**PROBLEM**

If the result of running a progression analysis (i.e., no significant progression) is negative, how sure are we that the visual field (VF) really hasn’t changed over time? Maybe the variability (noise) of the VFs is so high that the progression (signal) is not detectable by our analysis

**IDEA**

Monte Carlo simulation! We can predict the relative (i.e., between different patients) power of the analysis by simulating different rates of VF progression.

**METHODS**

First, a pre-existing dataset comprising 30 patients with stable glaucoma recruited from the glaucoma clinics at the Queen Elizabeth Health Sciences Centre in Halifax (Nova Scotia, Canada) was used to represent the **empirical** test-retest variability associated with threshold static perimetry (HVA) in a real-world experimental setting. Each patient underwent 12 VF measurements in 12 consecutive weekly sessions. The test interval was adapted to 0.5 year for the purpose of our simulation.

Then, each series of measurements was randomly reordered before a negative linear trend starting from *-0.02 dB/y* to *-44.2 dB/y* in *2x* steps was introduced to 5 randomly-sampled locations. \*Pointwise VF sensitivity was allowed to deteriorate beyond zero in a time series. VF locations that met the following criteria were excluded from being sampled:

1. Blind-spot locations — corresponding to locations 26 and 35 in a VF map generated by the visualFields package.
2. Mean (i.e. across 12 VF measurements) VF sensitivity not greater than 10 dB.

Permutation of Pointwise Linear Regression (PoPLR) analysis was run each time after a negative linear trend was ‘injected’. The result was regarded as true positive if the p-value derived from PoPLR was smaller than 0.05. Given that each VF series was randomly reordered 100 different times at each rate of progression, there were 100 p-values associated with each rate of progression. The proportion of true positives was taken as the \**power*.

**EXAMPLE (Px 1)**

1. Location 27 (L27) and 4 other VF locations meet the inclusion criteria because they are not blind-spot locations and each of their mean sensitivities is greater than 10 dB. The following steps apply to all 5 locations (carried out concurrently), but let’s use L27 as an example.
2. In the original L27 time series, the sensitivity values across 12 VFs are as follows: 28, 30, 25, 27, 26, 26, 28, 29, 29, 29, 24 and 27.
3. First, these values are randomly reordered so the series becomes: 25, 24, 27, 30, 26, 29, 29, 28, 28, 26, 27 and 29. Let’s call this the \*‘clean slate’.
4. -0.02 dB/y (the smallest rate of change) is then introduced to the ‘clean slate’. Because the test interval is 0.5y, the time series now becomes: 25, 23.99, 26.98, 29.97, 25.96 ……… 28.88.
5. The ‘clean slate’ is reordered again before repeating step 4.
6. Step 5 is repeated many times so that a total of 100 iterations are performed. PoPLR p-value is derived at each iteration and power is computed by dividing the number of p-values < 0.05 by 100.
7. Step 4 to 6 are repeated with other rates of progression.

**IMPLICATIONS: The big idea?**

Researchers have spent a great deal of time quantifying variability in many different ways to help them tease out signal (progression) from the noise (variability) inherent in VF threshold measurements. The idea is that if we can better quantify variability — which is much easier to do than quantifying power (due to the absence of ground truth) — we can get a better (but vague) idea of whether or not a specific amount of (real) change is likely to be detected. For example, the ***spread*** between the 5th and 95th percentile of test-retest variation has been used as a measure of variability for different levels of underlying VF damage (<http://www.maths.lth.se/matstat/staff/georg/skrift/MinaPublikationer/AmJOphth1989.PDF>) which, in turn, is being used as a yardstick for power (see following scenario):

Question 🡪 What is the likelihood that a change of -3dB can be detected if the baseline TD is -15dB?

Answer 🡪 Very unlikely because the ***spread*** is nearly as wide as the perimetric dynamic range.

In other words, the goal of pursuing better ways to measure variability is to help us make better ***assumptions*** about **power**. However, we have not measured the power itself, so we can only make relatively coarse predictions at the end of the day (e.g., very likely, fairly likely, not very likely or not likely at all to be detected).

Is there an alternative method of predicting power? Yes, but rather than using variability as the means to power prediction, what if we simulate many random scenarios of progression and directly measure power? This way we can effectively (and approximately) ***quantify*** power and the prediction comes not only from variability but all other unknown random variables that may well affect power. What this all means is that:

1. In clinical practice, we can decide how much weight we are to give to a VF series in detecting clinically significant progression. For example, let’s assume that we consider MD -0.5 dB/y as clinically important. The outcome of the power simulation for patient A shows that there’s only a 40% chance that this rate of change will be detected by PoPLR as compared to 95% chance for patient B. It is then sensible for us to assign much more weight to structural signs in detecting progression in patient A than patient B.
2. In research (in a randomized controlled trial in particular), we can increase the ‘power’ of our study by including only participants with demonstrable, reasonable likelihood (e.g., >80%) of being detected if they progress by a pre-determined threshold amount (e.g., MD -0.2 dB/y). This way we can be surer that any null findings are not influenced (as much) by an underpowered analysis.