3.0 mm^2, at an mean age of 65 years (Figs. 6a–d). A "pseudoROC" curve was then constructed from the predicted probabilities in patients with glaucoma (sensitivity) and healthy control subjects (false-positive rate) at the given range of optic disc sizes (Figs. 6e, 6f). To compare the predictions of the model with the empiric data, the entire study sample was arbitrarily split into six approximately equal-sized groups according to optic disc size, and sensitivity and specificity estimates were derived within each of those subgroups (Figs. 6e, 6f).

DISCUSSION

Objective optic disc analyses can support and guide decision-making in clinical practice, but clinicians need to be aware of their strengths and limitations. In this article, we investigated the diagnostic performance of the GPS, a fully automated method of interpreting optic disc topography images acquired with the HRT, in an independent sample of healthy and glaucomatous eyes. We compared the GPS with the MRA, a widely used classification system.[20] One of the limitations of the MRA is its dependence on the position of the contour line. Several groups have shown that, with some optic discs, even expert users of the HRT may differ considerably in the placement of the contour lines,[21] introducing an element of subjectivity into an otherwise objective classification process. The contour line independence of the GPS is therefore an important advance.

GPS and MRA had similar diagnostic performances and agreed in most eyes (complete agreement in ~70% of eyes and at least partial agreement in ~95% of cases). When the large

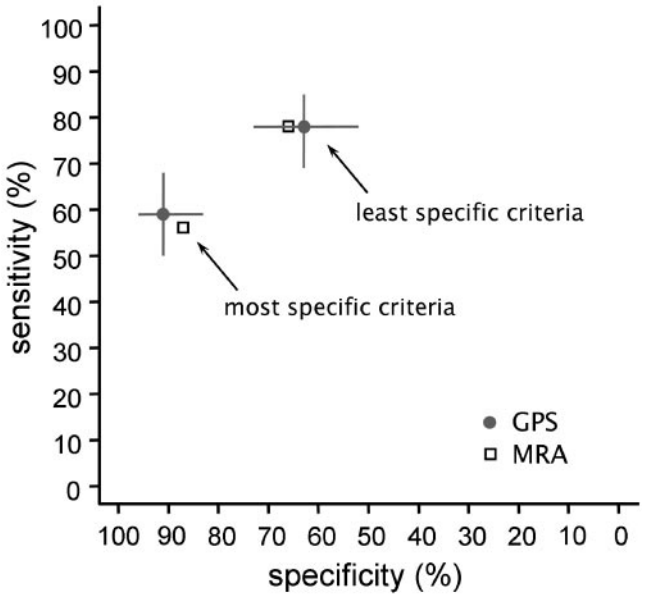

FIGURE 3. Sensitivities and specificities of GPS and MRA when borderline results were included as test negatives (most specific criteria) and positives (least specific criteria). The error bars (shown for the GPS only) indicate the 95% CIs of the obtained proportions.

TABLE 3. Agreement between Overall Classifications with GPS and MRA in Glaucoma Patients and Healthy Controls

	MRA				MRA		
GPS	**N**	**B**	**O**	**GPS**	**N**	**B**	**O**
Glaucoma (n = 117)				Controls (n = 88)			
N	17 (15%)	5 (4%)	4 (3%)	N	46 (82%)	6 (7%)	3 (3%)
B	5 (4%)	9 (8%)	8 (7%)	B	8 (9%)	11 (13%)	6 (7%)
O	3 (3%)	10 (9%)	56 (48%)	O	2 (2%)	3 (3%)	3 (3%)

N, within normal limits; B, borderline; O, outside normal limits. Data are the number of eyes; sample percentages are stated in parentheses.

proportion of borderline classifications (19% in the glaucomatous eyes, 28% in the control eyes) were included as test negatives, the GPS analysis performed with specificity close to 90% and sensitivity of just below 60%. When borderline classifications were included as test positives, sensitivity increased to ~80%, whereas specificity dropped to ~60%. With respect to the MRA, our findings were similar to those reported by Ford et al.[13] in a different dataset in which the glaucomatous eyes had a similar degree of glaucomatous visual field damage (median MD, −4.8 dB). Combining the two classifiers did not produce a meaningful gain in performance (i.e., the two analyses did not appear to provide complementary information).

Similar to the MRA, the GPS attempts to localize damage to six sectors of the optic disc. We therefore compared the frequency with which borderline or outside-normal-limits classifications are made with GPS and MRA in each of the six sectors. Whereas with the MRA the nasal–inferior and temporal–inferior optic disc sectors of the glaucomatous eyes were classified as borderline or outside normal limits more often than other sectors, no distinct spatial predominance was seen with the GPS. In fact, the GPS classifications within the individual disc sectors appeared strongly associated with each other, adding little if any information to that already available from the global classification.

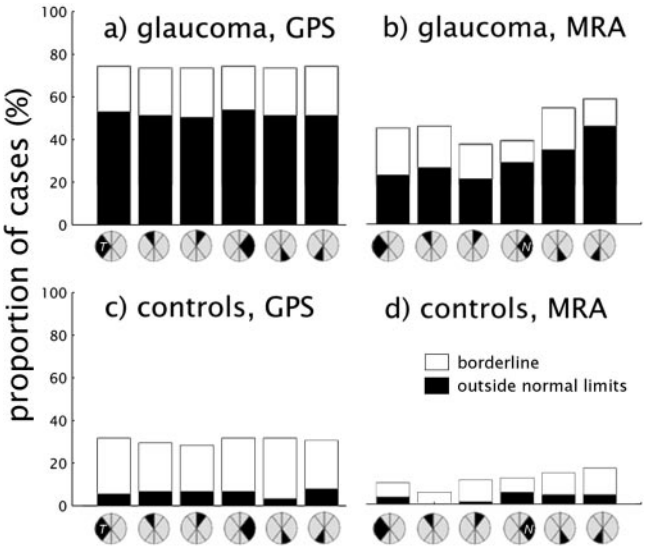

FIGURE 4. Sectoral classification with the GPS (*left*) and MRA (*right*) in glaucomatous eyes (*top*) and healthy control subjects (*bottom*). While the MRA classified the inferior optic disc sectors more often as outside normal limits or borderline, almost no such variations were observed with the GPS.

GPS models the shape of the optic disc on a simple geometric model from which it extracts several shape parameters.[7] The simplicity of its underlying model is an attractive feature of the analysis; each of the parameters has a straightforward morphologic interpretation. However, in a sizable proportion of eyes in our study, both glaucomatous (n = 4, 3%) and healthy (n = 7, 7%), the GPS algorithm did not find a satisfactory surface fit compatible with the optic disc topography and therefore failed to provide a classification. In the glaucoma group we did not find a single distinct cause for failure, but most of the unclassified discs in the healthy control group were small and crowded (Table 2). The Manchester Glaucoma Imaging Study,[8] from which our data were drawn, had excluded participants with refractive error outside the range of ±5.00 D equivalent sphere and ±3.00 D astigmatism. It is likely that subjects with high axial hyperopia and myopia have a relatively larger prevalence of optic discs outside the spectrum of typical appearances.[22,23] In these subjects, automated (GPS) or semiautomated (MRA) analyses are likely to be less reliable. If, as we believe, decision-support systems are particularly important in those cases that pose a challenge to clinicians, the performance of automated systems in atypical optic discs (for example in those with particularly small or large size) has not as yet been satisfactorily addressed by research.

The GPS derives its numerical index by a relevance vector machine, a state-of-the-art machine learning classifier potentially capable of solving complex classification problems.[9] Machine learning classifiers have previously been applied, for example by Bowd et al.[24] (who trained support vector ma-

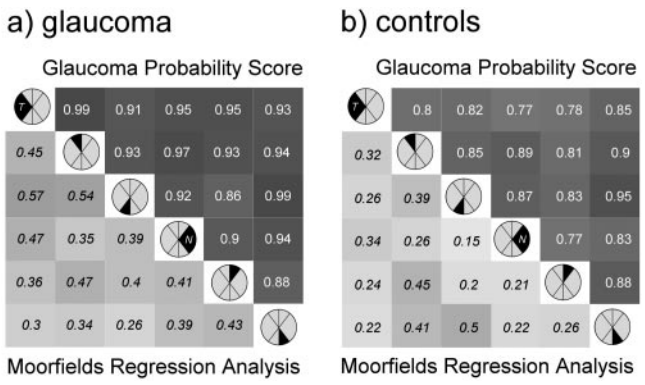

FIGURE 5. Association between disc sectors with the GPS and MRA, in patients with glaucoma (**a**) and healthy control eyes (**b**). Cells are gray-scale coded according to the level of association (Cramer V). Between-sector associations were much higher with the GPS than with the MRA, in both groups of subjects.

TABLE 4. Results of POLR Analysis

	Glaucoma Progression Score		Moorfields Regression Analysis	
	Odds Ratio (95% CI)	**R^2, % (*P*)**	**Odds ratio (95% CI)**	**R^2, % (*P*)**
Diagnosis (glauc/contr)	7.80 (4.27–14.2)	27 (<0.001)	5.20 (2.80–9.50)	25 (<0.001)
Disc area (0.1 mm^2)	1.21 (1.12–1.30)	38 (<0.001)	1.15 (1.07–1.23)	30 (<0.001)
Age (decades)	1.22 (0.1–16)	39 (=0.17)	1.53 (1.16–2.02)	34 (=0.002)

The R^2-value (Nagelkerke) gives the cumulative proportion of variance explained after adding the respective variable to the model. Adding disc area, for example, increased the explained variance from 27% to 38% (i.e. by 11%), in the GPS analysis. The *P* value refers to the likelihood ratio χ^2 test.

chines to discriminate between healthy and diseased eyes based on the conventional stereometric parameters), and by Hothorn and Lausen[25] who used classification trees to discriminate between optic disc surfaces in patients with glaucoma and healthy control subjects imaged with the HRT. Machine learning tools may be more robust and flexible compared with the traditional (and more familiar) statistical analyses whose validity may stand or fall with assumptions that may not be met in clinical populations. The MRA, for example, performs statistical hypothesis testing by comparing the measured rim area with that expected in an age-matched healthy subject with similar disc size, but there is evidence that at least three assumptions (linearity of the relationship between log rim and disc area, equality of variance across the entire range of values, Gaussian properties of the underlying distributions) of this analysis may not be met in clinical data. One of the potential advantages of the particular Bayesian classifier applied in the GPS is that it provides a probabilistic interpretation of optic disc status, rather than a simple binary classification into normal or abnormal.[9] However, in the absence of published details on the training sample, the precise statistical interpretation of the GPS, beyond that of an ordinal index of optic disc abnormality, remains somewhat unclear.

The most striking, though not unexpected, finding of our study is the large dependence of both GPS and MRA on optic disc size. Both GPS and MRA showed poor sensitivity to damage in small optic discs and poor specificity in large optic discs (Fig 6). These findings mirror those previously reported from others with the MRA[13,26] and are similar to those reported from expert observers.[27,28] A classification as borderline or outside normal limits in a small optic disc is much more likely to be a true finding than is the same result in a large optic disc (which is relatively more likely to be a false positive). With increasing optic disc size, the probability of a borderline or outside-normal-limits result increases in both the patients with glaucoma and the healthy control subjects, to a similar extent. Consequently, optic disc size primarily affects the *criterion* (i.e., the threshold used to discriminate between healthy and glaucomatous discs), rather than the absolute accuracy of the analyses (Fig. 6). We believe that the most likely explanation for the apparent criterion shift of the GPS is a sampling bias in the training data. If early damage is difficult to detect in small optic discs by ophthalmoscopy, such discs are likely to be relatively underrepresented among the patients with glaucoma attending secondary centers. Similarly, if damage is more readily detectable in large discs, eyes with large discs may be relatively overrepresented. A more representative sample of glaucomatous optic disc damage could be drawn from subjects exhibiting glaucomatous visual field damage in a screening study, so that disc-related features do not influence selection. However, the number of subjects necessary to obtain a sufficiently large sample makes this approach impractical. Sampling biases in the source population are likely to persist in study populations despite strict inclusion and exclusion criteria. The small but statistically significant difference in optic disc size between patients with glaucoma and healthy control subjects seen in our sample is compatible with this explanation. Although we believe that sampling biases provide the most likely, as well as the most parsimonious, explanation for the size dependence of both GPS and MRA, at least two alternative explanations should be considered. First, larger discs may be at a greater risk of the structural damage characteristic of glaucoma.[29] Second, glaucomatous damage itself may lead to expansion of the scleral canal[30] and therefore be responsible for the slightly larger optic disc size of the patients with glaucoma.

Independent of whether any disc size differences between patients with glaucoma and healthy control subjects are due to sampling bias, our data provide evidence for a large size-dependent criterion shift, with both the GPS and the MRA. Since even experts may find it more difficult to ascertain glaucomatous changes in small optic discs, and more difficult to rule out glaucomatous changes in a large optic disc, this issue may present a considerable problem in clinical practice. Ideally, an objective analysis would complement the clinicians' judgment and would be particularly valuable in those cases that pose a known challenge to subjective evaluation (for example particularly large or small discs). If our findings are confirmed by other centers, it would be worthwhile (and probably not difficult) to remove any size-related criterion shifts in the analysis.

An unexpected finding, to our knowledge not previously reported, was the apparent age dependence of the MRA results. Our data suggest that, with each decade of increasing age, the odds of a borderline or outside-normal-limits result with the MRA increase by roughly 50%, in patients with glaucoma and in healthy control subjects. Although the CIs around the odds ratio were large (16%–102%), age accounted for nearly 12% of the explained variance in the data and was confirmed as a significant predictor, even when glaucomatous and control eyes were evaluated in separate logistic regressions. Because the MRA explicitly accounts for the subjects' age in calculating the expected neuroretinal rim area,[10] we can only speculate that the age-related decline in rim area factored into the MRA is smaller than that observed in our sample. Because the age coefficient of the MRA was estimated by linear regression, it may have been attenuated by the large biological variation in neuroretinal rim area between subjects. It is important for our findings to be confirmed by other centers.

Finally, our findings underscore the importance of estimating covariate effects when evaluating the performance of diagnostic tests. Although the authors of the STARD initiative (Standards for the Reporting of Diagnostic Accuracy) have called for more complete reporting of relevant covariates (e.g., measures of disease severity) that may influence diagnostic performance,[31] reporting of such details has been incomplete in many previous studies on the HRT.[32] Because relevant covariates often vary between study samples, the quantitative modeling of their effects both on diagnostic performance[33]

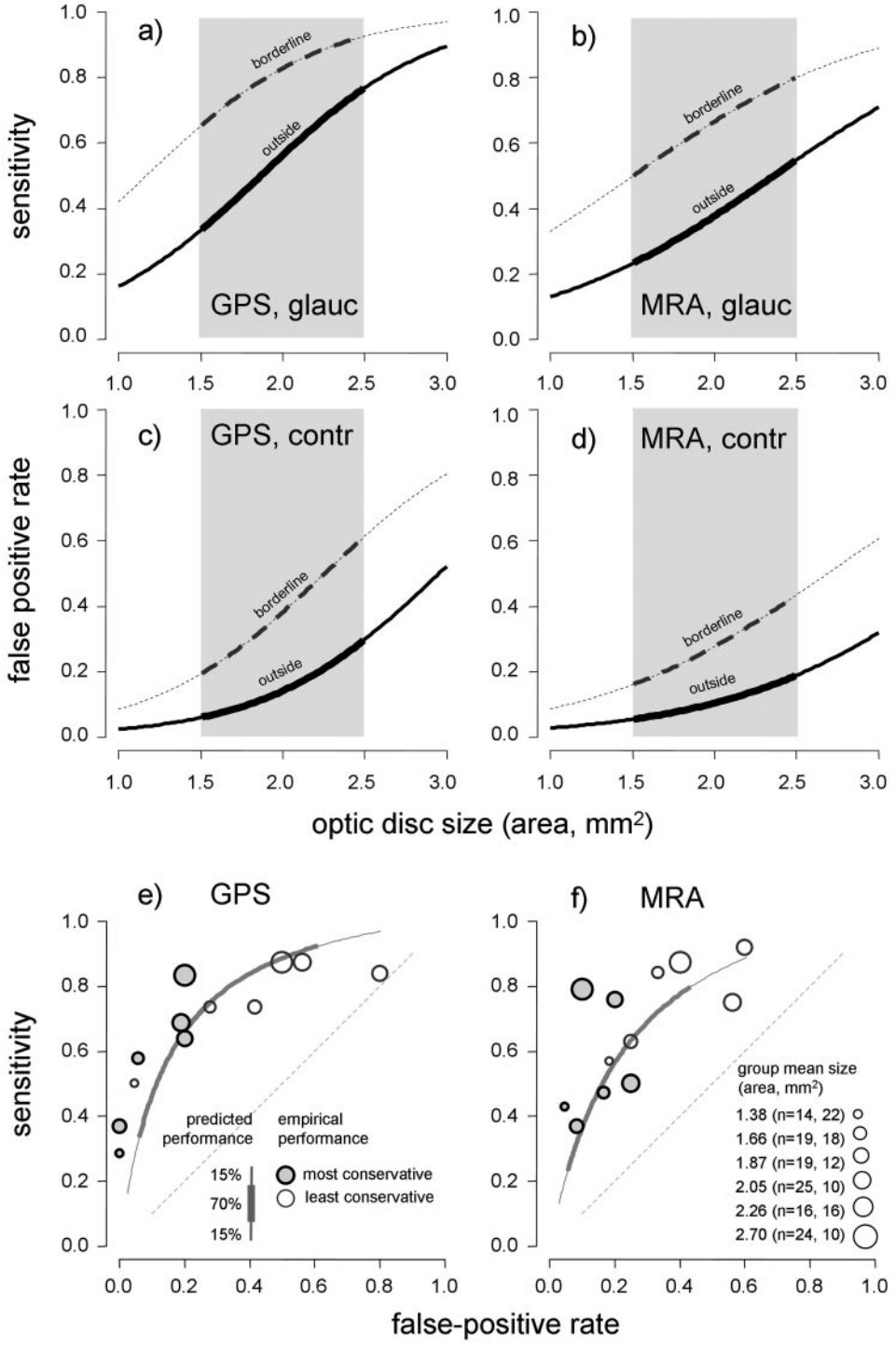

Figure 6. (**a**–**d**) Probability for a positive test outcome in patients with glaucoma (**a**, **b**) and healthy control subjects (**c**, **d**) with the GPS (**a**, **c**) and the MRA (**b**, **d**). The curves, estimated by POLR, show the probability of having a borderline or outside-normal-limits result at optic disc sizes ranging from 1.0 to 3.0 mm^2. The range of disc sizes corresponding to the central 70% of our data (1.5–2.5 mm^2) are shown by the *gray rectangle* and the heavier lines. (**e**, **f**) Estimates of (**a**–**e**) combined into "pseudoROC" curves (*gray line*) that shows the covariation of sensitivity and specificity with optic disc size with GPS (**e**) and MRA (**f**). *Shaded circles*: the sensitivity–specificity pairs from applying the most specific criteria (including borderline results as test negatives) to each of the six subgroups stratified by optic disc size; *open circles*: results for the least specific criteria (including borderline results as test positives). The heavy segment of the *gray line* shows the variation across the central 70% of optic disc sizes in our sample (corresponding to *gray rectangles* in **a**–**d**).

and on the tradeoff between sensitivity and specificity, may make it easier to compare findings between different centers. Moreover, covariate modeling may occasionally reveal important features of the tests that may otherwise not have been apparent.

Acknowledgments

The authors thank Anne Bjerre (University of Sheffield) and Dietmar Volz (Heidelberg Engineering) for technical assistance, Ivan M. Franch of The University of Manchester for advice on the Cramer V, and an anonymous reviewer for providing helpful comments on the statistical analysis.

References

1. Jonas JB, Budde WM. Diagnosis and pathogenesis of glaucomatous optic neuropathy: morphological aspects. *Prog Retin Eye Res*. 2000;19:1–40.
2. Lichter PR. Variability of expert observers in evaluating the optic disc. *Trans Am Ophthalmol Soc*. 1976;74:532–572.
3. Tielsch JM, Katz J, Quigley HA, Miller NR, Sommer A. Intraobserver and interobserver agreement in measurement of optic disc characteristics. *Ophthalmology*. 1988;95:350–356.
4. Varma R, Steinmann WC, Scott IU. Expert agreement in evaluating the optic disc for glaucoma. *Ophthalmology*. 1992;99:215–221.
5. Zangwill L, Shakiba S, Caprioli J, Weinreb RN. Agreement between clinicians and a confocal scanning laser ophthalmoscope in estimating cup/disk ratios. *Am J Ophthalmol*. 1995;119:415–421.
6. Azuara-Blanco A, Katz LJ, Spaeth GL, Vernon SA, Spencer F, Lanzl IM. Clinical agreement among glaucoma experts in the detection of glaucomatous changes of the optic disk using simultaneous stereoscopic photographs. *Am J Ophthalmol*. 2003;136:949–950.
7. Swindale NV, Stjepanovic G, Chin A, Mikelberg FS. Automated analysis of normal and glaucomatous optic nerve head topography images. *Invest Ophthalmol Vis Sci*. 2000;41:1730–1742.
8. Kwartz AJ, Henson DB, Harper RA, Spencer AF, McLeod D. The effectiveness of the Heidelberg Retina Tomograph and laser diagnostic glaucoma scanning system (GDx) in detecting and monitoring glaucoma. *Health Technol Assess*. 2005;9:1–148.
9. Tipping ME. Sparse Bayesian learning and the relevance vector machine. *J Mach Learn Res*. 2001;1:211–244.
10. Wollstein G, Garway-Heath DF, Fontana L, Hitchings RA. Identifying early glaucomatous changes: comparison between expert clinical assessment of optic disc photographs and confocal scanning ophthalmoscopy. *Ophthalmology*. 2000;107:2272–2277.
11. Burk RO, Vihanninjoki K, Bartke T, et al. Development of the standard reference plane for the Heidelberg retina tomograph. *Graefes Arch Clin Exp Ophthalmol*. 2000;238:375–384.
12. Agarwal HC, Gulati V, Sihota R. The normal optic nerve head on Heidelberg Retina Tomograph II. *Indian J Ophthalmol*. 2003;51:25–33.
13. Ford BA, Artes PH, McCormick TA, Nicolela MT, LeBlanc RP, Chauhan BC. Comparison of data analysis tools for detection of glaucoma with the Heidelberg Retina Tomograph. *Ophthalmology*. 2003;110:1145–1150.
14. McCullagh P. Regression models with ordinal data. *J Royal Stat Soc Series B.* 1980;42:109–142.
15. Venables WN, Ripley BD. *Modern Applied Statistics with S.* 4th ed. New York: Springer; 2002:204–205.
16. Ihaka R, Gentleman R. R: a language and environment for statistical computing. *J Comput Graph Stat*. 1996;5:299–314.
17. Harrell FE. Design Package. Nashville, TN: Vanderbilt University. Available at http://biostat.mc.vanderbilt.edu/s/Design
18. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*. 1983;148:839–843.
19. Liebetrau AM. *Measures of Association.* Newbury Park, CA: Sage Publications; 1983.
20. Wollstein G, Garway-Heath DF, Hitchings RA. Identification of early glaucoma cases with the scanning laser ophthalmoscope. *Ophthalmology*. 1998;105:1557–1563.
21. Garway-Heath DF, Poinoosawmy D, Wollstein G, et al. Inter- and intraobserver variation in the analysis of optic disc images: comparison of the Heidelberg retina tomograph and computer assisted planimetry. *Br J Ophthalmol*. 1999;83:664–669.
22. Jonas JB, Budde WM, Panda-Jonas S. Ophthalmoscopic evaluation of the optic nerve head. *Surv Ophthalmol*. 1999;43:293–320.
23. Jonas JB. Optic disk size correlated with refractive error. *Am J Ophthalmol*. 2005;139:346–348.
24. Bowd C, Chan K, Zangwill LM, et al. Comparing neural networks and linear discriminant functions for glaucoma detection using confocal scanning laser ophthalmoscopy of the optic disc. *Invest Ophthalmol Vis Sci*. 2002;43:3444–3454.
25. Hothorn T, Lausen B. Bagging tree classifiers for laser scanning images: a data- and simulation-based strategy. *Artif Intell Med*. 2003;27:65–79.
26. Medeiros FA, Zangwill LM, Bowd C, Sample PA, Weinreb RN. Influence of disease severity and optic disc size on the diagnostic performance of imaging instruments in glaucoma. *Invest Ophthalmol Vis Sci*. 2006;47:1008–1015.
27. Jonas JB, Fernandez MC, Naumann GO. Glaucomatous optic nerve atrophy in small discs with low cup-to-disc ratios. *Ophthalmology*. 1990;97:1211–1215.
28. Heijl A, Molder H. Optic disc diameter influences the ability to detect glaucomatous disc damage. *Acta Ophthalmol*. 1993;71:122–129.
29. Healey PR, Mitchell P. Optic disk size in open-angle glaucoma: the Blue Mountains Eye Study. *Am J Ophthalmol*. 1999;128:515–517.
30. Burgoyne CF, Downs JC, Bellezza AJ, Suh JK, Hart RT. The optic nerve head as a biomechanical structure: a new paradigm for understanding the role of IOP-related stress and strain in the pathophysiology of glaucomatous optic nerve head damage. *Prog Retin Eye Res*. 2005;24:39–73.
31. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ*. 2003;326:41–44.
32. Shunmugam M, Azuara-Blanco A. The quality of reporting of diagnostic accuracy studies in glaucoma using the Heidelberg retina tomograph. *Invest Ophthalmol Vis Sci*. 2006;47:2317–2323.
33. Medeiros FA, Sample PA, Zangwill LM, Liebmann JM, Girkin CA, Weinreb RN. A statistical approach to the evaluation of covariate effects on the receiver operating characteristic curves of diagnostic tests in glaucoma. *Invest Ophthalmol Vis Sci*. 2006;47:2520–2527.