INTRODUCTION

The advent of powerful imaging techniques, such as HRT, GDx, and OCT, with their ability to elucidate the structural details of glaucomatous damage, is changing the practice of glaucoma diagnosis and management. These methods potentially hold the promise of more sensitive and specific determinations of glaucomatous progression than sequential visual field measurements or longitudinal follow-up with ophthalmoscopy or stereo-photography. However, before attributing a difference in a patient’s measurement to progression it is necessary to quantify the variability associated with test-retest variability (i.e. measurement “error”) for a given imaging technology.

Experiments designed to evaluate test-retest variability usually follow the strategy of making consecutive measurements on an individual in a time period too short for glaucomatous progression to have occurred. This may be during the same measuring session, in a separate measuring session on the same day, and/or during a subsequent patient visit scheduled days or weeks following the initial measurement. One might expect that test-retest differences on different days would exceed those made during the same session, and this is a hypothesis that can be explicitly tested. In addition to test-retest variability, studies of this type may also be used to assess the influence of different technicians or different machines.

The aim of this manuscript is to describe, and provide examples of, some of the common methods of analyzing and presenting these types of data: estimation of the test-retest standard deviation, the coefficient of variation, and the intraclass correlation coefficient. These will be contrasted with the Bland-Altman method of assessing agreement. None of these methods is ideally suited to all situations.

THE VARIANCE OF MEASUREMENTS: BETWEEN PATIENTS VERSUS WITHIN PATIENTS (TEST-RETEST)

How does one assess the variability in a sample of patients’ measurements? Calculate the sample variance and then, perhaps, take its square-root to get the standard deviation. If the sample contains a single mean RNFL measurement apiece collected from a group of glaucoma patients with a single Stratus OCT machine and made by a single experienced technician, there may be nothing further to do. However, we recognize that the difference between the measurements of two patients will be comprised of two parts: (1) the actual difference in nerve fiber layer thicknesses of the individuals at the time they were measured and (2) the effects of imprecision of the OCT as a measuring device.

The former reflects what the patient’s normal RNFL thickness was before the onset of glaucomatous disease, the action of that disease over time, and some small additional thinning due to age. This is what the clinician would like to know to effectively manage the patient and monitor for progression. The latter is nuisance variability which, if large, undercuts the usefulness of the measuring device. One may think of the former as “signal” and the latter as “noise”.

As above, the experimental strategy for estimating the magnitude of this noise/nuisance variance is to repeated test patients over a period too short for any true clinical change to occur. The test-retest variance for each patient can then be calculated and pooled (averaged) over all patients to estimate the within patient/test-retest/nuisance variance associated with the error of the measuring method.

Data collected from such experiments are analyzed with variance components analysis (an application of Model II/mixed model ANOVA) and, for a suitably designed clinical experiment can simultaneously provide estimates of variance due to the true patient differences an instrument is designed to detect (between patient variance) plus a variety of “nuisance” variances: within session test-retest variance (repeat measurements on the same day or during the same sitting), between session test-retest variance (repeat measurements on different days or at different sittings), between instrument variance, between technician variance. It is important to keep in mind that these experiments assume that truly no biologic change has occurred between measurements. Intraocular pressure changes in a circadian fashion and if a reproducibility experiment evaluating, say, a new tonometer measured patients twice at different times of day, without taking this effect into account, measurement error would be over-estimated.
Performing a Reproducibility Analysis

The data from a simple reproducibility analysis can be analyzed with a spreadsheet. A very short example gives the flavor of how the calculations proceed, supposing two measurements of mean RNFL each made on the right eyes of five patients and no missing data.

<table>
<thead>
<tr>
<th>Patient/Eye</th>
<th>Mean measurement</th>
<th>RNFL</th>
<th>Variance*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>56.3</td>
<td>50.0</td>
<td>53.2</td>
</tr>
<tr>
<td>2</td>
<td>73.3</td>
<td>69.3</td>
<td>71.3</td>
</tr>
<tr>
<td>3</td>
<td>90.7</td>
<td>96.0</td>
<td>93.4</td>
</tr>
<tr>
<td>4</td>
<td>39.0</td>
<td>43.7</td>
<td>41.4</td>
</tr>
<tr>
<td>5</td>
<td>52.0</td>
<td>54.3</td>
<td>53.2</td>
</tr>
</tbody>
</table>

Mean variance 62.5

Variance 412.6

* The variance of two measurements can be calculated as \( \frac{1}{2} \times (\text{Mean}_1 - \text{Mean}_2)^2 \)

Within patient test-retest variance = (19.8+8+14+11+2.6) ÷ 5 = 11.1

Variance due to differences between patients = 412.6 – (11.1 ÷ 2) = 407.1

Notice that the naïve estimate of between eyes variance calculated from the five means contains a component of test-retest variance that must be subtracted off. In this case the “pooled within eye” variance of 11.1 is divided by two, reflecting the number of measurements per eye.

Another consideration is that while analyzing a simple reproducibility experiment with no missing data is easily accomplished, more complex experimental designs, including differing numbers of observations per eye, measurements of both eyes of participants, simultaneous estimation of variances of measurements made within and between sessions require analysis with a statistical package. Even with access to a sophisticated program such as SAS or SPSS inappropriate specification of the underlying statistical model, such as failure to account for nested designs or other analytical considerations that are not obvious, will produce incorrect estimates.

REPORTING AND INTERPRETING REPRODUCIBILITY ESTIMATES

Once a reproducibility experiment has been performed and variance components analysis carried out, there are a number of useful statistics used to communicate the data.

Test-retest Standard Deviation

The statistic that may be most helpful to a clinician making measurements of a patient is probably the test-retest standard deviation. In the example above that would be the square root of the pooled within patient standard deviation, \( \sqrt{11.1} \), or 3.3. Assuming test-retest differences follow a normal distribution (usually a defensible assumption), one can place a tolerance interval around a patient’s measurement with \( \pm 1.96 \times \text{test-retest SD} \), in this case about \( \pm 6.6 \). Some published reports include the 1.96 multiplier in the test-retest SD estimate and the reader may have to carefully examine the methods section or table footnotes to discern this. This tolerance interval provides the limits of expected measurement variability and a change over time larger than the tolerance limits provides evidence of clinical change.

If the test-retest standard deviation associated with a measuring device is larger than desirable but measurements are cheap, fast, and easy to make, a simple strategy for getting a better estimate of a patient’s data is to make multiple measurements and average them. This will reduce the width of the tolerance interval by a factor of the square-root of the number of measurements. In the example above, the tolerance interval associated with a measurement which is the average of three individual measurements would be reduced from \( 6.6 \) to \( 1.96 \times 3.3 \div \sqrt{3} = 3.7 \). This same consideration applies to the number of exams considered over time for a particular patient when evaluating for change. Further details are available in Budenz et al. Instrument manufacturers may provide sophisticated progression determination algorithms which incorporate these ideas.

Of course, as is the case with visual field measurements, test-retest standard deviations may be associated with the stage of disease. So a tolerance interval created with the test-retest SD estimated from normal control measurements may not apply to, say, patients with moderately severe disease.

Intraclass Correlation Coefficient

If one wishes to compare different measuring technologies, say OCT and GDx, the test-retest variance (noise) must be scaled to the actual differences between patients one wishes to quantify (signal). The standard method for doing this in the statistics literature is with the intraclass correlation coefficient (ICC). This statistic is unrelated to the Pearson correlation coefficient used to assess association between two continuous variables. The simplest form of the ICC is the ratio of the between patient variance to the sum of the between patient and the pooled within patient variances. In the example above, this would be \( 407.1 ÷ (407.1+11.1) = 0.97 \). The ICC can be thought of as the “signal” variance as a percent of the total variance and is frequently expressed as a percent (in this case 97.3%).

What is a good intraclass correlation coefficient? The minimum ICC is zero (all of the variance is due to test-retest) and the maximum is one (none of the variance is due to test-retest). The answer will depend on the specific application of the measuring technology; however, Fleiss provides generally applicable ranges: poor \( < 0.4 \); fair to good, \( 0.4-0.75 \); excellent \( > 0.75 \).
If one is simultaneously estimating multiple sources of variation in an experiment, say, test-retest within session, between session, between technicians, and between two models of the same instrument, variance component analysis is a simple extension of the ICC idea. The magnitude of the variance associated with each of these sources is estimated and can be expressed as a percent of the total variance.

The ICC has one central limitation as a method of expressing reproducibility, by its definition it is only useful for the type of population whose measurements contributed to the calculation. If reproducibility is assessed in controls whose true values all lie within a relatively narrow normal range, the numerator of the ICC (the between patient variance) will be small simply because there are no real differences between the individuals measured to detect. If one wants to get an ICC as large as possible, one simply has to measure subjects whose true values lie between the maximum and the minimum limits of the measuring device, say, normal controls to the most severely diseased patients.

Thus, the ICC is ideally suited to comparing reproducibility of measurements made with different devices on the same subjects. It is least useful (though not necessarily useless) when comparing measurements made on groups of patients with a wide range of disease states.

The ICC is closely related to the kappa statistic often used to summarize agreement of categorical gradings or other non-interval level data.

The Coefficient of Variation

The intraclass correlation coefficient scales test-retest variability to the differences between individuals an instrument is designed to detect. Another popular and intuitive method of quantifying reproducibility is to express the test-retest standard deviation as a percent of the average measurement. This statistic is called the coefficient of variation (CoV) and in the example above would be 11.1/62.5 × 100 = 17.8%.

The advantage and the defect of the CoV stem from the fact that it provides no information about how variable the group being measured is. If the test-retest SD was 10 microns and the group average was 80 microns, the CoV would be 12.5% regardless of whether mean RNFLs ranged from 50 to 110 or 79 to 81. In the former case, the ICC might be on the order of 96%, suggesting that the device could usefully discriminate between individuals, while in the latter case the ICC would be close to zero.

However, this does not mean that the CoV is independent of disease severity. If the test-retest variability remains the same while the parameter being measured, say mean RNFL, decreases with disease progression, the Coefficient of Variation will decrease with progression. This would be consistent with increased difficulty of demonstrating further progression in cases with severe disease.

While the CoV normally supplies the standard deviation as a percentage of the mean, an alternative which may also be useful is to perform the calculation with a known clinically meaningful difference, instead of the mean, in the denominator. This might be, for example, the difference between the average normal value and the average of patients with early manifest glaucoma. This quantity is one form of the “effect size” in a statistical sample size calculation.

Bland-Altman

In an influential 1986 paper in Lancet, Bland and Altman introduced into the medical literature a method for assessing the agreement between a new and an existing measuring technology. This method is graphical and intuitive and has had numerous applications in ophthalmology. Briefly, for each subject, one calculates the signed difference between measurements with the old and new device and the average of the two devices’ measurements. When the differences are plotted against the averages for a large group of subjects, the graph reveals immediately if the differences cluster around zero (as one would hope, if the two technologies are measuring the same thing) and if the differences change, in direction or magnitude as the averages increase from low to high. Bland and Altman also suggest using the standard deviation of the differences, S, as a measure of variance in between instrument agreement, and refer to ± 1.96 × S as “limits of agreement”.

Some authors have, in the setting of quantifying test-retest variability, applied the Bland-Altman limits of agreement calculation to two consecutive measurements made with the same device. There is a simple relationship between Bland-Altman’s S and the usual test-retest standard deviation with two measurements per person. Assuming that the average difference between two consecutive measurements should be zero (which is sensible in a test-retest setting), Bland-Altman’s S = usual test-retest SD * √2. If one collected three test-retest measurements per eye (or person) it is unclear how the Bland-Altman limits of agreement would be constructed. In the setting of test-retest reproducibility Bland himself recommends using the test-retest SD.

EXAMPLES FROM THE LITERATURE

The ophthalmology literature is replete with reproducibility and agreement studies. We will briefly discuss a few of them related to imaging in glaucoma.

Budenz et al examined between session reproducibility (up to 5 measurements per eye on different days) of standard and fast Stratus OCT measurements made on one eye each of 51 patients with a diagnosis of chronic glaucoma with or without field loss. They found no relationship between mean RNFL measurements (range: < 40 to >100 μm) and test-retest SDs, suggesting no need to assess reproducibility separately for
eyes with different disease severity. Expressing results as ICCs, CoVs, and test-retest SDs, they found excellent reproducibility of mean RNFL measurements (ICC = 0.96, CoV = 5.1%) which, as expected, was higher than for quadrants or individual clock hours. Between session test-retest SD of mean RNFL fast scans was 6.7 μm. However, examination of supplementary tables available on the AAO website reveals that almost all of this variability is explained by within session test-retest (5.2 μm for measurements made on the same day).

The influence of different operators and different OCT Stratus instruments on reproducibility was examined separately in normal and glaucomatous eyes by Sehi et al. Differences in mean RNFL between machines and operators accounted for 16.4% and 2.4%, respectively, of total variance in normal subjects and 2.3% and approximately zero percent in glaucomatous subjects. Intraclass correlations for normals and glaucoma measurements were 59% and 91%, respectively. As noted above this discrepancy is because of the much wider range of nerve fiber layer thicknesses in the glaucomatous eyes. The unexplained test-retest SD of 4.5 μm in this study’s normal eyes was identical to that in the glaucoma group.

Fredette et al. analyzed data collected on the same group of patients but for the pattern electroretinogram (PERG). In contrast to the OCT they found a significant correlation of test-retest standard deviations with the magnitude of the PERG amplitude. The amplitude ICC, 0.79, was in the range of good to excellent (> 0.75), but not as high as for OCT measurements. Test-retest SD had previously been found to be lower (by about 50 to 70%) in normal subjects than in glaucoma patients, which may be the more appropriate result clinically, since PERG is most likely to prove useful for measuring functional deficits in the earliest stages of glaucoma.

Strouthidis et al evaluated reproducibility of a variety of parameters of the HRT II with respect to operators and inter-versus intra-session measurements in a group comprised of both ocular hypertensives and primary open angle glaucoma patients. They also studied how HRT software affected the results. They found rim and cup depth to have consistently high ICCs and low CoVs. They observed little effect of source of variability or software on reproducibility in any of the parameters, although reference height differences and poor image quality were important factors.

DeLeon Ortega et al undertook a comprehensive comparison of the reproducibility of GDx-VCC, HRT II, and Stratus OCT in a group of four cohorts: normal controls, as well as glaucoma patients with early, moderate, and severe field loss. Their principal analytic method was Bland-Altman analysis. They concluded that GDx and HRT measurements of nerve fiber layer and optic nerve head parameters (respectively) had reproducibility superior to that of OCT. However, examination of their Bland-Altman plots especially in the severe group reveals that, for example, there is a wider range of thickness measurements made with OCT than GDx. Thus, it may be misleading to assess reproducibility by comparing the Bland-Altman limits of agreement (termed repeatability coefficients in DeLeon Ortega). They subsequently published and compared intraclass correlation coefficients for both devices. In fact, the ICCs of both instruments were excellent in all diagnosis groups; although it is true that in the controls and moderately severe visual field loss groups the small differences (controls: 0.98 vs 0.94; moderate: 0.99 versus 0.96) approached statistical significance.

**SUMMARY**

Reproducibility studies are designed to quantify test-retest variability, so that differences between sequential measurements of a patient due to true clinical change can be separated from those likely to be due to measurement error. No single method of expressing the results of such a study conveys all important information; however, a combination of the test-retest standard deviation, intraclass correlation coefficient (ICC), and the coefficient of variation (CoV) all convey useful information. The Bland-Altman method of assessing agreement has been applied to reproducibility studies, but is not optimal.

**REFERENCES**


“I believe that if one man gains spiritually the whole world gains with him and, if one man falls, the whole world falls to that extent”

— Mahatma Gandhi