City Research Online

City St George's, University of London

Citation: Montesano, G., Quigley, H. & Crabb, D. P. (2021). Improving the Power of Glaucoma Neuroprotection Trials Using Existing Visual Field Data. American Journal of Ophthalmology, 229, pp. 127-136. doi: 10.1016/j.ajo.2021.04.008

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/26509/

Link to published version: https://doi.org/10.1016/j.ajo.2021.04.008

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

Abstract (Separate Word file)
Click here to access/download;Abstract (Separate Word file);Abstract_AJO_Revised.docx

Improving the power of glaucoma neuroprotection trials using existing visual field data

Giovanni Montesano, Harry A. Quigley, and David P. Crabb

Abstract

Purpose

Selecting reliable visual field (VF) test takers could improve the power of randomised clinical trials in glaucoma. We test this hypothesis via simulations using a large real world dataset.

Design

Methodology analysis: assessment of how improving reliability affects sample size estimates.

Methods

A variability index (VI) estimating inter-test variability was calculated for each subject using the residuals of the regression of the mean deviation over time for the first six tests in a series of at least 10 exams for 2804 patients. Using data from the rest of the series, we simulate VFs at regular intervals for two years. To simulate the neuroprotective effect (NE), we reduced the observed progression rate by 20%, 30% or 50%. The main outcome measure was the sample size to detect a significant difference ($p < 0.05$) at 80% power.

Results

In the first experiment, we simulated a trial including one eye per subject, either selecting randomly from the database or prioritising patients with low VI. We could not reach 80% power for the low NE with the available patients, but the sample size was reduced by 47% and 57% for the 30% and 50% NE respectively. In the second experiment, we simulated two eyes per subject, one of which was the control eye. The sample size (smaller overall) was reduced by 20% and 60% for the 30% and 50% NE by prioritising patients with low VI.

Conclusions

Selecting patients with low inter-test variability can significantly improve the power and reduce the sample size needed in a trial.

Revised manuscript with changes unmarked - DOC or DOCX files only
Click here to view linked References
Click here to access/download;Revised manuscript with changes unmarked - DOC or DOCX files only;Montesano

Improving the power of glaucoma neuroprotection trials using existing visual field data

Giovanni Montesano[1,2], Harry A. Quigley[3], and David P. Crabb[1]

1. City, University of London - Optometry and Visual Sciences, London, United Kingdom
2. NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
3. Wilmer Institute, Johns Hopkins School of Medicine, Baltimore, MD, United States

Corresponding author: David P. Crabb

Email: david.crabb.1@city.ac.uk

Phone number: +44 (0)20 7040 0191

Fax number: None

Address: City, University of London

Northampton Square

London

EC1V 0HB

United Kingdom

Short title: Improving glaucoma trials using existing visual field data

Introduction

Glaucoma is the second leading cause of world blindness, estimated to affect over 100 million persons during the next decades[1]. Therefore, even modest improvements in glaucoma treatments would prevent blindness in thousands of patients[1, 2]. Presently the only proven treatment for ameliorating glaucoma progression is lowering of intraocular pressure (IOP)[3-5]. New treatments might entail better ways for lowering IOP (surgery/sustained delivery) or a neuroprotective agent. While interest in neuroprotective approaches is increasings[6-8], any new treatment needs to be scrutinised by a clinical trial before it can become widely adopted[9]. Primary outcomes for these trials need to be sensitive enough to detect glaucoma progression in the relatively short time span of the trial. Some investigators, often on the insistence of regulatory bodies, are adopting patient reported outcome measures (PROMs) as a primary trial outcome[10-12]. Yet PROMs have been shown to be insensitive to detecting the small changes in visual function that might occur over the short period of time of a clinical trial[13].

Currently, the best candidate for a primary outcome, one approved by the United States Food and Drug Administration[14] for example, is measurement of the visual field (VF) using standard automated perimetry – an established technology that has been in clinics for more than 30 years. Yet, VF assessment is onerous for some individuals[15] and the measurements themselves can be noisy. This could result in the specification of disease progression being challenging in these patients. As a result, the use of VF worsening as the primary outcome in neuroprotection trials has been considered to require large numbers of persons over several years' time by some investigators.[9] The fact that one large clinical trial of neuroprotection oral treatment was apparently unsuccessful[16] seemingly reinforced this pessimistic viewpoint, though the use of VF testing as an outcome was not the primary reason for study failure

A variety of solutions to improve the chances of detecting VF progression in a clinical trial have been suggested. If trials were to extract information from every single participant[17] by measuring the rate of VF loss, the outcome can be more adequately assessed than if event based outcomes (patients defined as progressing or not) are used[17]. Modelling experiments (computer simulations) have been used to show that sample sizes required for trials can be substantially reduced by evaluating differences in the rates of VF loss between groups using linear mixed-effects models[17]. Another way to improve the power of glaucoma trials is to increase the frequency of VF testing during follow-up[18] or schedule clusters of tests at the beginning/end of the trial period[19, 20]. Such methods were used successfully in the UK Glaucoma Treatment Study[3] and were proven effective in a trial duration of just two years.

Some have suggested that selective patient selection could produce a more rapid outcome, perhaps by recruiting only those patients who are more likely to progress: the elderly, those with exfoliation, or those with higher baseline IOP. However, evidence from clinical trials shows that highly selective recruitment criteria leads to excessively long pre-trial period and the need for many recruiting sites. Others suggest recruiting patients that show rapid VF progression in the recent past. This has an ethical and practical flaw, in that once one knows the patient has recently worsened, IOP must be further lowered, making continued rapid progression less likely. Each of the selective recruitment strategies also is subject to the weakness that any result could fail to generalize to the overall open angle glaucoma (OAG) population and require longer to recruit sufficient subjects, increasing cost.

A more effective method for inclusion of the least number of persons to detect a neuroprotective effect may be to identify and recruit patients with a lower *inter-test* VF variability (sometimes referred to as *between test* variability). The potential for reduction in sample size and study duration using such an approach was previously suggested[21] on a theoretical basis, showing that, for particular levels of *inter-test* field variability and neuroprotective effects, satisfactory sample sizes and study durations could be achieved. Throughout the present report, we presume

that all persons, both in the new treatment and control arm, will have appropriate and similar IOP-lowering therapy. In the present report, we test this approach using modelling experiments based on thousands of real VFs extracted from five different glaucoma clinics in England. We aim to confirm the potential improvement in power obtained by recruiting people by their past VF reliability. We also propose a practical strategy for trial recruitment from an electronic medical record (EMR).

Methods

Dataset

Visual field (VF) data were extracted from an EMR (Medisoft; Medisoft Ltd., Leeds, UK) from five regional National Health Service Hospital Trust glaucoma clinics in England in November, 2015 as described elsewhere[22, 23]. All patient data were anonymised at the point of data extraction and subsequently transferred to a single secure database held at City, University of London. Subsequent analyses of the data were approved by a research ethics committee of City, University of London. The study adhered to the Declaration of Helsinki and the General Data Protection Regulation of the European Union. All VFs were recorded on the Humphrey Visual Field Analyzer (Carl Zeiss Meditec, Dublin, CA, USA) using a Goldmann size III stimulus with a 24-2 test pattern and the Swedish Interactive Testing Algorithms (SITA Standard or SITA Fast). The aggregated database contained 576,615 VFs from 71,361 people recorded between April, 2000 and March, 2015.

For this study, we selected all patients with at least 10 VFs recorded over at least 4 years in one or both eyes. We excluded any patient whose EMR contained ocular surgery other than cataract removal during the follow up period. The use of IOP lowering medications was not consistently reported. However, given that all these patients are being followed up in glaucoma clinics, their IOP would be managed according to standard clinical practice. Qualifying subjects had a Mean Deviation (MD) worse than -2 dB in at least two VFs[24, 25]. It seems likely that subjects with this level of damage and frequency of VF testing were either strong glaucoma suspects or persons with glaucomatous optic neuropathy. Our selection yielded 5,149 eyes from 3,732 people (68,812 VF tests). We then excluded patients with at least one VF test felt to be unreliable due to false positive errors (FP ≥ 15%). No exclusion criteria were applied based on fixation losses or false negative errors. The final selection included 37,281 VF tests from 2,804 subjects. (Of these, 922 patients had sufficient numbers of fields that met the inclusion criteria for both eyes [1844 eyes, 24,316 VF tests].)

Variability index

To be widely applicable in clinical trial design, the variability index (VI) used for patient selection needs to be easily calculated from readily available clinical data. We used the variability of the MD of the first six VFs in the series. Simply put, for each subject we fitted a linear regression on the MD values over time (years) of the first 6 VF tests[23]. The differences between the actual MD value at each point in time and the corresponding predicted value from regression were then calculated (residuals). The VI is simply the standard deviation of these residuals. A higher VI therefore indicated larger *inter-test* variability in a patient's VF follow-up. MD was preferred to the other global indices such as the Visual Field Index (VFI) for comparability of our results with the literature [17] and its simplicity of calculation in the simulated VF (see below). Moreover, the values used to calculate the VFI can change from pattern deviation to total deviation according to the level of damage[26] and this could introduce inconsistencies in the results of the simulations[27]. Pattern Standard Deviation was also avoided because it does not have a monotonic relationship with the level of damage, reverting back towards normal values for more advanced glaucoma[28]. Finally, previous reports have shown MD to be superior to both VFI and PSD in detecting glaucoma progression[29].

Simulation of the visual field series in a clinical trial protocol

In trials, as opposed to everyday clinical practice, VFs are typically measured at regular intervals with a precise sampling scheme. Our simulations followed the sampling scheme (patient visits) used in the UKGTS trial, namely 16 fields over 2 years. Specifically, VF tests were performed at baseline and then at 2, 4, 7, 10, 13, 16, 18, 20, 22 and 24 months, with clustering of 2 fields (test-retest) at baseline and at 2, 16, 18 and 24 months[3].

The simulation must also account for the relationship between sensitivity (dB) and variability at each VF location. A method to quantify this variability, proposed by Russell et al.[30-32] , uses linear

regression of sensitivity over time fitted at each individual VF location for each eye (point-wise linear regression). The residuals from each regression are used to quantify the variability for each sensitivity value predicted by the point-wise linear regression, which is known to be larger at lower sensitivities[30-33]. However, simulating local noise alone is generally not sufficient accurately to reproduce the variability of the MD, which is mostly influenced by global fluctuations in the VF[17, 34, 35]. Such fluctuations affect the VF as a whole rather than acting only on specific locations. They can be determined by a series of factors, but are usually well characterised as a random process[17, 34, 35]. Wu et al.[35] introduced a method to capture such fluctuations, based on the use of "noise templates" mapped in a standardised probability space that is independent of the specific threshold values (see Supplementary Material). In our work, we wished to retain the relationship between the VI of each subject (from the standard deviation of the MD residuals of their first 6 tests) and their point-wise variability in the simulations. To this aim, we used the model proposed by Wu et al. with minor modifications. The full methodology is detailed in the supplementary material. Importantly, our simulations sampled (with replacement) noise templates derived only from the VF series of the subject being simulated. This ensured that the noise in the simulations better reproduced the behaviour typical of that specific patient. **Figure 1** shows our simulation paradigm, with two examples from two patients with different levels of noise.

Calculation of the effect of treatment

Next, to estimate the potential effect of therapy on VF worsening, we simulated multiple series of VFs for each subject using the simulation method described in the previous section. There were four possible series that were simulated for each patient: one series with no treatment effect (control) and three series with increasingly beneficial (hypothetical) neuroprotective effects on ameliorating speed (rate) of progression. These slowed the pointwise VF progression velocity by 0.10, 0.15 or 0.25 dB/year, respectively. These values represent therapeutic improvements of 20%, 30% and 50% (small, medium and large), respectively, based on the average progression rate among our included dataset, which was -0.51 ± 1.04 dB/year in MD. This progression rate was calculated on VF tests after the 6th test in the series for each subject using one eye per subject, the same used for experiment 1 (see later). To simulate a parallel two group clinical trial, subjects were randomly assigned to either the treatment arm (with one of the neuroprotective effect) or to a placebo arm (with no improvement). Following Wu et al.[17], we used a linear mixed effect model with random intercepts and slopes to make full use of the whole series of the 16 individual MD values calculated from the simulated VF tests. The MD for the simulated series was calculated using the **`visualFields`** package for R[36]. An interaction term in the model denoted the difference in progression slope between the two arms. The effect was detected when the p-value for this interaction term was < 0.05. This procedure for the random assignment and the calculation of the p-value was repeated 5000 times for different sample sizes. The power at each sample size was then the percentage of realizations for which an effect of the selected magnitude was detected in 5000 attempts. This overall method was used in two sets of experimental approaches.

Experiment 1: using one eye per patient

We used the described methods to assess the possible effect of prior subject variability on the sample size needed to determine specific treatment outcomes in two ways. The first, presented here, was meant to simulate the effect of a systemic treatment, in which the two eyes cannot be treated separately. In this framework, only one eye per patient was simulated and the comparison was performed between the average effect in the two independent arms of the trial. We ordered (ranked) all eyes according to their VI, as could be easily done in an EMR of a real clinic. Power curves were then calculated by recruiting a progressively increasing number (N) of subjects. Two recruiting approaches were compared: the first relied on random selection of N patients; the second selected the first N subjects in the database ordered by VI (i.e., the N least variable subjects with minimum inter-test variability). For each approach, the N subjects were then randomly split between the treatment and placebo arm, and the process was repeated 5000 times for selected N and each

selection methods. Ideally, this process should be repeated for every N from 1 to the half the size of the sample (1402), so that the same N could be used for both arms of the trials. However, for the practical implementation of the procedure, the calculations were performed at defined N, from 25 to 1402 every 50 subjects, with the last increment equal to 27 eyes.

Experiment 2: two eyes per patient

In this experiment, we simulated a treatment that could be applied to only one eye of the patient, leaving the other untreated. In this case, the fellow eye could be used as an internal control. To apply the selection based on variability, we computed a VI for each subject as the average of the VIs of the two eyes. This subject VI was used for the selection process, which was identical to experiment one. The linear model used to compare progression rates had to be modified to account for this design. In particular, the random effect was applied only at the subject level, since the two eyes were included in the two different arms of the trial. The fixed effect part of the model was the same. However, to model the individual differences in progression rates between the two eyes, we included the interaction term as part of the random slope. The calculations of the power curves were identical to experiment one. For this experiment, the N values of the power curves were from 25 to 922 every 50 subjects, with the last increment equal to 47 subjects.

Results

Median number of VF tests for included patients was 13 (Interquartile range [IQR]: 11, 15). Baseline values for patients were defined as those at the point of their 6th VF test, with a median (IQR) age of 68 (60, 75) years and MD of -6.14 (-11.05, -3.51) dB. Rate of progression was -0.13 ± 0.81 dB/year (mean ± SD) during the first 6 VF tests and -0.51 ± 1.01 dB/year from subsequent examinations. Median (IQR) VI per eye was 1.09 (0.72, 1.67) dB when calculated on the first 6 VFs, and 1.08 (0.74, 1.63) when calculated on the rest of the series. There was a significant correlation between the $\log_{10}$-transformed VIs calculated using the two parts of the series (correlation coefficient = 0.26, $p < 0.001$). Since the data may have been influenced by whether subjects had undergone cataract surgery during follow-up, we present the data stratified by cataract surgery in **Table 1**. For experiment 2, we could only use patients who had both eyes meeting the inclusion criteria. Descriptive statistics for this subset of eyes are also reported in Table 1. This subset included 1844 eyes from 922 subjects, with a median (IQR) number of VF tests of 12 (11, 15). Median (IQR) VI per subject (used in the simulations for experiment 2) was 1.14 (0.82, 1.69) dB when calculated on the first 6 VFs, and 1.18 (0.84, 1.72) when calculated on the rest of the series. There was a significant correlation between the $\log_{10}$-transformed subject VIs calculated with the two parts of the series (correlation coefficient = 0.32, $p < 0.001$) and between the $\log_{10}$-transformed VIs of the two eyes from the same subject (correlation coefficient = 0.46 for the first six VFs, 0.41 for the rest of the series, $p < 0.001$).

Table 2 reports relevant descriptive statistics of the sample for experiment 1 divided into the quartiles of the VI. People with lower Vis were generally younger, with earlier VF damage and showed a slower progression rate.

In experiment 1, persons with lower *inter-test* variability were more likely to determine a treatment benefit with lower sample sizes (**Figure 2, Table 3**). Sample size required to reach 80% power was reduced by 47% for an effect size of -0.15 dB/year (medium) and by 57% for an effect size of -0.25 dB/year. We could not reach 80% power with the random selection for the smallest effect size with either selection method. This also implies that, for this small effect, including all subjects led to less overall power than with a subset selected with the smallest variability criterion. The improvement was also observed in experiment 2, yielding a reduction of 20% and 60% for the medium and large therapeutic effect respectively (**Figure 3, Table 3**). Again, 80% power could not be reached with either method for the smallest effect size. In general, using two eyes per subject largely improved the power over one eye per subject (**Table 3**). In both experiments, we observed transient reduction in power in the smallest variability selection at different points in the curve; this possibly reflects the noise in the selection process, i.e. people that were classified as being less variable in the first six tests did not perform as well during the trial (**Figure 2 and 3**).

Discussion

In our first experiment, we demonstrated that selection of participants with lower *inter-test* VF variability can markedly increase the power of a neuroprotection clinical trial where VFs are the outcome measure. No other feature of a prospective sample for such trials has been shown to reduce their effective patient population and duration. The selection of persons for a trial based on their VF *inter-test* variability far outweighs other potential selection criteria for producing a useful outcome of glaucoma trials. We recognize that patients with low *inter-test* variability (better test takers) are a minority of the overall clinical population, depending upon the criterion selected. For example, patients with a VI of less than 1 dB represented less than half of the patient in the sample (44%, 1243 subjects). Yet, in our approach, we retained the beneficial effect of reducing sample size by basing selection simply on smaller (better) variability indices, instead of using hard thresholds as cut-offs. Our selection method could be based on a ranking list of patients from an EMR containing VF data, meaning it could be easily implemented. The VI could be calculated from the MD values of the VF series and patients selectively recruited with better VIs.

In our second experiment, we showed that this improvement is mostly maintained when a paired comparison between the two eyes of the same patient is performed. In this set of simulations, we assumed that the neuroprotective treatment could be applied independently to one of the two eyes, leaving the other as a control. Such an assumption might not hold true in a realistic scenario if there were possible systemic effects of local treatments. However, if any spurious neuroprotective effect on the fellow eye can be safely disregarded, such an approach is effective in substantially improving power compared to the first experiment, where only one eye per subject was analysed. This result was somewhat unexpected, because it implies some similarity in the rate of progression between the two eyes of the same patient, so that one eye could serve as an internal control for the other. The observed inter-eye correlation of the rate of progression was not strong (Correlation coefficient = 0.54, $p < 0.001$), but it was enough to yield a gain in power. Moreover, a recent report also showed, in a different population of glaucoma patients, that a correlation exists between the VF progression rates of the two eyes from the same patient[37]. The effect of our improved selection method based on the VI was generally smaller for this experiment. This could be a consequence of the fact that the VI was calculated at a person level by averaging the value for the two eyes. This could have diluted the efficacy in the detection of more variable eyes. However, the fact that an improvement is still measurable strengthens the idea that most of the variability in the MD is due to global fluctuations, which are likely a feature of the subject rather than the individual eye. Indeed, there was a significant correlation between the VI of the two eyes within the same subject (Correlation coefficient = 0.61, $p < 0.001$).

Others have used computer based simulations to assess sample sizes for glaucoma treatment trials using trend-based analyses, Wu et al.[17] indicated that such a method is much more powerful than an event based approach. However, there are some differences with our work. They used mixed models to calculate sample sizes required to obtain an improvement in VF progression at 90% power, compared to our criterion of 80% power. Their method also differed by simulating a series of 10 VF tests over two years, as opposed to 16 in our simulations following the UKGTS scheme. The median (IQR) progression rate assumed in their dataset, -0.57 [-0.98, -0.24] dB/year (Median [IQR]) and their baseline MD (-3.23 [-5.53, -1.85] dB) were very similar to ours. Also, our analysis was based on more than 1.9 million VF data points (for experiment 1) from patients from five regionally different clinics compared to one cohort of data used by Wu et al (321 eyes from 240 patients). Discrepancies in our conclusions are most likely explained by the different approach to simulations and modelling. Wu et al.[17] used a sigmoid function to fit the VF series and calculate the residuals. Their model had more parameters than a linear regression and this could have reduced the estimated variability from the residuals used to inform the simulations[38]. Most importantly, they simulated the neuroprotective effect by completely halting the progression of a proportion of randomly selected patients. In contrast, we applied more realistic, graded reductions in the

progression rate. This means that for any effect size, we retained the distribution of rates of progression from the real dataset and simply shifted their average value. Consequently, even people in the neuroprotection arm were allowed to progress to lower sensitivities, albeit with a slower rate on average, and this alone would systematically increase the variability. This effect would have been reduced if we had simulated zero progression for some of the patients, as in Wu et al[17].

There are other differences between our simulations compared to that of Wu et al. For example, there is an issue with 'positive slopes', a situation that arises where the VF sensitivity seemingly 'improves' over time as a likely result of measurement variability or a learning effect. Halting the progression rate, as assumed by Wu et al, implies eliminating positive slopes altogether; this is not realistic because positive slopes are a common feature in series of VFs in clinical settings and trials[18, 39, 40]. Positive slopes are a 'real' source of variability when trying to quantify the neuroprotective effect and should therefore, in our opinion, be included in the simulations. Setting progression rates to have a null (zero) slope would not only affect the average progression rate but also reduce the variability of the slopes in the simulations. To show the effect of this different approach, we provide, in the supplementary material, additional simulations for experiment 1 performed by halting the progression rate for different proportions of patients, according to the desired neuroprotective effect. The estimated sample sizes did not show important changes for the random selection. However, any advantage of our hierarchical selection method was lost. This is explained by considering that halting progression in a proportion of patients leads to a proportional reduction in the global progression rate. This means the linear difference in the rate of progression, the one measured by the linear mixed model used in the analysis, would change according to the progression velocity in the selected sample. In our case, people with lower VIs also showed slower rates of progression. Therefore, when a proportional change in progression velocity is modelled, the simulated linear difference would be smaller for people with lower VIs, nullifying the effect of the reduced perimetric noise. Ultimately, it is impossible to determine how a neuroprotector would affect measured VF loss in a real trial and both scenarios are possible. However, these additional simulations show that, even in the worst case scenario of a pure proportional effect, our selection method would not cause any loss in power.

There is another subtle, yet important, difference between our novel simulations compared to the approach used by Wu et al involving the implementation of the linear mixed model used for the power calculation. Wu et al. only included random intercepts in their modelling (private communication with the corresponding author), as opposed to our approach which also included a random effect on the slopes. Removing the random slope term has been shown to bias the standard error of the estimates for the population slope, leading to overoptimistic sample size calculations[41]. To show this effect, we performed additional simulations for experiment 1 (supplementary material) showing that excluding the random slope term from the model yielded much smaller sample sizes, very close to the estimates provided by Wu et al. We think this highlights an important hazard when planning a clinical trial: inadequate statistical modelling might cause underestimation of the true variability of the phenomenon under investigation, biasing sample size calculations and preventing generalisability of results.

Our simulations suggest that some treatment effects can be detected with manageable sample sizes. However, smaller effects will require substantial numbers of patients and potentially multicentre trial design. Given our patient group, a 20% reduction in the rate of VF progression could not be detected with 80% power. At 50% reduction, the standard random method of selection required 514 patients in the treatment and control groups each to reach sufficient power. In contrast, with our proposed optimal selection, the number of patients can be reduced to 222. This result represents a practical and achievable recruitment target even at individual glaucoma centres. Furthermore, existing patient records at such a center can immediately identify patients with low variability, dramatically speeding recruitment time, which is a large expense in any clinical trial. Different schemes for the frequency of VF tests were not explored in this work. Instead, for

simplicity, we adopted the one used for the UKGTS, which yielded a good separation of the two arms of the trial in as little as 18 months[3]. However, testing different schemes could also have beneficial effects on the power of the trial and could be the subject of future work. Wu et al.[20] partially addressed this question, showing that both the UKGTS and clustered VF testing schemes can reduce the sample size requirements of trials when compared to regular sampling.

We noticed the rate of progression in the first six VF tests (-0.13 ± 0.80 dB/year) was better (less negative) than the rest of the series (-0.52 ± 1.05 dB/year). This may be explained by a long-term learning effect[42] as shown in the supplementary material, where we show a larger percentage of positive progression rates before the 6th VF. We also considered this effect to be explained by the role of cataract surgery during follow-up, but our data do not support this. For example, among those receiving cataract surgery, there was no systematic difference in the average progression rate calculated on the first 6 VF tests compared to the patients that did not receive any surgery (see Supplementary material and Table 1). Other factors that might have influenced the progression rates in a real world long term follow-up are the reduced adherence to treatment over time and progressive loss of treatment efficacy (e.g. tachyphylaxis for drops, waning effect of laser treatment and failure for surgical procedures).

Our work has potential limitations. First, our definition of glaucomatous VF loss was restricted to using MD values, a relatively non-specific indicator of glaucoma damage. Furthermore, by using a criterion of MD ≤ -2 dB loss, we may have included subjects with minimal damage who were less likely to have progressive glaucomatous neuropathy. However, there was a wide range of VF loss (see Supplementary material and Table 2) and the average progression rate calculated after the 6th VF test was similar to the value obtained by Wu et al.[17] and compatible with other curated datasets of glaucoma subjects[18, 43]. Another important aspect to discuss is the definition of our VI. Being based on the standard deviation of the residuals of a linear regression of MD over time, VI could be influenced by the level of VF loss[30, 44]. This could lead to a substantial selection bias towards people with earlier damage or with slower rate of progression. We performed additional analyses (Supplementary material and Table 2) to show that this effect was present but small in magnitude for our sample. Specifically, we show that the MD and the rate of progression became slightly more negative as more people with higher VIs were included, but this change was small when compared to the variability of those same indices within the selected sample. In turn, this suggests that the main determinant of variability for the MD is global fluctuations, which are not necessarily correlated with the level of damage[35]. Moreover, in our sample, the correlation between VI and baseline MD (at the beginning of the simulated trial) was statistically significant, but extremely weak (increase of 0.015 in VI for every dB loss of MD, $p < 0.001$, correlation coefficient = -0.17). However, as previously explained, the importance of this selection bias largely depends on the assumptions made about the effect of the neuroprotective treatment (i.e., additive or proportional on the observed progression rate) and this will need to be the focus of further research. The VI measured on the first six VFs was not a perfect predictor of the VI for the remaining of the test series (Table 2). This is noteworthy because it implies that the repeatability of the performance could be worse during the trial than previously calculated. Finally, we focussed on a narrow set of scenarios for our main experiments. Although we explore other possibilities in the supplementary material, deciding which conditions are likely to be best mimic reality is problematic. This limitation should be kept in mind when interpreting the results of both ours and other simulation experiments.

There are further practical limitations to our results. The calculation of the VI requires six VF tests.; however, we believe this was the best compromise in preliminary simulation experiments, as VIs calculated with shorter series were much less effective in improving the power of our simulated trial. Yet, such a number of VFs is likely to be available for patients followed for a reasonable amount of time. These patients are likely to have their treatment optimised and therefore show a slower progression rate. In this case as well, if the neuroprotection effect is assumed proportional to the rate of progression, the amount of measurable difference might be reduced.

In conclusion, we showed that recruitment of subjects with lower inter-test measurement variability is an efficient way to maximise the power and minimize sample size of a trial for a new treatment for glaucoma.

Acknowledgments

Funding/Support: None.

Financial disclosures:

- Giovanni Montesano:
  consultant for CenterVue
- Harry A. Quigley:
  consultant for Sensimed, IDx, Gore, Intense and Equinox;
  research support from Kali Care and Heidelberg Engineering
- David P. Crabb:
  consultant for CenterVue and Apellis;
  speaker's fee from Santen, Allergan, THEA, Bayer;
  unrestricted research funding from Santen, Allergan and Apellis;
  patents: T4 and ANSWERS

Other acknowledgements: we greatly thank Dr. Zhichao Wu for the constructive discussion about the simulation and modelling approach.

References

1. Tham YC, Li X, Wong TY, Quigley HA, Aung T, Cheng CY. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. Ophthalmology 2014;121:2081-90.
2. Quigley HA. Glaucoma Neuroprotection Trials Are Practical Using Visual Field Outcomes. Ophthalmol Glaucoma 2019;2:69-71.
3. Garway-Heath DF, Crabb DP, Bunce C, et al. Latanoprost for open-angle glaucoma (UKGTS): a randomised, multicentre, placebo-controlled trial. Lancet 2015;385:1295-304.
4. Heijl A, Leske MC, Bengtsson B, et al. Reduction of intraocular pressure and glaucoma progression: results from the Early Manifest Glaucoma Trial. Arch Ophthalmol 2002;120:1268-79.
5. Anderson DR, Normal Tension Glaucoma S. Collaborative normal tension glaucoma study. Curr Opin Ophthalmol 2003;14:86-90.
6. Guymer C, Wood JP, Chidlow G, Casson RJ. Neuroprotection in glaucoma: recent advances and clinical translation. Clin Exp Ophthalmol 2019;47:88-105.
7. Lawlor M, Danesh-Meyer H, Levin LA, Davagnanam I, De Vita E, Plant GT. Glaucoma and the brain: Trans-synaptic degeneration, structural change, and implications for neuroprotection. Surv Ophthalmol 2018;63:296-306.
8. He S, Stankowska DL, Ellis DZ, Krishnamoorthy RR, Yorio T. Targets of Neuroprotection in Glaucoma. J Ocul Pharmacol Ther 2018;34:85-106.
9. Sena DF, Lindsley K. Neuroprotection for treatment of glaucoma in adults. Cochrane Database Syst Rev 2017;1:CD006539.
10. Azuara-Blanco A, Burr J, Ramsay C, et al. Effectiveness of early lens extraction for the treatment of primary angle-closure glaucoma (EAGLE): a randomised controlled trial. Lancet 2016;388:1389-1397.
11. Gazzard G, Konstantakopoulou E, Garway-Heath D, et al. Laser in Glaucoma and Ocular Hypertension (LiGHT) trial. A multicentre, randomised controlled trial: design and methodology. Br J Ophthalmol 2018;102:593-598.
12. King AJ, Fernie G, Azuara-Blanco A, et al. Treatment of Advanced Glaucoma Study: a multicentre randomised controlled trial comparing primary medical treatment with primary trabeculectomy for people with newly diagnosed advanced glaucoma-study protocol. Br J Ophthalmol 2018;102:922-928.
13. Jones L, Garway-Heath DF, Azuara-Blanco A, Crabb DP, United Kingdom Glaucoma Treatment Study I. Are Patient Self-Reported Outcome Measures Sensitive Enough to Be Used as End Points in Clinical Trials?: Evidence from the United Kingdom Glaucoma Treatment Study. Ophthalmology 2019;126:682-689.
14. Weinreb RN, Kaufman PL. The glaucoma research community and FDA look to the future: a report from the NEI/FDA CDER Glaucoma Clinical Trial Design and Endpoints Symposium. Invest Ophthalmol Vis Sci 2009;50:1497-505.
15. Glen FC, Baker H, Crabb DP. A qualitative investigation into patients' views on visual field testing for glaucoma monitoring. BMJ Open 2014;4:e003996.
16. Weinreb RN, Liebmann JM, Cioffi GA, et al. Oral Memantine for the Treatment of Glaucoma: Design and Results of 2 Randomized, Placebo-Controlled, Phase 3 Studies. Ophthalmology 2018;125:1874-1885.
17. Wu Z, Crabb DP, Chauhan BC, Crowston JG, Medeiros FA. Improving the Feasibility of Glaucoma Clinical Trials Using Trend-Based Visual Field Progression Endpoints. Ophthalmol Glaucoma 2019;2:72-77.
18. Chauhan BC, Malik R, Shuba LM, Rafuse PE, Nicolela MT, Artes PH. Rates of glaucomatous visual field change in a large clinical population. Invest Ophthalmol Vis Sci 2014;55:4135-43.
19. Crabb DP, Garway-Heath DF. Intervals between visual field tests when monitoring the glaucomatous patient: wait-and-see approach. Invest Ophthalmol Vis Sci 2012;53:2770-6.

20. Wu Z, Medeiros FA. Impact of Different Visual Field Testing Paradigms on Sample Size Requirements for Glaucoma Clinical Trials. Sci Rep 2018;8:4889.
21. Quigley HA. Clinical trials for glaucoma neuroprotection are not impossible. Curr Opin Ophthalmol 2012;23:144-54.
22. Liu X, Kelly SR, Montesano G, et al. Evaluating the Impact of Uveitis on Visual Field Progression Using Large-Scale Real-World Data. Am J Ophthalmol 2019;207:144-150.
23. Kelly SR, Bryan SR, Crabb DP. Does eye examination order for standard automated perimetry matter? Acta Ophthalmol 2019;97:e833-e838.
24. Boodhna T, Crabb DP. Disease severity in newly diagnosed glaucoma patients with visual field loss: trends from more than a decade of data. Ophthalmic Physiol Opt 2015;35:225-30.
25. Crabb DP, Saunders LJ, Edwards LA. Cases of advanced visual field loss at referral to glaucoma clinics - more men than women? Ophthalmic Physiol Opt 2017;37:82-87.
26. Bengtsson B, Heijl A. A visual field index for calculation of glaucoma rate of progression. Am J Ophthalmol 2008;145:343-53.
27. Rao HL, Senthil S, Choudhari NS, Mandal AK, Garudadri CS. Behavior of visual field index in advanced glaucoma. Invest Ophthalmol Vis Sci 2013;54:307-12.
28. Blumenthal EZ, Sapir-Pichhadze R. Misleading statistical calculations in far-advanced glaucomatous visual field loss. Ophthalmology 2003;110:196-200.
29. Gardiner SK, Demirel S. Detecting Change Using Standard Global Perimetric Indices in Glaucoma. Am J Ophthalmol 2017;176:148-156.
30. Russell RA, Crabb DP, Malik R, Garway-Heath DF. The relationship between variability and sensitivity in large-scale longitudinal visual field data. Invest Ophthalmol Vis Sci 2012;53:5985-90.
31. Saunders LJ, Russell RA, Crabb DP. Measurement precision in a series of visual fields acquired by the standard and fast versions of the Swedish interactive thresholding algorithm: analysis of large-scale data from clinics. JAMA Ophthalmol 2015;133:74-80.
32. Wu Z, Medeiros FA. Comparison of Visual Field Point-Wise Event-Based and Global Trend-Based Analysis for Detecting Glaucomatous Progression. Transl Vis Sci Technol 2018;7:20.
33. Henson DB, Chaudry S, Artes PH, Faragher EB, Ansons A. Response variability in the visual field: comparison of optic neuritis, glaucoma, ocular hypertension, and normal eyes. Invest Ophthalmol Vis Sci 2000;41:417-21.
34. Bryan SR, Eilers PH, Lesaffre EM, Lemij HG, Vermeer KA. Global Visit Effects in Point-Wise Longitudinal Modeling of Glaucomatous Visual Fields. Invest Ophthalmol Vis Sci 2015;56:4283-9.
35. Wu Z, Medeiros FA. Development of a Visual Field Simulation Model of Longitudinal Point-Wise Sensitivity Changes From a Clinical Glaucoma Cohort. Transl Vis Sci Technol 2018;7:22.
36. Marin-Franch I, Swanson WH. The visualFields package: a tool for analysis and visualization of visual fields. J Vis 2013;13.
37. Anderson AJ, Gardiner SK. Using the Rate of Glaucomatous Visual Field Progression in One Eye to Help Assess the Rate in the Fellow Eye. Ophthalmology Glaucoma 2020;3:360-368.
38. Otarola F, Chen A, Morales E, Yu F, Afifi A, Caprioli J. Course of Glaucomatous Visual Field Loss Across the Entire Perimetric Range. JAMA Ophthalmol 2016;134:496-502.
39. Wright DM, Konstantakopoulou E, Montesano G, et al. Visual Field Outcomes from the Multicenter, Randomized Controlled Laser in Glaucoma and Ocular Hypertension Trial (LiGHT). Ophthalmology 2020;127:1313-1321.
40. Heijl A, Buchholz P, Norrgren G, Bengtsson B. Rates of visual field progression in clinical glaucoma care. Acta Ophthalmol 2013;91:406-12.
41. Heisig JP, Schaeffer M. Why You Should Always Include a Random Slope for the Lower-Level Variable Involved in a Cross-Level Interaction. European Sociological Review 2019;35:258-279.
42. Gardiner SK, Demirel S, Johnson CA. Is there evidence for continued learning over multiple years in perimetry? Optom Vis Sci 2008;85:1043-8.
43. Heijl A, Bengtsson B, Hyman L, Leske MC, Early Manifest Glaucoma Trial G. Natural history of open-angle glaucoma. Ophthalmology 2009;116:2271-6.

44. Russell RA, Garway-Heath DF, Crabb DP. New insights into measurement variability in glaucomatous visual fields from computer modelling. PLoS One 2013;8:e83595.

Figure legends

Figure 1. Trial simulations, example from two subjects with different variability. Real data on the left: the first 6 tests are used to calculate the variability index (VI), as the standard deviation of the residuals of the regression line. The second part of the VF series (shaded) is used to seed the simulations. Simulations on the right: obtained using the method proposed by Wu et al. Notice how, in these cases, the simulations preserve the difference in inter- test variability between the subjects.

Figure 2. Power curves for experiment one (one eye per subject). The dashed horizontal line represents the usual 80% threshold used in clinical trials. The coloured labels report estimated number of trial participants to reach 80% power with the two selection methods.

Figure 3. Power curves for experiment two (two eyes per subject). The dashed horizontal line represents the usual 80% threshold used in clinical trials. The coloured labels report the estimated number to reach 80% power with the two selection methods. The horizontal axis is scaled differently than Figure 2. Both eyes of each subject were included.

Click here to access/download;Table - drawn in Word, .DOC & .DOCX;Table_1.docx

	Cataract surgery		
One eye per subject (Experiment 1)			
	None (N = 2001)	**Before the 6th VF test (N = 406)**	**After the 6th VF test (N = 329)**
Baseline age (years)	67 [57, 74]	71 [65, 76]	73 [66, 78]
Baseline MD (dB)	-5.91 [-10.63, -3.39]	-7.22 [-12.85, -4.38]	-6.58 [-12.39, -3.52]
Variability Index (dB)	1.07 [0.69, 1.63]	1.34 [0.91, 1.89]	1.00 [0.68, 1.66]
Number of tests	12 [11,15]	14 [11,16]	12 [11,14]
Rate of Progression (dB/year)			
First 6 tests	-0.07 [-0.41, 0.25]	-0.18 [-0.56, 0.13]	-0.12 [-0.49, 0.19]
After the 6th test	-0.33 [-0.79, -0.02]	-0.34 [-0.79, 0.02]	-0.46 [-1.04, -0.07]
Two eyes per subject (Experiment 2)			
	None (N = 1276)	**Before the 6th VF test (N = 309)**	**After the 6th VF test (N = 261)**
Baseline age (years)	68 [58, 75]	73 [66, 77]	75 [68, 79]
Baseline MD (dB)	-5.95 [-10.79, -3.35]	-7.69 [-13.00, -4.51]	-6.58 [-12.45, -3.58]
Variability Index (dB)	1.12 [0.71, 1.66]	1.39 [0.93, 1.94]	0.95 [0.65, 1.71]
Number of tests	12 [11,15]	13 [11,16]	12 [11,14]
Rate of Progression (dB/year)			
First 6 tests	0.01 [-0.35, 0.34]	-0.14 [-0.53, 0.21]	-0.10 [-0.44, 0.21]
After the 6th test	-0.29 [-0.77, 0.01]	-0.40 [-0.82, 0]	-0.47 [-0.97, -0.08]

Table 1. ***Summary demographics*** *Median [IQR] stratified according to when and if each group received cataract surgery. All values are reported as the Mean ± Standard Deviation. Baseline age and MD refer to the beginning of the 'trial', so at the time of the sixth exam in the sequence. The top part of the table describes summary statistics for the database used for Experiment 1, the bottom part the one used for Experiment 2.*

Click here to access/download;Table - drawn in Word, .DOC & .DOCX;Table_2.docx

	Group 1	Group 2	Group 3	Group 4
			Median [IQR]	
Baseline age (years)	67 [57, 74]	67 [60, 74]	69 [60, 76]	70 [62, 76]
RoP (dB/year)	-0.26 [-0.61, -0.01]	-0.34 [-0.74, -0.02]	-0.39 [-0.91, -0.03]	-0.45 [-1.06, -0.02]
Baseline MD (dB)	-4.1 [-7.06, -2.94]	-5.08 [-9.01, -3.54]	-6.02 [-10.28, -4.06]	-8.07 [-11.25, -5.51]
			Mean ± SD	
Baseline age (years)	65.02 ± 11.72	66.25 ± 11.14	66.77 ± 11.91	68.01 ± 11.17
RoP (dB/year)	-0.38 ± 0.74	-0.46 ± 0.89	-0.56 ± 0.98	-0.65 ± 1.32
Baseline MD (dB)	-6.29 ± 5.53	-7.21 ± 5.58	-8.12 ± 5.71	-9.07 ± 4.8

Table 2. Sample characteristics according to the Variability Index. *Patients in the main experiment (Experiment 1) were divided into four groups based on the quartiles of the distribution of the Variability Index. The descriptive statistics are reported both as Median [Interquaritile range (IQR)] and Mean ± Standard Deviation (SD). Patients with a lower Variability Index had generally a slower rate of progression (RoP) and a lower Mean Deviation (MD). All values refer to the part of the VF series used to simulate the trial, i.e. from the 7th visual field test onwards.*

Click here to access/download;Table - drawn in Word, .DOC & .DOCX;Table_3.docx

Total number of subjects to reach 80% power		Effect (Progression reduction) -0.1 dB/year (20%)	-0.15 dB/year (30%)	-0.25 dB/year (50%)
Experiment 1 **One eye**	**Smallest variability**	Not reached	1038 [EP: 91%]	286 [EP: 95%]
	Random	Not reached	1680	566
Experiment 2 **Two eyes**	**Smallest variability**	Not reached	458 [EP: 85%]	124 [EP: 95%]
	Random	Not reached	615	201

Table 2. Results of simulations. *Number of subjects to reach 80% power for the three different neuroprotective effects. The numbers represent the total amount of subjects, i.e. the sum of the subjects from both arms of the trial. In brackets, we reported the equivalent power (EP), which is the power obtained with the smallest variability selection at the same sample size required to obtain 80% power with the random selection. * 80% not reached, reporting the total number of subjects; EP not calculated*

Figure 1

Click here to access/download;Figure - label number in Description field; no PDFs;Figure_1.tiff

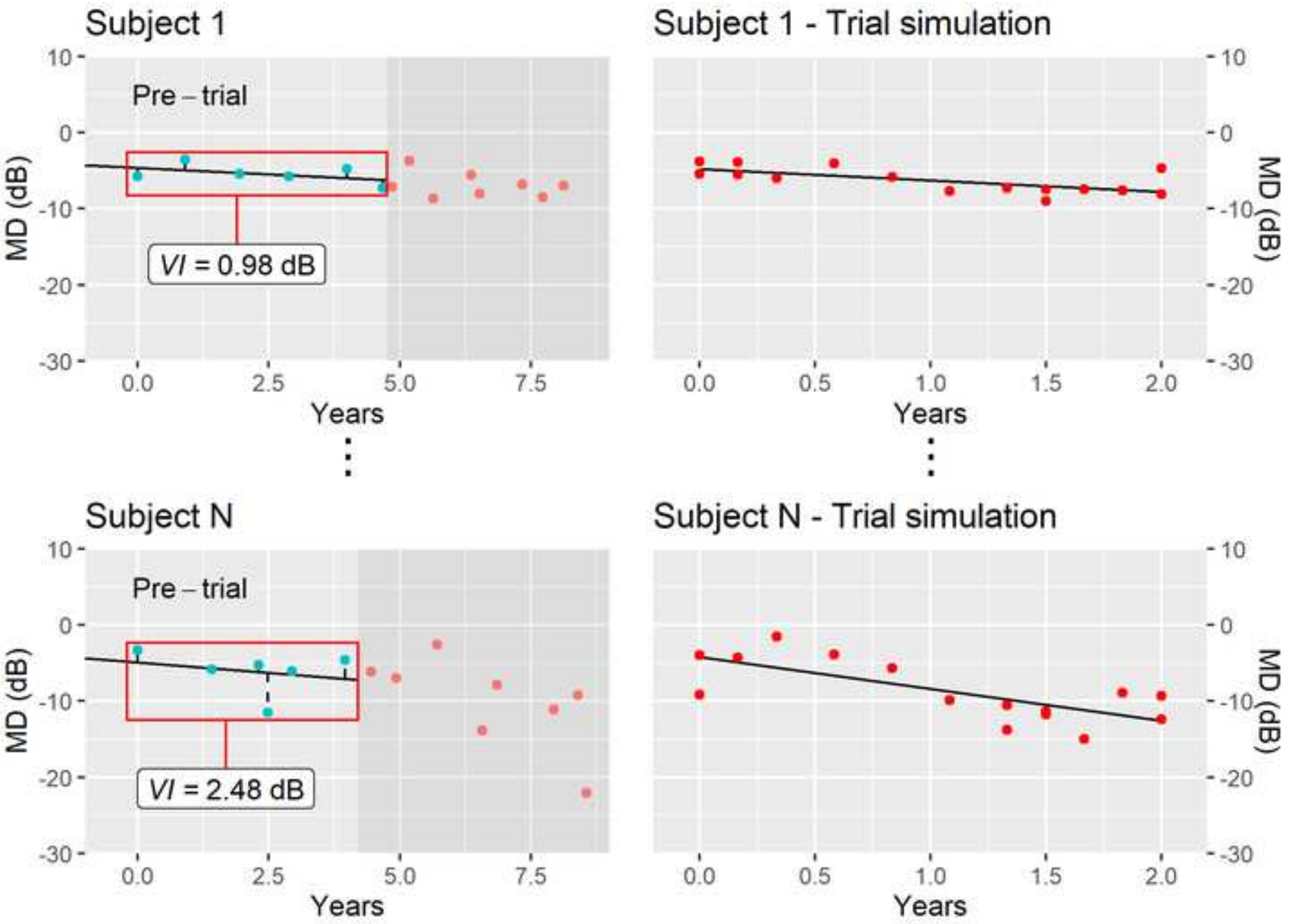


Figure 2

Click here to access/download;Figure - label number in Description field; no PDFs;Figure_2.tiff

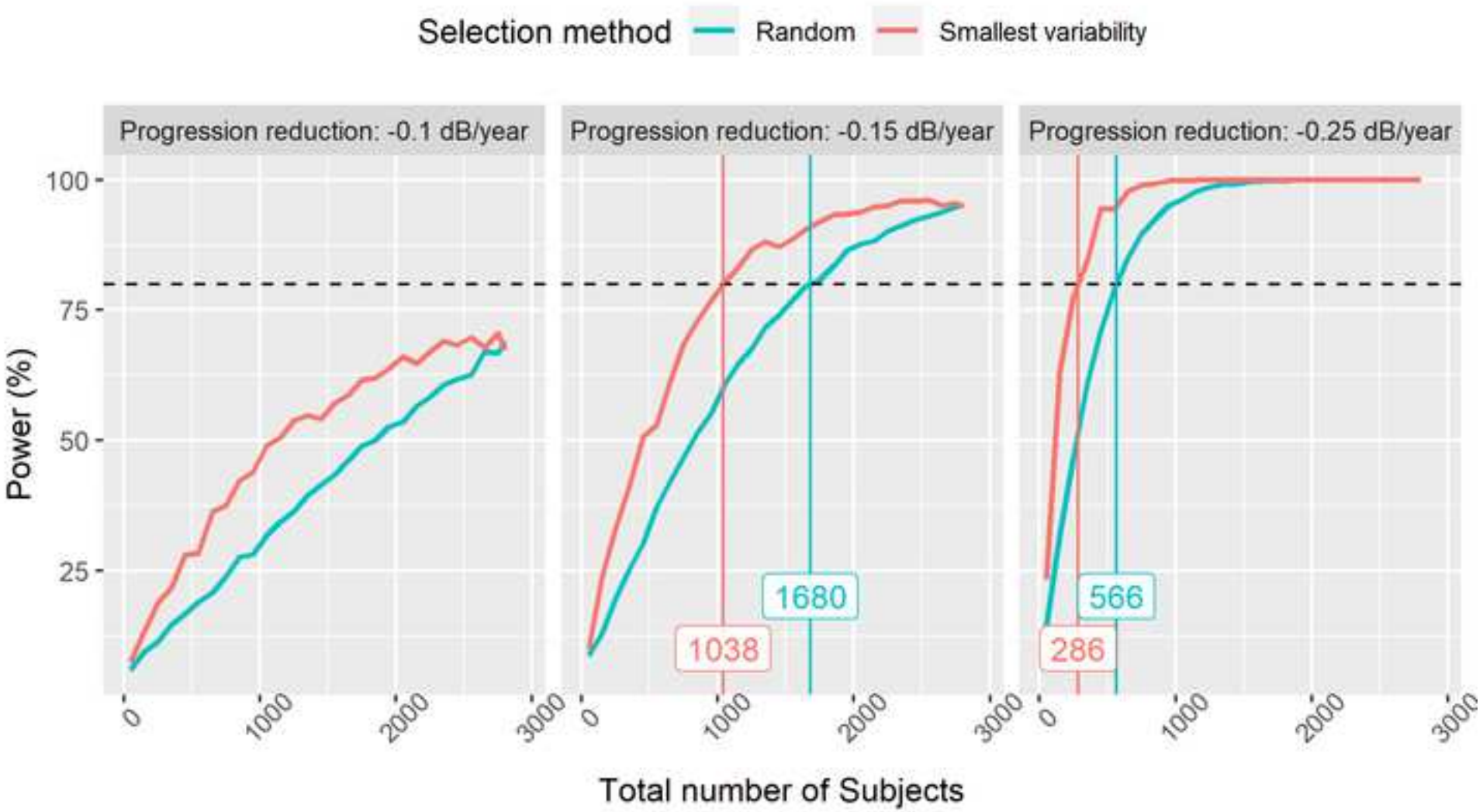


Figure 3

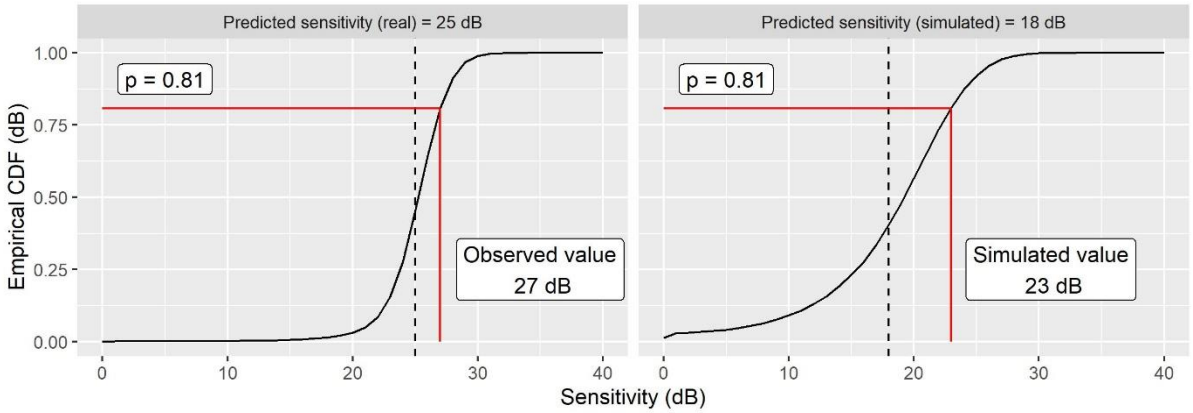
Click here to access/download;Figure - label number in Description field; no PDFs;Figure_3.tiff

Reduction in the rate of progression of visual field damage is a required endpoint for neuroprotection trials in glaucoma. However, variability in visual field tests can reduce the statistical power and require very large sample sizes. Selecting patients with lower variability on visual field tests can reduce the number of participants needed, improving the statistical power of neuroprotection trials.

Supplementary material

Simulation approach

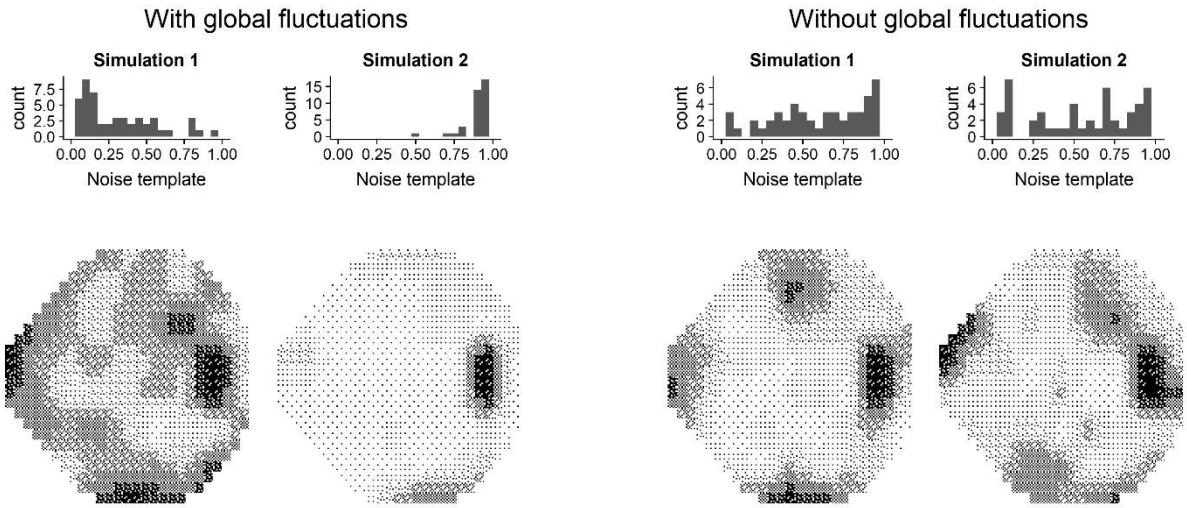
Simulations were based on methodology proposed by Wu et al.[1]. In brief, the empirical distribution function (eCDF) of the residuals was calculated from point-wise regressions in the whole dataset for each predicted sensitivity value, rounded to nearest positive integer. All residuals were transformed into a probability value based on the CDF for the corresponding predicted sensitivity (Supplementary Figure 1, left panel). Repeating this for all visual field (VF) locations allows for each VF to be transformed into, as referred to by Wu et al, a "noise template" composed of probability values. These templates are then applied to simulate a new VF, independently of the specific simulated sensitivity, by inverting the eCDF corresponding to the "true" sensitivity (Supplementary Figure 1, right panel). Importantly, we used a linear regression instead of a sigmoid function to calculate the residuals and generate the predicted sensitivity values for the simulations. This choice was determined by the easier implementation of a change in slope based on the simulated neuroprotective effect. Instead, Wu et al.[1] simulated the neuroprotective effect by completely halting progression in the neuroprotected arm for a variable percentage of people (defined as "responders").



Supplementary figure 1. *Empirical CDF of the residuals for two different predicted sensitivity values. The observed fluctuation (27 dB) is transformed into a probability (0.81) given the CDF corresponding to the predicted sensitivity (25 dB). This can then be mapped onto a CDF for a different predicted sensitivity (18 dB) generating a new value for the simulation (23 dB). Notice that the same offset in the probability space generates a different amount of offset in dB depending on the CDF, accounting for the differences in variability for each predicted sensitivity.*

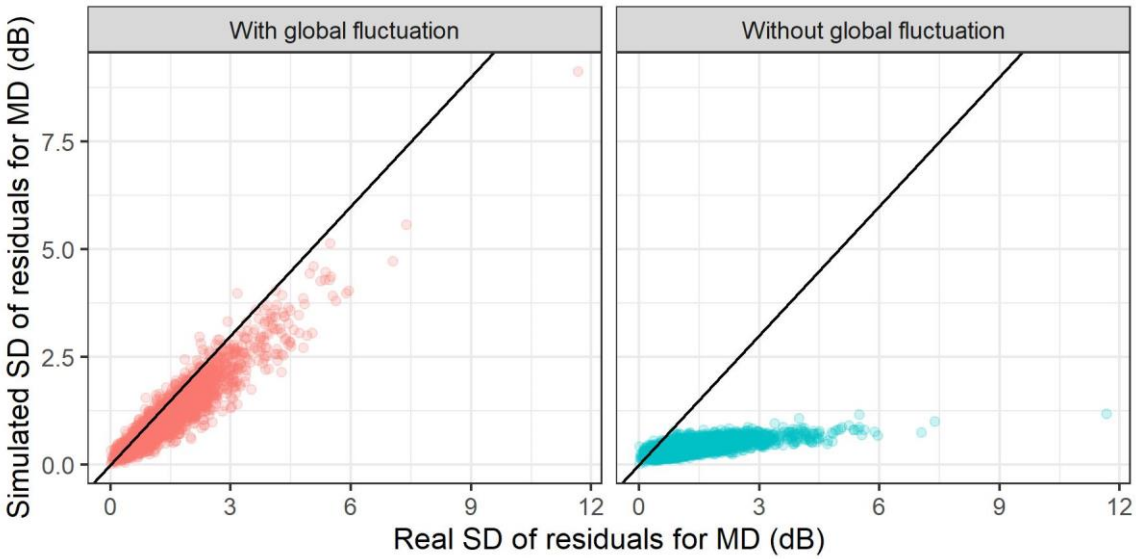
This process also captured the correlation among locations within a VF (global fluctuations) since tests with positive fluctuations will have a noise template composed of probability values closer to 1, whereas tests with negative fluctuations will generate noise templates with values towards 0. We think this is an important component of the capturing the inter-test variability in series of VF measurements. The value of considering these fluctuations is seen in the examples in Supplementary Figure 2. In our simulations, when simulating a given subject, we sampled (with replacement) the noise templates from that subject. This preserved both the global fluctuations and the variability typical of that subject. This was essential for our selection method based on variability to have an effect on the power of the simulated trial. To increase the number of available templates, we randomly shuffled the probability values within each template at every extraction. This does not compromise the global fluctuations since the distribution of probability values within the template remains unchanged. For each eye and each neuroprotective effect, we simulated 100 VFs series. Then, to calculate the power curves, we randomly split the selected sample between the placebo

and treatment arm. At each realization, we randomly selected one of the 100 simulated VFs for each eye at each time-point for the desired neuroprotective effect (null for the placebo arm).



Supplementary figure 2. *Four simulation examples generated with and without accounting for global fluctuations from the same field. The probability values for each noise template are shown in the top panels. Global fluctuations are removed in the examples on the right by sampling randomly from the probability values of different templates from the same subject. Notice how the two visual fields on the right are much more similar compared to those on the on the left, showing that not accounting for global fluctuations can reduce the variability observed in real test series.*

Notably, the global fluctuations were the main component of the observed differences in variability between individuals when measuring the Mean Deviation (MD). Indeed, From Supplementary Figure 3, we can see that only simulating local noise (without the global component) leads to severe underestimation of the variability of MD. Although a reverse calculation using the noise template perfectly reproduced the original VFs, random sampling of the noise templates produced a slightly smaller variance in the simulated series and this is a partial limitation of the methodology.

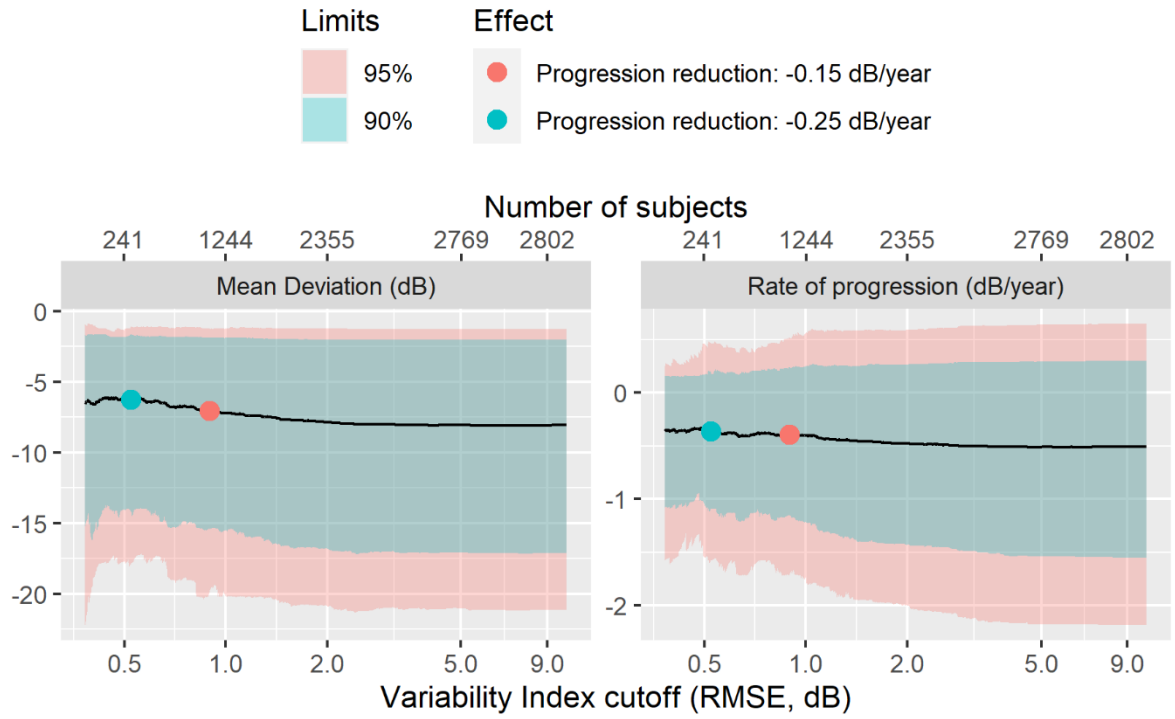


Supplementary figure 3. *Results obtained as the average of 100 simulations, performed at the same time points as the observed VF tests in the actual series, with not change to the progression slopes. Despite some mild underestimation of the variability, the use of global fluctuations effectively*

accounts for most of the differences in variability observed between subjects in the real series. On the contrary, disregarding such global fluctuations leads to severe underestimation of the variability of the MD.

Effect of variability cut-offs on the characteristic of the selected sample

Measurement variability in VFs is associated with the level of VF loss; eyes with more VF damage return more variability. Here we estimate the size of this the effect of our selection on the composition of the recruited sample. Supplementary Figure 4 shows the effect of the variability cut-off used for recruitment on the MD and speed (rate) of progression (dB/year) values of the selected sample (numerical values refer to experiment 1 and are given in Table 2). MD used for this calculation is the predicted value at the time of the 6th VF test (inclusion MD). Values for the rate of progression were the ones used to simulate the trial. As expected, the average MD was slightly higher (better VFs) when the variability cut-off was more restrictive, but the range of defects included (outlined by the 90% and 95% bands in the figure, left panel) is always very large and does not vary greatly across the whole range of the variability cut-offs. The average rate of VF loss (dB/year, in the right panel) was also slightly less negative at lower variability cut-off values. However, the selection process affected its variability range much more significantly (funnelling), in that more extreme progression slopes were less represented among subjects with a lower variability index (Supplementary Figure 4 and Table 2). This however applied to both positives and negative slopes, indicating that subjects that are more variable are more likely to yield 'extreme' progression slopes.



Supplementary Figure 4. *Effect of variability cut-off on the distribution of MD and rate of progression values within the selected sample. The solid black line represents the average value as the variability threshold is increased, whereas the shaded coloured areas represent the 90% and 95% quantile limits. The two coloured dots represent the variability thresholds needed to include the number of subjects required to reach 80% power with the progressive selection method (from the power curves reported in Figure 4).*

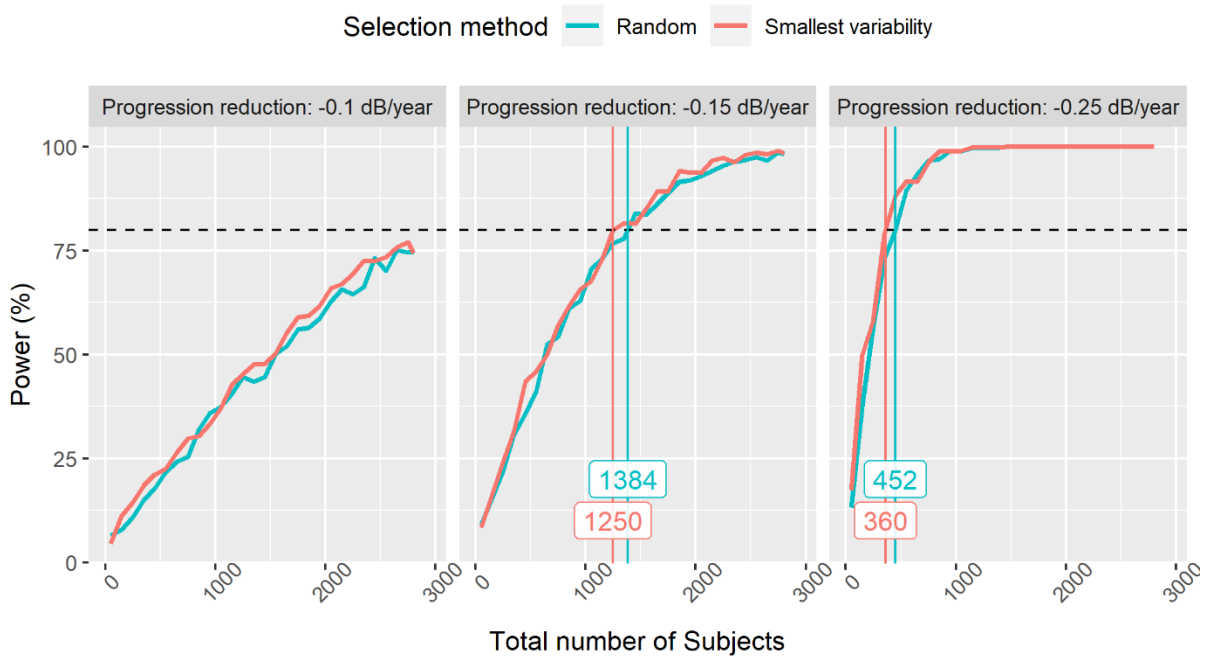
Effect of learning on the rate of progression

The learning effect can be visualized by showing the percentage of positive slopes when progression is calculated on a sliding window of four consecutive VF tests in a sequence, each starting at different points along that sequence. For example, the slope attributed to the first visit will be calculated on the first 4 exams. Then, the sliding window is advanced and the slope attributed to the second visit is calculated on visits 2 to 5, and so on. As the sliding window of four exams moves further into the sequence of each individual test series, the percentage of positive slopes in the whole sample can be calculated. Supplementary figure 5 shows that this percentage diminishes in time and reaches a 'steady' value at the 7th exam. This is consistent with previous reports[2].

Supplementary Figure 5. *Percentage of positive slopes calculated with a moving window of 4 exams along each series. The calculated slope is attributed to the first test in the window. The width of the bars is proportional to the sample size. In both patients who received cataract surgery during the recorded follow-up and those who did not receive it, the number of positive slopes reaches a steady value at approximately the 7th test. The width of the bars is proportional to the sample size.*

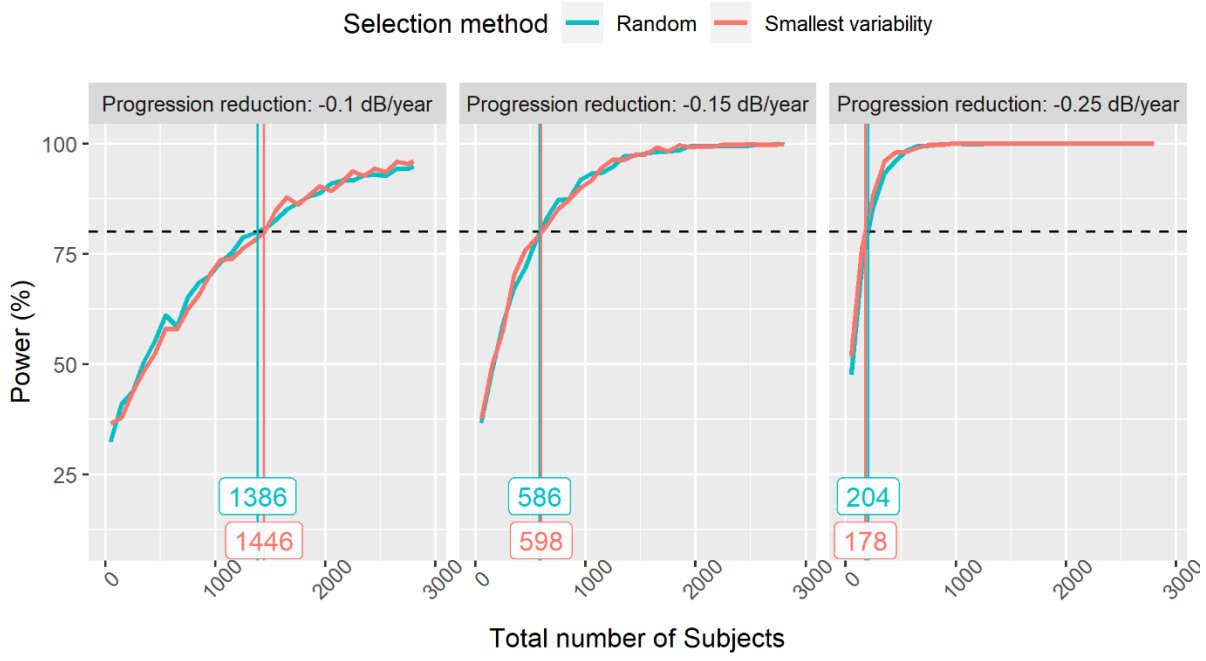
Additional simulations experiments

Supplementary Figure 6 reports the power curves for experiment 1 obtained by completely halting progression for an increasing proportion of eyes according to the desired neuroprotective effect, as in Wu et al. [1]. Notice that the effect is proportional to the progression rates of the selected sample. In this simulation scenario, the increase in power with the hierarchical selection is lost, because people with lower VIs also exhibit slower average progression effect. Hence, the simulated linear neuroprotective effect (i.e. the difference in progression slope between the two arms) is smaller when less variable subjects are selected.



Supplementary Figure 6. *Power curves obtained by simulating a proportional neuroprotective effect (N = 1000 simulations).*

Supplementary Figure 7 reports the power curves for simulations performed as in Supplementary Figure 6, but with the linear mixed model only including a random intercept term and not a random slope term. The power in higher due to the biased standard error for the estimate of the population rate of progression (and their difference between treatment arms). The estimated sample sizes are also very similar to those reported by Wu et al. [1].



Supplementary Figure 7 *Power curves obtained by simulating a proportional neuroprotective effect and without modelling random slopes (N = 1000 simulations).*

References

1. Wu Z, Medeiros FA. Development of a Visual Field Simulation Model of Longitudinal Point-Wise Sensitivity Changes From a Clinical Glaucoma Cohort. Transl Vis Sci Technol 2018;7:22.
2. Gardiner SK, Demirel S, Johnson CA. Is there evidence for continued learning over multiple years in perimetry? Optom Vis Sci 2008;85:1043-8.