Article

Signal-to-Noise Ratios for Structural and Functional Tests in Glaucoma

Stuart K. Gardiner[✉], Brad Fortune[✉], and Shaban Demirel[✉]

Devers Eye Institute, Legacy Health, Portland, OR

Correspondence: Stuart K. Gardiner, Devers Eye Institute, Legacy Health, 1225 NE 2nd Ave, Portland, OR 97232, USA. e-mail: sgardiner@deverseye.org

Received: 26 June 2013

Accepted: 8 September 2013

Published: 29 October 2013

Keywords: statistics, glaucoma, perimetry, optical coherence tomography

Citation: Gardiner SK, Fortune B, Demirel S. Signal-to-noise ratios for structural and functional tests in glaucoma. *Trans Vis Sci Tech.* 2013; 2(6):3, <http://tvstjournal.org/doi/full/10.1167/tvst.2.6.3>, doi:10.1167/tvst.2.6.3

Purpose: Standard automated perimetry (SAP) demonstrates high variability. Structural tests such as optical coherence tomography (OCT) may be more repeatable. However, comparisons of their ability to detect glaucomatous change are challenging due to different units and dynamic ranges. This study demonstrates a signal-to-noise analysis to perform comparisons within a common framework.

Methods: Longitudinal data were used from 226 eyes of 130 subjects with nonendstage glaucoma (mean deviation [MD] from -19.50 to 2.89 dB). Subjects were tested twice a year for a total of at least six visits. For each eye, MD from SAP and average retinal nerve fiber layer thickness (RNFLT) from OCT were regressed linearly against time. 'Signal' was defined as the rate of change over time, while 'noise' was defined as the SD of residuals from this trend. Individual longitudinal signal-to-noise ratios were calculated. A summary quantification was also calculated, using the 10th percentile of these rates within the cohort as signal and the SD of residuals pooled across all eyes as noise.

Results: Individual signal-to-noise ratios were significantly better for OCT RNFLT than for SAP MD ($P < 0.0001$). The summary quantification of signal-to-noise ratio was better for OCT RNFLT (-1.35 y^{-1}) than for SAP MD (-0.74 y^{-1}).

Conclusions: RNFLT measured by OCT had a better longitudinal signal-to-noise ratio than MD from SAP.

Translational Relevance: The longitudinal signal-to-noise ratio provides a means to perform a fair comparison between different techniques, which is robust to differences in scale and measurement units. Longitudinal studies in glaucoma should consider reporting signal-to-noise ratios to facilitate interpretation and comparison of results.

Introduction

It has often been noted that there are no gold standards when it comes to the assessment of eyes with glaucoma. Every test has intrinsic variability, and so it is impossible to say with any certainty how accurate any given test is either for diagnosing glaucoma or for monitoring disease progression. Clinician assessment of stereophotographs exhibits only moderate interobserver reproducibility, with kappa variously reported as being $\kappa = 0.55\text{--}0.78$ ¹ or $\kappa = 0.45\text{--}0.74$,² depending on the task and study population. In one study, agreement in detecting progression using stereophotographs was as low as $\kappa = 0.2$, and in 40% of cases the photo judged as showing increased damage had in fact been taken at

the start of the study.³ This lack of a gold standard makes objective assessment of different clinical testing techniques challenging. For example, structural testing using Heidelberg Retina Tomography and functional testing using Standard Automated Perimetry (SAP) have been reported as having low agreement in classifying eyes as healthy or glaucomatous, with $\kappa = 0.39$,⁴ and it is often impossible to say which of the devices provides the correct classification for any given eye. Indeed, both tests may be accurate, but reflect different aspects of the pathophysiology or have different dynamic ranges. Alternatively, they could be measuring the same aspects of the pathophysiology, but with high variability masking the underlying level of agreement.⁵

Due to these various issues, an objective and self-

contained framework is needed to assess the efficacy of a test. One formulation that can be used is to assess the signal-to-noise ratio of a test. A test with zero variability is of no use if its results remain constant as the disease progresses, whereas a test whose variability covers its entire dynamic range is also ineffective for clinical purposes. A measure of progression that is to be useful for monitoring patients should have a high signal-to-noise ratio, allowing more accurate assessment of the rate of change over time, and a greater number of discriminable steps as the disease progresses.^{6,7}

The problem with applying signal-to-noise analysis to glaucomatous progression is in defining ‘signal’ and ‘noise’, in the absence of a gold standard for progression. Artes et al. described a form of signal-to-noise analysis for identification of glaucomatous visual field defects.⁸ Their definition of signal was based on the interhemifield difference between total deviation values averaged within glaucoma hemifield test zones, and noise was based on the SD of the permutation distribution of these values within a repeated-testing dataset. They hypothesized that “gradients in space can be used as a first approximation for change over time,”⁸ and, hence, that signal-to-noise ratios derived from longitudinal data would be similar to those derived using their technique. However, they acknowledged in response to a subsequent letter to the editor that there are limitations with using this cross-sectional technique to predict longitudinal signal-to-noise ratios.⁹

In this paper, we used longitudinal data to derive estimates of both rates of change and variability, to derive a longitudinal signal-to-noise ratio. Signal is defined as the rate of change for an eye, whereas noise is defined as the SD of residuals from the trend over time for that eye. Hence, a more negative longitudinal signal-to-noise ratio provides evidence of a better technique for determining the rate of change, since this rate would be less than zero for a progressing eye. The technique is used to compare a functional measure with a structural measure, without requiring a separate gold standard. This allows us to assess our hypothesis that the objective nature of structural testing, compared with the subjective nature of perimetry due to its dependence on subject responses, results in structural testing having a better longitudinal signal-to-noise ratio.

Methods

Data from 445 eyes of 227 subjects with nonend-stage glaucoma (i.e., with areas of remaining visual

function) or high risk ocular hypertension in the ongoing Portland Progression Project were used.¹⁰ Inclusion criteria were a diagnosis of nonendstage glaucoma, suspected glaucoma or high-risk ocular hypertension as determined by their clinician. Due to the lack of a gold standard, such classifications can vary considerably between techniques and between clinicians, and so no distinction was made between groups in this study. Exclusion criteria were a history of ocular surgery (except for uncomplicated cataract removal), other ocular pathologies likely to affect the visual field, or inability to maintain fixation and produce reliable test results. The study adhered to the tenets of the Declaration of Helsinki, and all protocols were approved and monitored by the institutional review board of Legacy Health. All subjects gave informed consent before undergoing testing, after having the risks and benefits of participation explained to them.

Subjects were tested every 6 months with SAP using the Humphrey Field Analyzer II (Carl Zeiss Meditec Inc., Dublin, CA) with the SITA standard algorithm, 24-2 test pattern and standard testing protocols; and with Optical Coherence Tomography (OCT) using the 6° circle scan protocol of the Heidelberg Spectralis spectral-domain OCT (Heidelberg Engineering GmbH, Heidelberg, Germany). Both tests were performed on the same day. Only visits with both reliable SAP visual fields ($\leq 15\%$ false positives, $\leq 30\%$ false negatives and fixation losses) and good quality OCT scans (≥ 15 signal strength) were included. Mean Deviation (MD) was used to summarize the visual field information from SAP. Average peripapillary retinal nerve fiber layer thickness (RNFLT) was assessed from OCT scans. Trained technicians manually corrected the accuracy of the instrument’s native automated layer segmentations when the software algorithm used had obviously erred from the inner and outer borders of the RNFL to an adjacent layer (such as a refractive element in the vitreous instead of the internal limiting membrane, or to the inner plexiform layer instead of the outer border of the RNFL). To be included in the analysis, a minimum of six eligible visits (i.e., spanning approximately 2.5 y or more) was required for each eye, so that a reliable measure of the rate of change could be obtained.

Mean deviation and RNFLT were regressed against time using ordinary least squares linear regression. Mean deviation is age-corrected, and so the rate of change of MD over time was taken from the regression and used as the measure of functional signal for that

eye. Average RNFLT is not age-corrected, and declines at a rate of $-0.075 \mu\text{m}/\text{y}$ for healthy eyes (taken from the normative database used by the Spectralis software, per personal communication). Therefore, the rate of change of RNFLT plus $0.075 \mu\text{m}/\text{y}$ was used as the measure of structural signal for that eye. Note that while the P values associated with ordinary least squares regression are generally inappropriate for longitudinal data due to potential autocorrelation between residuals on consecutive visits,¹¹ estimates of the rate of change remain valid.

For both functional and structural tests, residuals from the trend over time were calculated for each eye, and then the SD of these residuals was used as the measure of noise for that eye. This gives a measure of variability that is unaffected by the possibility of progression over time for that eye (on the assumption that such progression is linear within the observation period). The resulting longitudinal signal-to-noise ratios were calculated for each eye, for both MD and RNFLT, and histograms constructed. As a formal comparison, a Wilcoxon matched-pairs test was performed on the absolute value of the normalized rates.

A summary quantification of the signal-to-noise ratios was also generated. Signal was defined as the 10th percentile of the rates of annual change within the cohort (Slope_{10}). This signal quantification was chosen instead of using the average rate of change in order to reduce the effect of stable eyes, which constitute a large proportion of the dataset, and, hence, to ensure that signal is representative of a more typical rate of change in a progressing eye. It also reduces the effect of outliers (which could be caused by noise rather than being reliable measures of disease progression) when compared with using the worst rate of change in the dataset. Noise was defined as the SD of residuals from the individual-eye trends over time, pooled over all eyes in the dataset. As a secondary analysis, these summary quantifications were repeated using only those sequences where the MD was never worse than -3 dB , to reduce the confounding effect of variability increasing as visual field damage worsens in SAP.

Results

Two hundred and twenty-six eyes of 130 subjects had a series of at least six visits for which eligible, reliable results were available for both SAP and OCT. Table 1 shows the characteristics of the cohort, restricted to these eligible series. At baseline, 32 eyes (14%) had MD outside normal limits ($P < 0.05$), of which five were worse than -10 dB . At the end of

Table 1. Characteristics of the Dataset, Including Results From SAP and OCT at the Start and End of Each Series

	Mean	SD	Range
Series length (tests)	6.5	0.6	6–8
Age, y	65.5	9.7	37.6–85.7
Baseline MD from SAP, dB	−0.50	2.60	−15.46–2.50
Baseline pattern SD from SAP, dB	2.19	2.23	0.91–15.64
Baseline RNFLT from OCT, μm	90.3	14.3	37.5–126.2
Final MD from SAP, dB	−0.72	2.96	−19.50–2.89
Final pattern SD from SAP, dB	2.38	2.30	0.87–13.76
Final RNFLT from OCT, μm	87.5	15.0	35.8–125.8

each subject's series, 38 eyes (17%) has MD outside normal limits, still with five worse than -10 dB . The mean rate of change of MD was $-0.04 \text{ dB}/\text{y}$, and the mean rate of change for RNFLT was $-0.94 \mu\text{m}/\text{y}$ (or $-0.83 \mu\text{m}/\text{y}$ after the age-correction was applied). The Pearson correlation between rates of structural and functional change within an eye was 0.21 ($P = 0.002$, t -test).

Figure 1 shows histograms of the individual longitudinal signal-to-noise ratios (i.e., the rate of annual change divided by the SD of residuals for each eye individually) for SAP MD and OCT RNFLT. The mean longitudinal signal-to-noise ratios were -0.045 y^{-1} for SAP MD and -0.601 y^{-1} for OCT RNFLT. It can be seen that there is a greater range of values for OCT than SAP, as seen for example in the greater proportion of eyes with a rate of change worse than -1.5 times the noise. The signal (rate of change) when expressed as a multiple of the noise tends to be greater for RNFLT. Indeed, the values were significantly worse (less negative) for SAP than OCT, with P less than 0.0001.

As noted above, only 32 of the 226 eyes had MD outside normal limits at baseline. For these subjects, the mean longitudinal signal-to-noise ratios were -0.303 y^{-1} for SAP MD and -0.810 y^{-1} for OCT RNFLT. For the subjects within normal limits at baseline, who might be expected to progress more slowly and, hence, have smaller signal, the mean longitudinal signal-to-noise ratios were -0.003 y^{-1} for SAP MD and -0.567 y^{-1} for OCT RNFLT.

The summary quantification of noise, measured as

Figure 1. Histograms of the longitudinal signal-to-noise ratio (rate of change divided by SD of residuals from the trend line), for MD from SAP and for age-corrected RNFLT from OCT.

the SD of residuals from the trend over time pooled over all eyes, was 0.58 dB for MD. Note that this is lower than estimates of pointwise variability, due to the averaging inherent in the calculation of MD. Indeed, the SD of pointwise residuals averaged 1.5 dB, which gives a 95% confidence interval (CI) for test–retest differences at an individual visual field location of ± 9.0 dB, similar to that reported in the literature. For example, Chauhan et al. reported that the 90% CI for test–retest was 9.6 dB for a baseline deviation of 0 dB.¹² The summary quantification of noise for RNFLT was 1.76 μm , which is also similar to that reported in the literature. Mwanza et al. have reported that the test–retest SD for average RNFLT was 1.67 μm in people with glaucoma.¹³ The summary quantification of signal, defined using the 10th percentile of observed rates of change as described in the Methods section, was -0.43 dB/y for SAP MD and -2.37 $\mu\text{m}/\text{y}$ for OCT RNFLT. Together, these give summary quantifications of longitudinal signal-to-noise ratio as -0.74 y^{-1} for SAP MD, and -1.35 y^{-1} for age-corrected OCT average RNFLT.

Variability in SAP results is known to increase as damage worsens.^{12,14} In an attempt to alleviate this potential confound, the analysis was repeated using only the 207 eyes of 124 subjects whose MD never went below -3 dB at any point in their sequence of visual field testing. In this subset, the noise was 0.47 dB for MD, and 1.65 μm for RNFLT. Using the same summary quantification of signal as before (based on the entire dataset, since the eyes with the most signal would otherwise be excluded from the analysis), and

these lower measures of noise, the longitudinal signal-to-noise ratios were -0.90 y^{-1} for SAP MD, and -1.48 y^{-1} for age-corrected OCT RNFLT.

Discussion

An ideal metric for monitoring glaucoma would be accurate (directly representative of the true disease status), precise (taking one of many different possible values rather than just a small number of possibilities), and repeatable (low test–retest variability). Unfortunately, it is currently impossible to assess test accuracy, since there is no gold standard available for describing the true disease status or its rate of change in any given eye. An assay of the exact number of remaining retinal ganglion cells may one day provide such a gold standard, but even that may not be sufficient if an unknown number of those cells are present but dysfunctional. Precision is easier to assess. For example, frequency-doubling technology (FDT) perimetry using the Matrix perimeter (Carl Zeiss Meditec Inc.) has only 15 possible final threshold estimates available to describe the functional status at any given visual field location, whereas OCT reports RNFLT on a continuous scale.

Repeatability has most frequently been assessed using test–retest variability,^{7,12,15,16} with as short a period of time as possible between the repeated testing in order to minimize the possibility of true change having occurred during that time period. In this study, repeatability is assessed using the residuals from the trend over time, which is essentially the same

as assessing test–retest variability using tests that are up to several years apart, but accounting for possible change that may have taken place over that period. This enables a larger dataset to be used without requiring impractical amounts of short-term data collection, improving estimates of the amount of variability present. The estimates of noise derived here are similar to those from the literature using short duration test–retest studies.^{12,13}

Perhaps a larger advantage of using longitudinal data, rather than short duration test–retest data, is that it allows the amount of change over time to be used as a measure of signal. A test whose outcome displays very little change over time or has a narrow dynamic range is of limited use for a progressive neuropathy such as glaucoma, even if the variability is low. As an extreme example, consider multiplying all results from a device by zero; there would be zero variability, but there would also be no remaining signal. Detection of change, and perhaps as importantly the rate of change, is one of the two main priorities for a clinical test used to monitor a chronic disease such as glaucoma. The other priority is its ability to discriminate between healthy and diseased eyes. The cross-sectional technique proposed by Artes et al.⁸ can be used to assess signal-to-noise ratios for the ability to discriminate healthy from diseased eyes, but the technique demonstrated in the current study may provide a better framework within which to assess signal-to-noise ratios for detection of progression.

Choices made in this study may have had a minor effect on the results, even though such effects were minimized wherever possible. Firstly, RNFLT was assessed from OCT scans after manual refinement of the instrument's automated segmentations, to remove obvious errors and inconsistencies. Such refinements might not be routinely performed in the clinical setting. Secondly, linear fits over time were used for both methodologies, despite the fact that SAP sensitivities on a decibel scale have been reported to decline exponentially whereas RNFLT declines linearly,^{11,17–19} and so the residuals for MD may be larger (giving a larger estimate of 'noise') than they would be with a different longitudinal model. The linear fit was retained principally because it provides a more intuitive measure of signal, since in an exponential longitudinal model the rate of change is not constant. A linear model appears to be adequate for examining changes in MD over time so long as the series length is not excessive.²⁰ Thirdly, it is known that variability

increases with damage for SAP,^{12,14,21} and so the resultant estimate of noise may be an overestimate compared with the true amount of noise that is present at the earliest stages of the disease. Fourthly, unreliable test results were excluded from the analysis. Eight and half percent of SAP visual fields were excluded based on excessive false positives, false negatives, or fixation losses, whereas only a single OCT scan was excluded based on an unacceptable quality score. This, together with potential learning effects (which should not be present in this dataset due to the subjects' prior experience with SAP), could make the signal-to-noise ratio for SAP worse in practice than reported here. Finally, the cohort consisted primarily of subjects with relatively early disease, and it has been suggested that structural testing can detect damage sooner than functional testing. It is possible that the signal-to-noise ratio for SAP could change later in the disease process.

It is perhaps surprising that the signal-to-noise advantage of OCT RNFLT over SAP MD is not greater. OCT is an objective test, in as much as it does not rely on inherently variable psychophysical responses, which are probabilistic due to the nature of the psychometric function. It has also been suggested that RNFLT from OCT may provide a better metric for change than SAP MD in early stages of the disease,²² which is the dominant severity in this cohort, as evidenced by the average MD of -0.72 dB. It can be concluded that the variability in OCT RNFLT is still considerable, and further advances are needed in both image quality and the segmentation of anatomical layers before its full potential is realized.

It is imperative when comparing different testing modalities, not only using different units but measuring completely different targets, to have a common framework within which they may be compared. Signal-to-noise ratios may provide such an ideal conceptual framework, and could usefully be reported by studies of methods for assessing the rate of change in glaucoma. In the current study, based on estimated rates of change, RNFLT as measured by OCT had a better longitudinal signal-to-noise ratio than MD from SAP.

Acknowledgments

Supported by grants from the National Institutes of Health R01-EY19674 (SD), and The Legacy Good Samaritan Foundation, Portland, OR.

Disclosure: **S.K. Gardiner**, None; **B. Fortune**, None; **S. Demirel**, None

References

1. 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.
2. Reus NJ, de Graaf M, Lemij HG. Accuracy of GDx VCC, HRT I, and clinical assessment of stereoscopic optic nerve head photographs for diagnosing glaucoma. *Br J Ophthalmol*. 2007;91:313–318.
3. Jampel HD, Friedman D, Quigley H, et al. Agreement among glaucoma specialists in assessing progressive disc changes from photographs in open-angle glaucoma patients. *Am J Ophthalmol*. 2009;147:39–44.e31.
4. Ng D, Zangwill LM, Racette L, et al. Agreement and repeatability for standard automated perimetry and confocal scanning laser ophthalmoscopy in the diagnostic innovations in glaucoma study. *Am J Ophthalmol*. 2006;142:381–386.
5. Gardiner S, Johnson C, Demirel S. The effect of test variability on the structure–function relationship in early glaucoma. *Graefes Arch Clin Exp Ophthalmol*. 2012;250:1851–1861.
6. Wall M, Woodward KR, Doyle CK, Zamba GJ. The effective dynamic ranges of standard automated perimetry sizes III and V and motion and matrix perimetry. *Arch Ophthalmol*. 2010;128:570–576.
7. Jampel HD, Vitale S, Ding Y, et al. Test-retest variability in structural and functional parameters of glaucoma damage in the glaucoma imaging longitudinal study. *J Glaucoma*. 2006;15:152–157.
8. Artes PH, Chauhan BC. Signal/noise analysis to compare tests for measuring visual field loss and its progression. *Invest Ophthalmol Vis Sci*. 2009;50:4700–4708.
9. Artes PH, Chauhan BC. Author response: signal/noise ratios to compare tests for measuring visual field progression. *Invest Ophthalmol Vis Sci*. 2010;51:6893–6894.
10. Gardiner SK, Johnson CA, Demirel S. Factors predicting the rate of functional progression in early and suspected glaucoma. *Invest Ophthalmol Vis Sci*. 2012;53:3598–3604.
11. Pathak M, Demirel S, Gardiner SK. Nonlinear, multilevel mixed-effects approach for modeling longitudinal standard automated perimetry data in glaucoma. *Invest Ophthalmol Vis Sci*. 2013;54:5505–5513.
12. Chauhan B, Johnson C. Test-retest variability of frequency-doubling perimetry and conventional perimetry in glaucoma patients and normal subjects. *Invest Ophthalmol Vis Sci*. 1999;40:648–656.
13. Mwanza J-C, Chang RT, Budenz DL, et al. Reproducibility of peripapillary retinal nerve fiber layer thickness and optic nerve head parameters measured with Cirrus HD-OCT in glaucomatous eyes. *Invest Ophthalmol Vis Sci*. 2010;51:5724–5730.
14. Henson D, Chaudry S, Artes P, Faragher E, Ansons A. Response variability in the visual field: comparison of optic neuritis, glaucoma, ocular hypertension, and normal eyes. *Invest Ophthalmol Vis Sci*. 2000;41:417–421.
15. Piltz J, Starita R. Test-retest variability in glaucomatous visual fields. *Am J Ophthalmol*. 1990;109:109–110.
16. Bjerre A, Grigg J, Parry N, Henson D. Test-retest variability of multifocal visual evoked potential and SITA standard perimetry in glaucoma. *Invest Ophthalmol Vis Sci*. 2004;45:4035–4040.
17. Hood D, Kardon R. A framework for comparing structural and functional measures of glaucomatous damage. *Prog Retin Eye Res*. 2007;26:688–710.
18. Harwerth R, Quigley H. Visual field defects and retinal ganglion cell losses in patients with glaucoma. *Arch Ophthalmol*. 2006;124:853–859.
19. Garway-Heath D, Caprioli J, Fitzke F, Hitchings R. Scaling the hill of vision: the physiological relationship between light sensitivity and ganglion cell numbers. *Invest Ophthalmol Vis Sci*. 2000;41:1774–1782.
20. Gardiner SK, Demirel S, De Moraes CG, et al. Series length used during trend analysis affects sensitivity to changes in progression rate in the ocular hypertension treatment study. *Invest Ophthalmol Vis Sci*. 2013;54:1252–1259.
21. Chauhan B, House P. Intratest variability in conventional and high-pass resolution perimetry. *Ophthalmology*. 1991;98:79–83.
22. Medeiros FA, Zangwill LM, Bowd C, Mansouri K, Weinreb RN. The structure and function relationship in glaucoma: implications for detection of progression and measurement of rates of change. *Invest Ophthalmol Vis Sci*. 2012;53:6939–6946.