Understanding Disparities Among Diagnostic Technologies in Glaucoma

Carlos Gustavo V. De Moraes, MD; Jeffrey M. Liebmann, MD; Robert Ritch, MD; Donald C. Hood, PhD

**Objective:** To investigate causes of disagreement among 3 glaucoma diagnostic techniques: standard automated achromatic perimetry (SAP), the multifocal visual evoked potential technique (mfVEP), and optical coherence tomography (OCT).

**Methods:** In a prospective cross-sectional study, 138 eyes of 69 patients with glaucomatous optic neuropathy were tested using SAP, the mfVEP, and OCT. Eyes with the worse and better mean deviations (MDs) were analyzed separately. If the results of 2 tests were consistent for the presence of an abnormality in the same topographic site, that abnormality was considered a true glaucoma defect. If a third test missed that abnormality (false-negative result), the reasons for disparity were investigated.

**Results:** Eyes with worse MD (mean [SD], −6.8 [8.0] dB) had better agreements among tests than did eyes with better MD (−2.5 [3.5] dB, \( P < .01 \)). For the 94 of 138 hemifields with abnormalities of the more advanced eyes, the 3 tests were consistent in showing the same hemifield abnormality in 50 hemifields (53%), and at least 2 tests were abnormal in 65 of the 94 hemifields (69%). The potential explanations for the false-negative results fell into 2 general categories: inherent limitations of each technique to detect distinct features of glaucoma and individual variability and the distribution of normative values used to define statistically significant abnormalities.

**Conclusions:** All the cases of disparity could be explained by known limitations of each technique and interindividual variability, suggesting that the agreement among diagnostic tests may be better than summary statistics suggest and that disagreements between tests do not indicate discordance in the structure-function relationship.

Arch Ophthalmol. 2012;130(7):833-841

Different techniques, each focusing on a particular feature of glaucomatous damage, have been developed to help diagnose and observe patients with glaucoma. However, these techniques are not always consistent in determining the extent and location of injury. Such disparity may cause confusion among physicians and delay treatment for patients who could benefit from early diagnosis and intervention.

The assessment of structure and function relationships in glaucoma and of differences in the performance of diagnostic tests has been the focus of extensive research for more than 2 decades. In general, previous studies have focused on the degree of agreement among various glaucoma diagnostic tests, but little is known about possible causes that underlie their disagreement. A better understanding of these causes could help physicians choose the most appropriate combination of tests for detecting damage and could help them decide which test to rely on in cases of disparity. Progress in this area has been hindered by the lack of a gold standard to define glaucoma damage.

To circumvent this problem, this study is based on the assumption that if 2 different tests agree regarding the presence and topographic location of an abnormality, it is likely that there is true glaucoma damage. Under this assumption, if the result of a third test is normal, we conclude that it is more likely to be a false negative. This gives us a method for evaluating the bases for disagreement among tests.
Herein, we attempted to better understand the disagreement among 3 different types of diagnostic examinations: standard automated achromatic perimetry (SAP), the multifocal visual evoked potential technique (mfVEP), and optical coherence tomography (OCT) of the retinal nerve fiber layer (RNFL).

**METHODS**

This prospective study was approved by The New York Eye and Ear Infirmary and Columbia University institutional review boards. Written informed consent was obtained from all the participants, and the study followed the tenets of the Declaration of Helsinki. The patients enrolled were referred between August 1, 2007, and July 31, 2008, by 3 glaucoma specialists.

Before testing, patients underwent a complete ophthalmic evaluation and optic disc stereophotography. Eyes with glaucomatous optic neuropathy, with or without SAP defects, were referred for testing. The reason why an abnormal SAP finding was not required to define glaucoma was that we were testing the agreement among tests, and one of these tests was SAP; and the use of this (or any other) test to define the presence of damage would have incurred selection bias. Glaucomatous optic neuropathy was defined as a vertical cup-disc ratio greater than 0.6, asymmetry of the cup-disc ratio greater than 0.2 between eyes, and the presence of localized RNFL or neuroretinal rim defects in the absence of any other retinal or neuro-ophthalmic abnormalities that could explain such findings.

All the patients were tested using the mfVEP and OCT on the same day, whereas SAP tests were all performed within 3 months. Details of each technique are described as follows.

**STANDARD AUTOMATED ACHROMATIC PERIMETRY**

SAP was performed using the Humphrey Visual Field Analyzer II (24-2 Swedish Interactive Threshold Algorithm standard strategy) (Carl Zeiss Meditec, Inc). All the visual field tests had reliability indexes of less than 30% fixation losses, false-positive responses, or false-negative responses. All the eyes had visual acuities of at least 20/40 and refractive errors less than 6.0 diopters (D) sphere or 2.0 D cylinder.

**MULTIFOCAL VISUAL EVOKED POTENTIAL TECHNIQUE**

The mfVEP was performed using a software program (VERIS 4.3; Electro-Diagnostic Imaging Inc). Details of the technique have been described elsewhere.3 Examples of mfVEP responses, together with probability plots obtained from a patient with open-angle glaucoma, are shown in Figure 1. The points in these plots are positioned in the center of each of the 60 sectors of the display. A colored square indicates that the mfVEP response was statistically significant at either the 5% (>1.96 SD, desaturated color) or 1% (>2.58 SD, saturated color) level compared with normal values.14 On the interocular plot, the color indicates whether it was the response of the left (red) or right (blue) eye that was significantly smaller than that of the fellow eye. On the monocular plot, the color indicates that the signal-noise ratio values of the left (red) or right (blue) eye for that location were significantly smaller.

**OPTICAL COHERENCE TOMOGRAPHY**

The thickness of the parapapillary RNFL was measured using OCT (Stratus; Carl Zeiss Meditec, Inc) using version 4.0 software and the fast scan protocol. During a single recording, 3 scans are made around a ring 3.4 mm in diameter with a spatial resolution of 236 points and then are averaged. The commercial software provides various summary statistics of the resulting RNFL scan and a comparison of these statistics with a normative database. The primary measure used was the average thickness of each of the 12 clock-hour segments. The software compares these values with an age-matched normative database and indicates whether they fall in the top 5% (coded white), top 95% (green), bottom 5% (yellow), or bottom 1% (red).

**DEFINITION OF ABNORMAL HEMIFIELDS**

Assessment of the topographic agreement among tests was based on hemifield damage. Because how we define an abnormal hemifield may affect the detection rates and agreement among tests, we chose to use definitions widely reported in the literature, including from our previous work.13-17 For the SAP and mfVEP tests, cluster criteria were used as previously defined.13 In particular, for SAP (total deviation plot), a hemifield was defined as abnormal if a cluster had 2 or more contiguous points at P < .01 or 3 or more contiguous points at P < .05 with at least 1 point at P < .01.18 To avoid rim artifacts, the cluster could contain no more than 1 point from the outer ring of the 24-2 SAP points. We chose to use the total deviation plot because it shows the comparison of threshold sensitivities between individuals and age-matched controls,19 similar to the analysis provided by OCT13 and the mfVEP.3 Adjustment of the depth of the island of vision (a procedure that defines the pattern deviation plot)20 would not allow equal diagnostic conditions for the 3 tests. In addition, Artes et al20 recently demonstrated that total deviation plots may be more informative for determining early SAP loss in glaucoma. The mfVEP finding from a hemifield was considered abnormal if it contained a cluster on the monocular or interocular plot of 2 or more contiguous points at P < .01 or 3 or more contiguous points at P < .05 with at least 1 point at P < .01. For OCT, the clock-hour plot on the fast RNFL report was used. The OCT finding for a hemifield was considered abnormal if there was 1 red (1% level) or 2 yellow (5% level) clock hours anywhere in a hemisphere.15
To assess the influence of disease severity on the present results, we divided fellow eyes into 2 groups based on the level of asymmetry of the SAP mean deviation. We called the eye with the more negative mean deviation the “worse eye” and the fellow eye the “better eye.”

We assumed that if the results of 2 tests were consistent for the presence of an abnormality (as defined previously herein for each technique) in the same hemifield (superior or inferior), then that abnormality was likely a true glaucoma defect (Figure 2). If a third test missed that abnormality, we attempted to understand the reason(s) why.

RESULTS

One hundred thirty-eight eyes of 69 patients (mean [SD] age, 63.0 [13.7] years; age range, 20-83 years) were enrolled. Thirty-nine patients (57%) were women and 61 (88%) were of European ancestry. Worse and better eyes had mean (SD) mean deviations of –6.8 (8.0) and –2.5 (3.5) dB, respectively (P/H110.01, paired t test).

WORSE EYES

The agreement of abnormal hemifields among the 3 tests is summarized in Figure 3. Ninety-four of the 138 hemifields (68%) tested were abnormal on 1 or more of the 3 tests. Thirty-nine patients (57%) were women and 61 (88%) were of European ancestry. Worse and better eyes had mean (SD) mean deviations of –6.8 (8.0) and –2.5 (3.5) dB, respectively (P < .01, paired t test).

BETTER EYES

The agreement of abnormal hemifields among the 3 tests is summarized in Figure 6. The 3 tests combined showed 66 abnormal hemifields from a total of 138 hemifields (48%). The 3 tests were consistent in 14 of the 66 abnormal hemifields (21%), and at least 2 tests were abnormal in 27 of these 66 hemifields (41%).

Considering the cases in which SAP and OCT findings were both abnormal (55 hemifields), OCT findings were normal in 7 of these hemifields (12%, shown in green in Figure 3). In all 7 cases, the OCT result showed areas elsewhere in the printout where the RNFL was thicker than the normative database (coded white, 5%); 6 of these were in the temporal sector. In 2 cases, a borderline defect (1 yellow clock hour) was present in the corresponding location. In all cases, the RNFL profile showed localized thinning compared with adjacent locations and corresponding areas of fellow eyes (Figure 5).

Figure 2. Venn diagram showing the agreement among 3 hypothetical diagnostic tests (A, B, and C). “W” represents the number of cases where the 3 tests showed abnormality in the same location. “V + W” (orange oval) represents cases where A and B were consistent. This is likely to represent a true defect. However, in “V” cases (red circle), test C did not show any abnormality in that region, which represents a false-negative result.

Figure 3. Venn diagram showing the agreement among the multifocal visual evoked potential technique (mVEP), standard automated achromatic perimetry (SAP), and optical coherence tomography (OCT) in showing abnormal hemifields in eyes with worse mean deviation. The colored circles represent missed cases for the mVEP (lime green), SAP (blue), and OCT (green).
The monocular analysis also did not show significantly reduced amplitudes compared with the normative database, although often the corresponding region showed abnormal points (Figure 7).

Considering the cases in which OCT and mfVEP findings were consistent (15 hemifields), SAP missed 1 abnormal hemifield (7%, shown in blue in Figure 4A). This case also had points with reduced threshold sensitivities corresponding to the same location shown by the other 2 tests but that did not reach statistical significance in the probability plot.

Considering the cases in which SAP and mfVEP findings were consistent (16 hemifields), OCT missed 2 abnormal hemifields (12%, shown in green in Figure 4A). In both cases,
the OCT result showed areas in which the RNFL was thicker than the normative database, 1 of which was in the temporal sector. In the other case, a borderline defect (1 yellow clock hour) was present. Localized RNFL thinning was also observed in both cases when looking at the RNFL profile.

**COMMENT**

We investigated the agreement among 3 diagnostic tests used in a glaucoma practice and evaluated potential causes of disparity among them. We emphasize that it was not the purpose of this study to investigate the performance of each technique or to determine which test is better for glaucoma diagnosis. Rather, we sought to understand potential causes of false-negative results for each technique to help physicians make better use of the technology with which they are most familiar.

For the mfVEP, failure to detect damage shown by OCT and SAP was due to previously described limitations of the interocular and monocular analyses. First, the interocular analysis has a substantially higher detection rate than does the monocular analysis. Second, the mfVEP will miss damage in regions outside the test pattern and in regions near the periphery of the test region owing to the sparse sampling in these regions. Third, abnormal points were often seen in the...
region of abnormality on SAP and OCT, but these points did not meet the cluster criteria. In the cases in which SAP missed defects shown by OCT and the mfVEP (6% for worse eyes and 7% for better eyes), a nonsignificant reduction in threshold sensitivity was present in the corresponding topographic site shown by the other tests. The values in the total deviation plot were smaller than those in age-matched controls but did not deviate enough to be flagged as statistically significant. One possible explanation could be that these eyes started with abnormally high baseline sensitivities so that the same amount of loss that could have been enough to reach statistical significance in other patients was not enough in this specific case.\(^21,22\) This hypothesis is also supported by the finding of abnormal points in the pattern deviation plot in 2 of 3 cases, as this analysis adjusts the depth of the island of vision and may show more subtle losses in these cases. In addition, the fact that all 3 cases missed were in the superior hemifield suggests that higher population variability in the superior field, likely to be related to droopy eyelids and ptosis, for example, increases the confidence limits and makes it more difficult to achieve statistical significance at the 5% level.

Finally, in the cases in which OCT did not detect abnormalities shown by the mfVEP and SAP (12% in both groups), localized thinning that did not reach the cluster definition of abnormality was often present. Similar to the explanation previously herein for misses on SAP, perhaps these eyes started with a thicker baseline RNFL so that a greater amount of nerve fiber loss would be required before reaching significance compared with the normative database.\(^21,22\) Evidence for this comes from the finding that most of these eyes showed a relatively thick RNFL in the temporal sector, a region that shows minimal changes until the late stage of glaucoma.\(^24\) It is possible that a relatively thick superior and inferior RNFL was present before disease; hence, explaining why the RNFL thinning did not reach statistical significance. Also, small areas of localized thinning shown on the OCT RNFL thickness map could be missed.

One limitation of this study was that although we avoided using SAP results to define glaucoma, we used clinical evaluation of the optic disc to define glaucomatous optic neuropathy. Although we did not use disc photography as one of the diagnostic tests for comparisons, it is possible that the performance of a structural test (OCT) may have been overestimated. Therefore, a selection bias may have occurred. Of course, a bias is possible as long as any type of structural or functional test is used to define glaucoma. Also, the use of less stringent reliability criteria for SAP results (<30%) could have affected the rates of agreement between this technique and others. This underscores the concept that the agreement among techniques is influenced by examination quality (“noise” or “variability”). The potential explanations for the false-negative results or misses fell into 2 general categories. First, all the techniques are aimed at detecting distinct features of glaucoma and have inherent limitations. For example, it is known that the mfVEP detects more central defects than does SAP because it tests more points in the central field than does a 24-2 test.\(^5\) On the other hand, it misses points outside the central 10° owing to sparser sampling. Similarly, the 24-2 SAP may miss RNFL defects depicted by OCT in the nasal parapapillary region given the poor sampling of the visual field temporal to the blind spot. To avoid confusion over contradictory test results, the physician needs to be aware of these inherent limitations of a test.

Second, these techniques define abnormality based on the normal distribution of values from controls. Normative databases for different technologies may not reflect the populations under investigation, as ethnicity and age, for example, are not always considered in these comparisons. The ability of a given technique to detect statistically significant abnormalities will depend on where a given patient was situated in the normal distribution before glaucoma started to progress. For example, a patient whose RNFL thickness before glaucoma developed was in the upper boundaries of normality (ie, above average) but whose visual sensitivities were in the lower boundaries (ie, below average) will most likely have an abnormality detected using a test that measures RNFL thickness rather than using a functional test. This point was made and illustrated by Hood and Kardon.\(^21,22\) Consistent with this hypothesis, prospective longitudinal studies have shown that eyes with normal SAP\(^25\) and OCT\(^26\) results at baseline but whose indexes were closer to the lower boundaries of normality are more likely to be diagnosed as having glaucoma in the long term using these same techniques.

How one defines abnormality using a given technology can largely influence its sensitivity and specificity and consequently the rates of agreement with other technologies. Less stringent criteria, requiring condensed areas of abnormality or with smaller depth, are more likely to reveal greater sensitivities at the expense of worse specificity. More stringent criteria, conversely, lead to opposite findings. Consequently, the agreement between 2 tests tends to increase when loose criteria are used in both cases.

**Figure 6.** Venn diagram showing the agreement among the multifocal visual evoked potential technique (mfVEP), standard automated achromatic perimetry (SAP), and optical coherence tomography (OCT) in showing abnormal hemifields in eyes with better mean deviation. The colored circles represent missed cases for the mfVEP (lime green), SAP (blue), and OCT (green).
and to decrease when stringent definitions are used. Physicians should look at the strictness of the criteria used to define abnormalities when facing potentially discordant cases.

One potential way to circumvent problems is to use continuous probability scales. For example, Wall et al\(^2\) developed a continuous probability scale to display visual field damage and showed that contiguous defects following a retinotopic pattern were more prevalent and larger using this method than using the standard percentiles provided by the total deviation plot of StatPac software (Carl Zeiss Meditec, Inc). A second approach to this problem is to superimpose the topographic results of the different tests. Hood and Raza\(^2\) showed that

Figure 7. Example of a patient in whom standard automated achromatic perimetry and optical coherence tomography showed a consistent defect but the multifocal visual evoked potential technique (mfVEP) results were classified as normal. A, Superior visual field defect and corresponding inferior nerve fiber layer thinning (red arrows). B, The mfVEP result showing an extensive and deep defect in the left eye that hindered identification of the defect in the right eye. Note that there were abnormal points in the monocular analysis that did not meet the cluster definition of abnormality (red arrows).
borderline SAP results (Figure 4) can fall in the significantly abnormal region of the OCT results if properly displaced. Under these conditions, no mismatch occurs among tests, although technically the SAP result was classified as normal.

Finally, one way to improve the agreement among tests would be to compare measurements in the same individual at different time points. As an alternative, methods that use intereye or interhemifield comparisons may increase detection rates. This is already true for the OCT (comparison of the RNFL profile between eyes)28 and mfVEP (interocular analysis) printouts.29 For example, the case illustrated in Figure 5, despite displaying a thicker RNFL compared with the normative database, had unquestionable localized RNFL loss in the intereye analysis (black arrow in the RNFL profile).

In summary, we found that despite overall good agreement among the mfVEP, OCT, and SAP tests in eyes with glaucomatous optic neuropathy, the results were inconsistent regarding the presence and topographic location of damage. By examining the cases in which 1 test missed a defect confirmed by the other 2, we identified likely causes of disagreement among tests that were related primarily to the characteristics of the individual diagnostic testing modalities. No evidence of a disparity was noted in the testing results that could not be explained using our current understanding of the clinical pathogenesis of glaucoma. There was no evidence of a significant disparity in testing outcomes in eyes with glaucomatous damage as confirmed by at least 2 testing modalities. This finding suggests that the agreement among diagnostic tests may be better than summary statistics suggest and that disagreements among tests do not indicate discordance in the structure-function relationship. By better understanding the limitations of a particular test, the physician should be able to choose the test results on which to depend on when confronted by conflicting results for a particular patient.

Submitted for Publication: November 11, 2011; final revision received February 3, 2012; accepted February 27, 2012.

Correspondence: Carlos Gustavo V. De Moraes, MD, Department of Ophthalmology, The New York Eye and Ear Infirmary, 310 E 14th St, New York, NY 10003 (demoraesmd@gmail.com).

Financial Disclosure: None reported.

Funding/Support: This study was supported in part by grants EY09076 and EY02115 from the National Institutes of Health and by the HRH Prince Ahmed Al-Saud Research Fund of the New York Glaucoma Research Institute. Dr De Moraes is the Edith C. Blum Foundation Research Scientist.

Online-Only Material: This article is featured in the Archives Journal Club. Go to http://www.archophthalmol.com to download teaching PowerPoint slides.

REFERENCES


