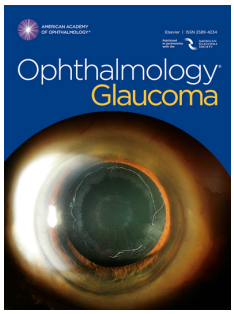
Journal Pre-proof



Comparing the rate of retinal nerve fibre layer and visual field loss as outcomes in glaucoma trials

Giovanni Montesano, David P. Crabb, Giovanni Ometto, David F. Garway-Heath

PII: S2589-4196(26)00010-4

DOI: https://doi.org/10.1016/j.ogla.2026.01.010

Reference: OGLA 785

To appear in: *Ophthalmology Glaucoma*

Received Date: 6 October 2025

Revised Date: 7 January 2026

Accepted Date: 16 January 2026

Please cite this article as: Montesano G, Crabb DP, Ometto G, Garway-Heath DF, Comparing the rate of retinal nerve fibre layer and visual field loss as outcomes in glaucoma trials, *Ophthalmology Glaucoma* (2026), doi: https://doi.org/10.1016/j.ogla.2026.01.010.

Comparing the rate of retinal nerve fibre layer and visual field loss as outcomes in glaucoma trials

Giovanni Montesano[1,2]; David P. Crabb[2]; Giovanni Ometto[1,2]; David F. Garway-Heath[1]

1. NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
2. City St George's, University of London, Optometry and Visual Sciences, London, United Kingdom

Corresponding author: Giovanni Montesano
Email: g.montesano@ucl.ac.uk
162 City Road,
London,
EC1V 2PD

Keywords: glaucoma; visual field; OCT; RNFL; clinical trials; neuroprotection; outcome measures.

Funding/Support: this work and GM are supported by funding from Glaucoma UK and the Royal College of Ophthalmologists. This research was supported by the NIHR Biomedical Research Centre at Moorfields Eye Hospital and the UCL Institute of Ophthalmology. The RAPID data collection was funded by the United Kingdom's National Institute for Health Research Health Technology Assessment (HTA) Project Funding: 11/129/245.

Financial Disclosures: a conflict-of-interest statement is provided separately for each author.

Abstract

Purpose: to compare the statistical power of structural and visual field (VF) outcomes for randomised clinical trials (RCTs) in glaucoma.

Design: analysis of retrospectively collected data.

Participants: Eighty-two glaucoma patients were recruited to a test–retest study, during which up to ten 24-2 SITA Standard VF and circumpapillary retinal nerve fibre layer (cpRNFL) Spectralis OCT scans were collected in separate sessions over 3 months.

Methods: Eyes with at least three sessions with a reliable VF (false positives < 15%) and cpRNFL scan (quality index ≥ 25 dB) were selected (127 eyes, 68 patients) to model the test-retest variability and the structural floor effect. These estimates were combined with a published realistic structure-function progression model from the United Kingdom Glaucoma Treatment Study to simulate longitudinal RCTs (30% neuroprotective effect). Simulations only included data from eyes with early to moderate VF loss (Mean Deviation, MD, ≥-10 dB, 107 eyes, 65 patients). Simulations were repeated 5000 times to estimate sample size requirements to detect a significant difference ($p < 0.05$) in the rate of change of MD and average cpRNFL thickness, estimated with a linear mixed effect model. We also tested the power of a significant outcome with either metric ($p < 0.025$). A supplementary analysis was performed including eyes with early VF loss only (MD ≥-6 dB).

Main outcome measures: sample size at 80% power for the linear rate of MD, cpRNFL and their combination.

Results: at 80% power, the required sample size (patients [95%-Confidence Interval]) was 38% smaller for the MD rate (292 [300, 283]) than the cpRNFL rate (470 [481, 459]). The sample size for the combined outcome was only marginally smaller than the MD alone (275 [283, 268]). The supplementary analysis on eyes with early VF loss showed similar results.

Conclusions: using realistic modelling of structure-function progression and test-retest data, MD progression showed higher statistical power cpRFNL as an outcome measure for clinical trials.

The main objective glaucoma management is to prevent further loss of vision by reducing the speed of disease progression. Progression of glaucoma is primarily monitored with visual field (VF) testing, a direct quantification of the patients' field of vision. The introduction of optical coherence tomography (OCT) imaging has also allowed the monitoring of structural changes of the optic nerve head and the circumpapillary retinal nerve fiber layer (cpRNFL). However, the correlation between functional and structural changes is imperfect[1,2], because of variability in test results and inherent characteristics of structural and functional parameters. The biggest discrepancy between structure and function arises from the structural floor effect[1,2]: large changes in VF can occur without apparent loss of RNFL, especially in more advanced disease. This floor effect is partially a consequence of the difference in scale (logarithmic for VF, linear for OCT), but it is also a direct effect of non-neural tissue which contributes to the RNFL thickness measured by imaging[1].

At present, lowering the intraocular pressure (IOP) is the only recognized approach to treat glaucoma. However, there is active research in the discovery and validation of non-IOP related neuroprotection, with some compounds already being tested in phase III randomized clinical trials (RCTs). The most common outcome measure is the rate of VF progression[3,4]. Linear mixed effect models (LMM) are often used to identify statistically significant differences in the average rate of progression of the Mean Deviation (MD), a global summary metric of VF loss[5–9]. OCT metrics, such as cpRNFL, have been considered as alternative outcomes. Imaging derived metrics are particularly attractive because their perceived lower variability and the belief that early structural loss precedes functional changes. However, rigorous comparisons do not confirm this view[10,11]. In a recent analysis of data from the United Kingdom Glaucoma Treatment Study (UKGTS)[3], we have also shown that the true rate of functional and structural progression are largely the same, once the confounding effects of measurement scale and structural floor are minimized[2].

One important aspect to consider for clinical outcome measures, especially in RCTs, is their statistical power to detect a treatment effect. This is often quantified with the help of mathematical approximations or computer simulations[5–7,12,13]. These, however, rely on accurate modelling of test variability and progression. There is evidence to support that MD progression can be accurately described by a linear decay[14,15] and its test-retest variability is well characterized[16–19]. In contrast, cpRNFL and other structural metrics exhibit a non-linear behavior over time, also as a consequence of the floor effect[20]. The relationship between structural and functional rates of change is also complex. These intricacies have often been overlooked in previous research[12,21] and can greatly affect a fair comparison between structural and functional metrics.

Our recent description of structural and functional progression in UKGTS offers a comprehensive framework for realistic simulation of glaucoma progression[2]. We combine this improved framework and test-retest data from a cohort of glaucoma patients to provide accurate estimates of the statistical power of VF and OCT in glaucoma neuroprotection RCTs.

Methods

RAPID test-retest dataset

Eighty-two clinically stable glaucoma patients were recruited to a test–retest study.[22] The study was undertaken in accordance with good clinical practice guidelines and adhered to the tenets of the Declaration of Helsinki. The study was approved by the North of Scotland National Research Ethics Service committee (reference no. 13/NS/0132), and NHS Permissions for Research were granted by the Joint Research Office at University College London Hospitals NHS Foundation Trust on December 3, 2013. All patients provided written informed consent before the screening investigations were carried out. Criteria for inclusion were: reproducible VF loss with congruent damage to the optic nerve head; no other condition that could lead to VF loss; age > 18 years old; visual acuity of at least 20/40; refractive error within ±8 diopters (D); an IOP of < 30 mmHg; a VF mean deviation (MD) better than −16 decibels (dB) in the worse eye and better than −12 dB in the better eye. Patients performed VF testing and OCT imaging in up to 10 separate appointments over a period of 3 months, during which no meaningful progression of the disease was expected. VF testing was undertaken with a Humphrey Field Analyzer (HFA) using a SITA Standard strategy with a 24-2 pattern. Unreliable tests were repeated on the same day (with a break of at least 30 minutes). Circumpapillary RNFL OCT imaging (cpRNFL-OCT, 12 degrees scan diameter) was carried out using a SPECTRALIS Spectral Domain OCT (software version 5.2.4) in follow-up mode using the same baseline test.

For this study, we selected eyes that had at least three episodes in their test-retest session with a corresponding reliable VF (false-positive errors ≤ 15%)[23] and a cpRNFL-OCT scan with a quality index ≥ 25 dB. Note that, in each session, the operator was allowed multiple scan attempts, to achieve the highest possible quality. If more than one OCT scan was available at the same visit, we chose the one with the highest quality scan index. Data from all available eyes meeting these criteria were used for the characterization of the floor effect and of the test-retest variability, stratified by damage (see later). However, for the simulations, we only included eyes with an average MD ≥ -10 dB, replicating previous studies[12]. This was meant to prevent a large influence from the perimetric and structural measurement floor, although the structural floor effect was explicitly modeled (see later). The variability was quantified as the standard deviation (SD) of the test-retest series. OCT data were missing for 3 patients (6 eyes); one additional patient (both eyes) did not have any OCT scans of sufficient quality. No visits were excluded because of unreliable VF tests (see flowchart in **supplementary material** for details). The final selection for the simulations was composed of 881 tests performed in 107 eyes of 65 subjects. The descriptive statistics for this sample are reported in **Table 1**. It should be noted that most of this sample (85/107 eyes) had early damage (MD ≥ -6 dB). However, a **supplementary analysis** was also performed with simulations including eyes with a MD ≥ -6 dB.

Simulation experiments

Simulation model

The simulations were based on the modelling described in Montesano et al.[2,13] For VF, the model describes the observed linear rate of MD progression as a combination of a sign-reversed exponential distribution, representing the distribution of 'true' negative progression rates, and a Gaussian distribution, modelling the uncertainty introduced by test variability. The mean of the Gaussian distribution also captures the effect of learning, i.e. a positive bias in the rates of progression from patients' initial inexperience with the test. When fitted on patients' data, the model can estimate the distribution of 'true' rates of progression in a population. In Montesano et al.[2], we extended this model to study the functional and structural progression in the UKGTS. For structural data, the learning was set to zero and the data were transformed into a dB scale, to replicate the scale of the MD data. The model also estimates the correlation between the 'true' rates of structural and functional progression. Of the different implementations of the model in Montesano et al.[2], these simulations used the one where the average measurement floor was removed from the structural data before taking the logarithm to transform in the dB scale. The elements of the model relevant for the simulations are described in detail below. A flowchart is provided as **supplementary material**.

Structural and functional progression

For each eye, the true rate of MD and cpRNFL progression was sampled from a sign-reversed exponential distribution. The structural and functional rates were sampled as correlated observations, as explained in Montesano et al.[2], using the within-eye correlation estimated from the UKGTS data (0.75).

For the MD, the mean rate was -0.38 dB/year, the average 'true' rate reported for a large cohort of glaucoma patients under active management[13]. For the cpRNFL, the mean rate was 61% of the MD rate (i.e. -0.23 dB/year). As explained above, this is the cpRNFL rate in dB scale, after removing the average floor effect. The 61% ratio was derived from the model estimates in UKGTS. In Montesano et al.[2], we have shown that this difference in the true rate of MD and cpRNFL progression is likely an artifact arising from the fact that the MD is the average of dB values, whereas the transformed cpRNFL is the logarithm of the average cpRNFL thickness. Indeed, the average true rates of functional and structural progression were very similar when the MD was replaced with the logarithm of the average of un-logged sensitivity values. However, to replicate a typical clinical trial scenario, our simulations use MD and cpRNFL and therefore retain this artifactual difference in true rate. Note that the proportional effect of IOP was essentially identical between structure and function, regardless of the functional metric used. This suggests that the effect of treatment has a similar proportional effect on progression. We make this assumption in our simulations (see below).

Each set of simulated true rates was paired randomly with an eye in the RAPID cohort. The average MD and cpRNFL calculated from the test-retest data were used as the baseline true MD and cpRNFL in the simulations. Simulated true values were generated for 16 visits over 2 years, following the testing schedule in UKGTS, with clustering of two test at 0, 2, 16, 18 and 24 months (same testing schedule for VF and OCT). For MD, the linear rate was used to directly calculate the simulated true values. For the cpRNFL, additional transformations were required to simulate realistic progression: 1) a structural floor level was randomly generated (see later) and subtracted from the baseline; 2) the floor-corrected baseline was transformed into a dB scale ($Baseline_{dB} = 10 \times log_{10}(Baseline_{\mu m})$); 3) the simulated structural rate was used to generate simulated true cpRNFL values over time, in dB scale; 4) the dB values were reconverted in linear scale before adding the generated floor value back in. An example of the simulation for an individual eye is shown in **Figure 1**.

The structural floor effect cannot be determined directly for individual eyes. The statistical distribution used to sample the floor value was determined by fitting a linear structure-function model, similar to the one proposed by Hood and Kardon[1], using the un-logged average MD and the average cpRNFL from the RAPID cohort (**Figure 2**). As previously mentioned, this part of the modelling did not exclude patients with an MD < -10 dB, to obtain a better estimate of the structural floor (127 eyes, 65 patients). In the model, the un-logged MD is the independent variable and the intercept is an estimate of the floor. The distribution for the floor effect was a Gaussian with mean equal to the estimated intercept (56.7 µm) and standard deviation equal to the residual standard error (12.1 µm). Whenever the sampled floor value was higher than the assumed cpRNFL baseline thickness, the value was replaced with the 2.5% confidence quantile of the estimated average baseline ($Baseline_{\mu m} - 1.96 \times SD/\sqrt{N}$), where SD is the test-retest standard deviation for that eye. It should be noted that, while both cpRNFL and MD are measured with noise, these results are obtained from averaging at least 3 test results per eye, with 102 / 127 eyes having 5 or more test results available, improving the accuracy of our estimates.

Simulation of test variability

We used the test-retest data from the RAPID cohort to generate a population model for the expected average variability according to the level of MD and cpRFNL loss. Similarly to the floor effect, we did not exclude patients with an MD < -10 dB when constructing our population model of variability. For MD, we fitted a generalized linear model (logarithmic link function for the mean with a Gamma distributed error) predicting the test-retest variance (calculated for each eye) according to the average MD. We used a quadratic relationship to describe the data (see **Figure 2**). Note that, because of the logarithmic link function, the predicted values for the variance cannot be negative. For the cpRNFL, we did not find any significant relationship with the average thickness (**Figure 2**, p = 0.951).

To capture the inter-eye variation in test-retest variability, we calculated the ratio between each eye's calculated and predicted test-retest variance. Note that, because they are

calculated for each eye and used as paired values, these variance ratios also capture the across-eye correlation between the magnitude of variability of structural and functional tests. Also note that, because there is no change in the assumed cpRNFL variability with the level of damage, the predicted variability is simply the average variance and this procedure returns the original SDs for the structural metrics. We then calculated the within-eye correlation between the standardized structural and functional test-retest residuals, to quantify how much the test-retest residuals correlated across visits in each eye.

We used the variability models defined above to add realistic noise to our simulations. We first generated a pair of correlated standardized residuals for each eye, using a standard bivariate Gaussian distribution and the within-eye residual correlation calculated for each eye. For clustering visits, we also accounted for the correlation between simulated test repeats on the same day. Medeiros et al.[24] reported a within cluster correlation of approximately 0.2 for both OCT and VF tests. Their tests were, however, not performed on the same day. The same-day test correlation is likely to be similar for OCT, but higher for VF, due to an overall performance effect. In the UKGTS cohort used in Montesano et al.[2], we estimated a same-day VF correlation of 0.33, calculated by adding the visit effect as a random intercept term to a standard LMM, together with subject specific random intercepts and slopes, similarly to Bryan et al.[15] This same LMM was also used to estimate the outcome of a trial accounting for such correlations, and is described in the next paragraph. We therefore simulated a same-day correlation of 0.2 for OCT tests and 0.33 for VF tests.

In summary, for the simulations, we calculated the predicted variance of structural and functional tests (a function of the simulated MD for VF tests; the average variance for cpRNFL). We scaled this predicted variance by the specific variance ratio of the eye being simulated. These variances were transformed into standard deviations and used to scale the standardized residuals calculated in the previous step. These residuals were finally added to the simulated true values.

Simulated outcomes for randomized clinical trials

For each simulated trial, we generated simulated test series for an increasing number of eyes (from 100 to 2000, in steps of 100), sampled with replacement from the main selection cohort (MD ≥ -10 dB). These eyes were randomly assigned to the treatment or placebo arm. For the treatment arm, the true rate of progression was reduced by 30% (-0.27 dB/year for MD, -0.16 dB/year for cpRNFL), i.e. the same proportional change was applied to both structure and function. A 30% treatment effect was chosen since it is often reported as being clinically meaningful and detectable with feasible sample sizes in neuroprotection RCTs[5,6,12]. Note that larger effects would affect the sample size but not the relative differences in power of the two outcomes. A **supplementary analysis** assuming the same baseline rate and the same treatment difference (in dB/year) for both structure and function was also performed.

For each run of the simulation, following previous literature[5,6,12], we tested the difference in the average rates of structural and functional progression between the two arms using a LMM, with random intercepts and slopes. The LMM used either the MD or the linear cpRNFL as outcome variables. Note that this is different from the more complex model used to simulate realistic structural and functional progression, described above, and that the structural outcome for the LMM was the linear cpRNFL (μm/year). We also tested an alternative version of the LMM, that would model the same-day variability with the addition of a random intercept term for the test cluster, similar to the global visit effect proposed by Bryan et al.[15]. This LMM was the same used to calculate the 0.33 same-day correlation from UKGTS data and used for our simulations (see previous paragraph).

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 \text{Arm}_i + \beta_3 \left(t_{ij} \times \text{Arm}_i\right) + b_{0i}^{\text{Subj}} + b_{1i}^{\text{Subj}} t_{ij} + b_{0j}^{\text{Visit}} + \varepsilon_{ij}$$

In the formula, β_0, β_1, β_2, β_3 are the fixed effects, with β_3 representing the difference in rate between the two arms. The random slopes are represented by the term b_{1i}^{Subj}. The two random intercept terms b_{0i}^{Subj} and b_{0j}^{Visit} represent the subject and visit effect respectively. All random effects and the residuals are assumed to follow a Gaussian distribution. The standard LMM is identical, but missing the b_{0j}^{Visit} term. The same-day correlation can be estimated as

$$\text{Same-day correlation} = \frac{\sigma^2_{\text{Visit},0}}{\sigma^2_{\text{Visit},0} + \sigma^2_{\varepsilon}}$$

The simulations were repeated 5000 times. A p-value < 0.05 was considered statistically significant. The statistical power was calculated as the percentage of simulations with a statistically significant difference. The standard error for the power were calculated as $SE = \sqrt{P_{p<0.05} \times \left(1 - P_{p<0.05}\right)/N}$, where N is the number of simulations and $P_{p<0.05}$ is the proportion of p-values below the significance threshold. Following Wu and Medeiros[12], a combination outcome (significant difference in either the MD or the cpRNFL progression) was also tested. For the combination outcome, the significance threshold was lowered to 0.025, to maintain the same false-discovery rate. A null-hypothesis simulation is provided as supplementary material, confirming a false discovery rate close to the expected 5% for all outcomes. Statistical testing was not performed to compare the power curves, because simulations allow for arbitrarily large sample sizes.

Results

The within-eye residual correlation between MD and cpRNFL was, on average, very small (Mean ± SD: -0.07 ± 0.4). The power and sample size calculations obtained from the simulations are reported in **Figure 3** and **Table 2**. In general, MD showed higher statistical power than cpRNFL as an outcome. At 80% power, the estimated sample size was 38% lower

for MD than for cpRNFL. The combined outcome (significant difference in either the MD or cpRNFL progression) performed marginally better than either outcome in isolation, but overall very similarly to MD alone. Modelling the same-day test correlation (i.e. the global visit effect) did not have any meaningful impact on the statistical power of either outcome, beyond small random fluctuations in the simulations (**Table 2**).

Importantly, neglecting the same-day correlations did not cause any systematic bias in the estimates (**Figure 4**). The LMMs explicitly accounting for the visit effect were able to correctly estimate these correlations, on average, with a much larger variability in the estimate for the cpRNFL outcome. For the largest sample size, in the control arm, the average estimated rate of progression across simulations (reported as mean ± SD of the trial results, not of individual-level data) was -0.38 ± 0.02 dB/year and -0.91 ± 0.02 µm/year for the MD and cpRNFL respectively. The estimated average difference due to treatment across trial simulations was 0.11 ± 0.02 dB/year and 0.26 ± 0.02 µm/year for the MD and cpRNFL.

The supplementary analysis including only eyes with early damage (MD ≥-6 dB) showed similar results. The supplementary analysis assuming the same baseline rate and the same treatment difference (in dB/year) for both structure and function also confirmed our main results. These results are provided as **supplementary material**.

Discussion

In our simulation experiments, MD progression performed better than average cpRNFL progression as a clinical outcome for RCTs in terms of statistical power and sample size requirements. The treatment effect on the rate of progression was measured with a LMM, a standard approach for establishing statistically significant differences in the rate of glaucoma progression[6–9]. Differently from previous literature, the power of LMMs was tested with simulated data derived from a realistic model of MD and cpRNFL progression in glaucoma. We used a combination of sophisticated modelling based on data from a landmark clinical trial (UKGTS) and extensive test-retest data collected from a cohort of glaucoma patients. Additionally, we have evaluated an implementation of the LMM that accounts for correlations between tests performed on the same visit. This version of the LMM was able to correctly identify these correlations, explicitly introduced in the simulations and based on published and experimental data[24]. However, this did not have a meaningful impact on the results in terms of statistical power and accuracy of the estimates.

Determining the optimal outcome measure is an essential step in the design of glaucoma RCTs. The recent development and testing of novel neuroprotective treatments has reignited the interest in efficient outcome measures[25,26]. VF testing is a proven and well-established technology to monitor glaucoma progression, and has a direct linkage to patients' vision-related quality of life[27–29]. Indeed, VF has been successfully used to establish the effectiveness of glaucoma treatment[3,4]. However, proving IOP independent neuroprotection

in patients actively treated to lower their IOP is particularly challenging because of the slower rate of progression compared to untreated patients. The detection of differences in the rate of MD progression with LMM has been proposed as a more powerful technique compared to traditional event-based analyses[5,7]. The rate of MD progression has also been shown to be predictive of event-based progression[7–9], an outcome generally accepted by regulatory bodies[30].

Owing to the considerable test-retest variability of VF tests[16–18], often dependent on patients' performance[15], imaging outcomes have been proposed as a more robust alternative to functional testing[12,21]. However, translating loss of cpRNFL thickness measured by OCT into functional loss is challenging. So far, standard OCT metrics have not been shown to be superior to VF as trial outcomes[12]. Moreover, the hypothesized advantage of structural endpoints is predicated upon their higher repeatability. However, good repeatability alone does not guarantee better detection of progression, especially when not considered in the context of a the measurements dynamic range, which might be limited by the floor effect[1,2,20].

Characterizing the interplay between structural and functional progression is challenging but crucial to develop realistic models of glaucoma progression, necessary for sample size calculations. In a recent analysis of UKGTS data[2], we have used empirical data from a clinical trial to show that the rate of progression and proportional effect of IOP are very similar for structural and functional loss once the differences in scaling and the structural floor effect are taken into account. We used a model designed to characterize the distribution of true rates of functional and structural progression, minimizing the effect of test variability on measured rates and, for VF tests, the effect of learning[13]. Taken together, those results support a description of progression based on a proportional loss of retinal ganglion cell axons. This would translate to a linear decay in a logarithmic (dB) scale for both structure and function. The constant proportional effect of IOP on the rate of structural and functional progression (i.e. a constant change in dB/year per mmHg) also justifies the modelling of the same 30% treatment effect for both metrics[13]. It however is possible that IOP-independent neuroprotective treatments might behave differently for structure and function, also based on their specific mechanism of action. This would need to be better characterized when data on non-IOP neuroprotective treatments become widely available. Our simulations also assumed a moderate correlation of 0.75 between the true rates of MD and cpRNFL progression, based on our empirical results from UKGTS[2]. A sensitivity analysis assuming a lower correlation (0.4) was also performed, and is provided as **supplementary material**. The results were largely similar to our main simulations, although the combined outcome performed marginally better.

Our modelling also allowed us to integrate the non-linear effect of the structural floor (see **Methods**). It is important to clarify that there is a substantial difference between the measurement floor in VF and OCT data. In the VF, the floor is the result of censoring the

measurement at 0 dB, although also likely close to the 'true' floor[31,32]; in contrast, the structural floor is an intrinsic property of the tissue being measured, whose minimum is biased by the presence of non-neural tissue[1,2,20]. For the VF, this means that global measurements, such as the MD, are only affected once the floor has been reached at one or more locations. The rate of proportional loss estimated for the cpRNFL is instead always distorted by the offset introduced by non-neural tissue in the measured thickness. For example, a change from 100 µm to 70 µm is a 30% reduction in cpRNFL thickness. However, subtracting the average floor (56.7 µm in our estimates) from both values, would imply a 70% loss in neural tissue. This effect was captured in our simulations (**Figure 1**). This also explains why simply log-transforming the data would not address the influence of the floor effect. This also implies that, differently from VFs, restricting the selection to patients with an earlier baseline damage, such as with an MD ≥ -10 dB, would have a limited impact in addressing this issue, even when the floor level is not reached for the duration of the trial. To show this, we have performed a **supplementary analysis**, restricting the inclusion to patients with early damage (MD ≥ -6 dB), with little change to our results despite some improvement in the power of cpRNFL. It should be noted that the vast majority of the originally selected sample (85 / 107 eyes) was already in this category, so these results are unsurprising. Methods exist to address the censoring floor in VF data[31], they have not yet been developed for structural measurements. Addressing this issue will likely involve a customized estimate of the floor based on the specific anatomy of each eye and might prove challenging. The measured rates of progression for structure and function are also influenced by how the summary metrics are calculated (average of dB values for MD, dB transformation of the average thickness for cpRNFL[2]). However, defining the treatment effect in proportional terms largely eliminates the influence of this discrepancy. This is shown by our supplementary analysis assuming the same baseline rate and the same treatment difference (in dB/year) for both structure and function, which confirmed our main results (**supplementary material**).

Our results have important implications for clinical trial design. Our simulations show a higher statistical power with MD based outcomes as opposed to cpRNFL. This is in contrast with Wu et al.[12], who showed very similar power for the two metrics. Our sample was selected to be similar in terms of baseline damage (MD ≥ -10 dB), but other differences could explain the discrepancy. The main difference is that Wu et al.[12] imposed a 30% reduction in both the rate of MD progression and the linear rate of cpRNFL loss. In our simulation, the effect of treatment was applied to the proportional rate of loss in both structure and function. This was justified by the very similar proportional effect of IOP in the UKGTS cohort[2]. Critically, the simplification adopted by Wu et al. is able to approximate the average change (the treatment effect was approximately 28% in linear scale for cpRFNL, see **Results** and **Figure 4**), but fails to capture the non-linear behavior of cpRNFL thickness change within the same eyes as it progresses and the variability across eyes with different levels of initial cpRNFL loss[20]. Wu et al.[12] also showed that a combined outcome would be

more powerful than either in isolation. This is in partial agreement with our results: despite a small improvement in statistical power, there was little difference compared to only using the MD (**Figure 3** and **Table 2**). However, the power of a combined outcome was sensitive to some of our assumptions (see **supplementary material**), showing marginally better power when the MD variability was increased or the correlation in the true rates of structure-function progression was reduced. It should be noted that this combined outcome would propose two alternative statistical hypotheses. We do not think that this framework would be easily accepted by a regulatory body, because if would prevent the definition of a clear primary outcome. A more promising option could be the integration of structural data to refine the assessment of VF progression[33,34] or to improve the precision of the VF test itself[35]. Of course, different simulated effect sizes would provide different sample size requirements. A 30% treatment effect was chosen for comparison with previous literature, in which it is often reported as both clinically meaningful and detectable with practically achievable sample sizes in neuroprotection RCTs[5,6,12]. Note that different effect sizes would not change the power comparison between the two outcomes.

Our analyses also address other important issues in the quantification of progression for clinical trials. Our simulations capture many of the complex features of test variability and the correlations between structural and functional metrics. Taking advantage of our well curated test-retest data collected over a short period of time, we modeled both the systematic change in variability with the level of VF loss and retained the heterogeneous variability of individual eyes. Our simulations also replicated the correlation of pairs of structural and functional measurements obtained on the same visit. Variability in structural and functional measurements is expected to be largely independent (patients performing poorly on a VF on one day would not necessarily exhibit a similar fluctuation in their structural measurement). Our data generally confirm this expectation, because the correlation between structural and functional residuals was, on average, close to zero (see **Results**). This indicates that such correlations could be disregarded in future modelling. This has implication for clinical practice as well, because it would allow the use of structural and functional assessments as independent metrics of glaucoma progression. It should be noted, however, that long-term correlations might still exist, especially when these are caused by changes in media opacity (such as dry eye and development of cataract or corneal opacities).

Another important aspect explored in our analysis is the effect of correlations within clusters of test repeats performed on the same visit. Medeiros et al. have shown a correlation of approximately 0.2 for both VF and OCT results for tests taken close together, but not on the same day[24]. The correlations were expected to be higher for VFs performed on the same day, because fluctuations in performance are likely to affect all the tests taken on a given day. This was confirmed by the data from the UKGTS cohort, in which the correlation was 0.33. Clarifying the impact of these correlations on the statistical power and the accuracy of the estimates from LMMs is crucial, since these are often neglected in most implementations[7–

[9,13]. Conveniently, LMMs can be easily modified to account for these correlations (see **Methods**), which were replicated in our simulated data. However, our results show no difference in statistical power and accuracy of the estimates when modelling these correlations (see **Figure 3**, **Figure 4** and **Table 2**), suggesting a small impact for clinical trial results for testing schedules like the one in the UKGTS. It should be noted, however, that the impact of these correlations would greatly vary based on the number of tests per cluster, the number of clusters in the series and their collocation within the testing schedule (large clusters at the beginning or the end of the trial would have a strong leverage on the slopes). Interestingly, the extended LMMs were able to correctly estimate, on average, the correlations in the simulated data (**Figure 4**), and might be a promising approach to evaluate their effect on different study designs.

Despite its relative complexity, our realistic model is mostly characterized by a series of parameters, provide in the Methods, Figures and Tables, that can be used to replicate our simulations. This has implications beyond the design of clinical trials, because simulations have become widespread tool to assess the clinical effectiveness of global OCT and VF metrics in clinical monitoring. These have often relied on simplistic assumptions, especially when modelling cpRFNL progression[12,21,36].

One limitation of our model is the lack of characterization of specific subgroups, the diversity of which is known to have a large impact on structural metrics[36–38]. This is, however, mostly a limitation of the available data rather than the methodology. Better characterization of these sources of variability could be easily integrated into our framework and improve the accuracy of the results. A limitation of our test-retest dataset is the inability to inform about long-term variability, which might be larger than short-term[39]. However, long-term fluctuations can only be evaluated over a long period of time, during which progression cannot be excluded, compromising the accuracy of their quantification. In UKGTS, the overall residual standard deviation (using the same longitudinal LMM used to estimate the GVE in the methods) was 1.19 dB for an average MD of -4 dB. The prediction from our variability model at the same MD would be 0.95 dB. Despite being similar, we performed a set of simulations in which we proportionally increased the predicted variability for MD to match the expected long-term variability from UKGTS (**supplementary material**). Despite not changing the cpRNFL variability in a similar way, the rate of MD progression still showed better statistical power. There was, however, a bigger advantage in using the combined outcome compared to our main results.

Naturally, the limited number of patients in our test-retest cohort meant that the same eyes, many of which were pairs from the same patient, had to be sampled multiple times in our simulations. While this procedure does, on average, replicate the distribution of the data, it does not generate fully independent eyes. However, the impact on the results would be small, because the correlated rates of MD and cpRNFL progression were generated independently for each simulated eye. It should also be mentioned that other analyses could

focus on point-wise progression for VF and sectoral changes for cpRNFL, as well as different OCT metrics, such as those based on detailed maps of the macular ganglion cell layer and RNFL. These might change the results of our power calculation. However, MD rates of progression are now preferred to point-wise event-based analyses since they have shown better statistical power[5,7]. This makes the average cpRNFL a good candidate for a fair comparison of the statistical power between structural and functional progression. In the current implementation, our simulations do not model the systematic correlation between baseline damage and rates of progression. While this would not affect our power calculations, it would have an important effect when exploring the impact of different inclusion criteria.

One important caveat is that all our calculations assume a model developed from a single trial. The UKGTS is peculiar in that it provides data on structural and functional glaucoma progression in patients without treatment or treated without escalation. Very few trials have collected similar data[4], making a full independent validation difficult. However, given the broad range of disease and IOP in UKGTS, its results are likely to be generalizable. Another important aspect to consider is that the model assumes the same average neuroprotective effect regardless of the level of damage. The predictions of the model are interpreted as indicating a consistent proportional loss of retinal ganglion cell axons and bodies, leading to a similar proportional rate of true structural and functional loss. We do not have reason to think that this proportional effect would change systematically across levels of damage and this interpretation is consistent with previous modelling based on histology from Harwerth et al.[40] Finally, the model allows disagreement between structural and functional progression, assuming a within-eye correlation of 0.75[2]. This, however, does not take into account systematic sources of structure-function discrepancy, which might influence these correlations. This influence would be difficult to quantify in practice, because of the confounding effect of the structural floor and test-retest variability.

In conclusion, our results show no advantage in using cpRNFL over MD as an outcome for clinical trials and a limited impact of their combination. Future efforts should focus on improving the analysis of cpRNFL change and on more effectively integrating the information from structural and functional data to improve statistical power.

Figure captions

Figure 1. Example of a simulation for an individual eye. A pair of correlated true rates of structural and functional progression (in dB) is sampled from their respective exponential distribution (top-left). The linear rate is applied directly to simulate the functional progression (bottom-left). The structural rate is converted in linear units, considering the structural floor as explained in the text (top-right). The residuals are simulated based on the correlation observed in the test-retest data for the selected eye (-0.55 in this example, i.e. anticorrelated). The black lines and dots represent the true and simulated values, the red

dots represent the residuals. MD = Mean Deviation; OCT = Optical Coherence Tomography; cpRNFL = circum-papillary Retinal Nerve Fibre Layer.

Figure 2. In all plots, the lighter colored points indicate eyes that were excluded from the simulations, because their MD was < -10 dB. They were however included in these calculations. The top panel shows the structure-function model used to estimate the distribution of floor values (dashed line and blue Gaussian on the left). The structure-function relation is shown with a solid black line. The two bottom panel show the variability model for the MD (left). The data are shown in blue, the model predictions are in black. No model was used for the structural data. The axes are in logarithmic steps. One eye had a very large cpRNFL test-retest variability (standard deviation = 17.15 microns). This was excluded when attempting the model fits described in the methods, but was used in the simulations. MD = Mean Deviation; cpRNFL = circum-papillary Retinal Nerve Fibre Layer; SD = Standard deviation.

Figure 3. Power curves for a 30% effect with the different outcomes. Results obtained with two types of linear mixed effect models are reported, one neglecting and one modelling the correlations among test results obtained on the same visit. The shaded areas represent the 95% confidence bands. MD = Mean Deviation; cpRNFL = circumpapillary retinal nerve fiber layer.

Figure 4. Parameter estimates obtained from the models during the simulations. The error bars represent ± one standard deviation. Results obtained with two types of linear mixed effect models are reported, one neglecting and one modelling the correlations among test results obtained on the same visit. The correlation values are only estimated with the second type of model. MD = Mean Deviation; cpRNFL = circumpapillary retinal nerve fiber layer.

References

1. Hood DC, Kardon RH. A framework for comparing structural and functional measures of glaucomatous damage. *Progress in retinal and eye research*. 2007;26(6):688-710. doi:10.1016/j.preteyeres.2007.08.001

2. Montesano G, Rabiolo A, Ometto G, Crabb DP, Garway-Heath DF. Relationship between intraocular pressure and the true rate of functional and structural progression in the united kingdom glaucoma treatment study. *Investigative ophthalmology & visual science*. 2025;66(1):32. doi:10.1167/iovs.66.1.32

3. Garway-Heath DF, Crabb DP, Bunce C, et al. Latanoprost for open-angle glaucoma (UKGTS): A randomised, multicentre, placebo-controlled trial. *Lancet (London, England)*. 2015;385(9975):1295-1304. doi:10.1016/S0140-6736(14)62111-5

4. Heijl A, Leske MC, Bengtsson B, Hyman L, Bengtsson B, Hussein M. Reduction of intraocular pressure and glaucoma progression: Results from the early manifest glaucoma

trial. *Archives of ophthalmology (Chicago, Ill : 1960)*. 2002;120(10):1268-1279. doi:10.1001/archopht.120.10.1268

5. Wu Z, Crabb DP, Chauhan BC, Crowston JG, Medeiros FA. Improving the feasibility of glaucoma clinical trials using trend-based visual field progression endpoints. *Ophthalmology Glaucoma*. 2019;2(2):72-77. doi:10.1016/j.ogla.2019.01.004

6. Montesano G, Quigley HA, Crabb DP. Improving the power of glaucoma neuroprotection trials using existing visual field data. *American journal of ophthalmology*. 2021;229:127-136. doi:10.1016/j.ajo.2021.04.008

7. Montesano G, Garway-Heath DF, Rabiolo A, De Moraes CG, Ometto G, Crabb DP. Validating trend-based end points for neuroprotection trials in glaucoma. *Translational vision science & technology*. 2023;12(10):20. doi:10.1167/tvst.12.10.20

8. Medeiros FA, Jammal AA. Validation of rates of mean deviation change as clinically relevant end points for glaucoma progression. *Ophthalmology*. 2023;130(5):469-477. doi:10.1016/j.ophtha.2022.12.025

9. De Moraes CG, Lane KJ, Wang X, Liebmann JM. A potential primary endpoint for clinical trials in glaucoma neuroprotection. *Scientific reports*. 2023;13(1):7098. doi:10.1038/s41598-023-34009-x

10. Malik R, Swanson WH, Garway-Heath DF. 'Structure-function relationship' in glaucoma: Past thinking and current concepts. *Clinical & experimental ophthalmology*. 2012;40(4):369-380. doi:10.1111/j.1442-9071.2012.02770.x

11. Gardiner SK, Mansberger SL, Fortune B. Time lag between functional change and loss of retinal nerve fiber layer in glaucoma. *Investigative ophthalmology & visual science*. 2020;61(13):5. doi:10.1167/iovs.61.13.5

12. Wu Z, Medeiros FA. Sample size requirements of glaucoma clinical trials when using combined optical coherence tomography and visual field endpoints. *Scientific reports*. 2019;9(1):18886. doi:10.1038/s41598-019-55345-x

13. Montesano G, Crabb DP, Wright DM, Rabiolo A, Ometto G, Garway-Heath DF. Estimating the distribution of true rates of visual field progression in glaucoma. *Translational vision science & technology*. 2024;13(4):15. doi:10.1167/tvst.13.4.15

14. McNaught AI, Crabb DP, Fitzke FW, Hitchings RA. Modelling series of visual fields to detect progression in normal-tension glaucoma. *Graefe's archive for clinical and experimental ophthalmology = Albrecht von Graefes Archiv fur klinische und experimentelle Ophthalmologie*. 1995;233(12):750-755. doi:10.1007/BF00184085

15. Bryan SR, Eilers PHC, Lesaffre EMEH, Lemij HG, Vermeer KA. Global visit effects in point-wise longitudinal modeling of glaucomatous visual fields. *Investigative ophthalmology & visual science*. 2015;56(8):4283-4289. doi:10.1167/iovs.15-16691

16. Artes PH, Iwase A, Ohno Y, Kitazawa Y, Chauhan BC. Properties of perimetric threshold estimates from full threshold, SITA standard, and SITA fast strategies. *Investigative ophthalmology & visual science*. 2002;43(8):2654-2659. https://www.ncbi.nlm.nih.gov/pubmed/12147599

17. 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. *Investigative ophthalmology & visual science*. 2000;41(2):417-421. https://www.ncbi.nlm.nih.gov/pubmed/10670471

18. Wall M, Doyle CK, Zamba KD, Artes P, Johnson CA. The repeatability of mean defect with size III and size v standard automated perimetry. *Investigative ophthalmology & visual science*. 2013;54(2):1345-1351. doi:10.1167/iovs.12-10299

19. Gardiner SK, Swanson WH, Mansberger SL. Long- and short-term variability of perimetry in glaucoma. *Translational vision science & technology*. 2022;11(8):3. doi:10.1167/tvst.11.8.3

20. Tomita R, Rawlyk B, Sharpe GP, et al. Progressive changes in the neuroretinal rim and retinal nerve fiber layer in glaucoma: Impact of baseline values and floor effects. *Ophthalmology*. 2024;131(6):700-707. doi:10.1016/j.ophtha.2023.12.032

21. Wu JH, Moghimi S, Walker E, et al. Time to glaucoma progression detection by optical coherence tomography and visual field in glaucoma individuals of african descent. *American journal of ophthalmology*. 2025;269:195-204. doi:10.1016/j.ajo.2024.07.020

22. Garway-Heath DF, Quartilho A, Prah P, Crabb DP, Cheng Q, Zhu H. Evaluation of visual field and imaging outcomes for glaucoma clinical trials (an american ophthalomological society thesis). *Transactions of the American Ophthalmological Society*. 2017;115:T4. https://www.ncbi.nlm.nih.gov/pubmed/29085257

23. Yohannan J, Wang J, Brown J, et al. Evidence-based criteria for assessment of visual field reliability. *Ophthalmology*. 2017;124(11):1612-1620. doi:10.1016/j.ophtha.2017.04.035

24. Medeiros FA, Malek DA, Tseng H, et al. Short-term detection of fast progressors in glaucoma: The fast progression assessment through clustered evaluation (fast-PACE) study. *Ophthalmology*. 2024;131(6):645-657. doi:10.1016/j.ophtha.2023.12.031

25. Quigley HA. Clinical trials for glaucoma neuroprotection are not impossible. *Current opinion in ophthalmology*. 2012;23(2):144-154. doi:10.1097/ICU.0b013e32834ff490

26. Quigley HA. Glaucoma neuroprotection trials are practical using visual field outcomes. *Ophthalmology Glaucoma*. 2019;2(2):69-71. doi:10.1016/j.ogla.2019.01.009

27. McKean-Cowdin R, Wang Y, Wu J, Azen SP, Varma R. Impact of visual field loss on health-related quality of life in glaucoma: The los angeles latino eye study. *Ophthalmology*. 2008;115(6):941-948.e1. doi:10.1016/j.ophtha.2007.08.037

28. Peters D, Heijl A, Brenner L, Bengtsson B. Visual impairment and vision-related quality of life in the early manifest glaucoma trial after 20 years of follow-up. *Acta ophthalmologica*. 2015;93(8):745-752. doi:10.1111/aos.12839

29. Jones L, Bryan SR, Crabb DP. Gradually then suddenly? Decline in vision-related quality of life as glaucoma worsens. *Journal of ophthalmology*. 2017;2017:1621640. doi:10.1155/2017/1621640

30. Weinreb RN, Kaufman PL. Glaucoma research community and FDA look to the future, II: NEI/FDA glaucoma clinical trial design and endpoints symposium: Measures of structural change and visual function. *Investigative ophthalmology & visual science*. 2011;52(11):7842-7851. doi:10.1167/iovs.11-7895

31. Montesano G, Garway-Heath DF, Ometto G, Crabb DP. Hierarchical censored bayesian analysis of visual field progression. *Translational vision science & technology*. 2021;10(12):4. doi:10.1167/tvst.10.12.4

32. Bryan SR, Vermeer KA, Eilers PHC, Lemij HG, Lesaffre EMEH. Robust and censored modeling and prediction of progression in glaucomatous visual fields. *Investigative ophthalmology & visual science*. 2013;54(10):6694-6700. doi:10.1167/iovs.12-11185

33. Russell RA, Malik R, Chauhan BC, Crabb DP, Garway-Heath DF. Improved estimates of visual field progression using bayesian linear regression to integrate structural information in patients with ocular hypertension. *Investigative ophthalmology & visual science*. 2012;53(6):2760-2769. doi:10.1167/iovs.11-7976

34. Medeiros FA, Leite MT, Zangwill LM, Weinreb RN. Combining structural and functional measurements to improve detection of glaucoma progression using bayesian hierarchical models. *Investigative ophthalmology & visual science*. 2011;52(8):5794-5803. doi:10.1167/iovs.10-7111

35. Montesano G, Lazaridis G, Ometto G, Crabb DP, Garway-Heath DF. Improving the accuracy and speed of visual field testing in glaucoma with structural information and deep learning. *Translational vision science & technology*. 2023;12(10):10. doi:10.1167/tvst.12.10.10

36. Beniz LAF, Jammal AA, da Costa DR, Mariottoni EB, Swaminathan SS, Medeiros FA. Impact of race and ethnicity on glaucoma progression detection by perimetry and optical coherence tomography. *Scientific reports*. 2024;14(1):30752. doi:10.1038/s41598-024-80481-4

37. Knight OJ, Girkin CA, Budenz DL, Durbin MK, Feuer WJ. Effect of race, age, and axial length on optic nerve head parameters and retinal nerve fiber layer thickness measured by cirrus HD-OCT. *Archives of ophthalmology (Chicago, Ill : 1960)*. 2012;130(3):312-318. doi:10.1001/archopthalmol.2011.1576

636 38. Poon LYC, Antar H, Tsikata E, et al. Effects of age, race, and ethnicity on the optic
637 nerve and peripapillary region using spectral-domain OCT 3D volume scans. *Translational*
638 *vision science & technology*. 2018;7(6):12. doi:10.1167/tvst.7.6.12

639 39. Urata CN, Mariottoni EB, Jammal AA, et al. Comparison of short- and long-term
640 variability in standard perimetry and spectral domain optical coherence tomography in
641 glaucoma. *American journal of ophthalmology*. 2020;210:19-25.
642 doi:10.1016/j.ajo.2019.10.034

643 40. Harwerth RS, Wheat JL, Fredette MJ, Anderson DR. Linking structure and function in
644 glaucoma. *Progress in retinal and eye research*. 2010;29(4):249-271.
645 doi:10.1016/j.preteyeres.2010.02.001

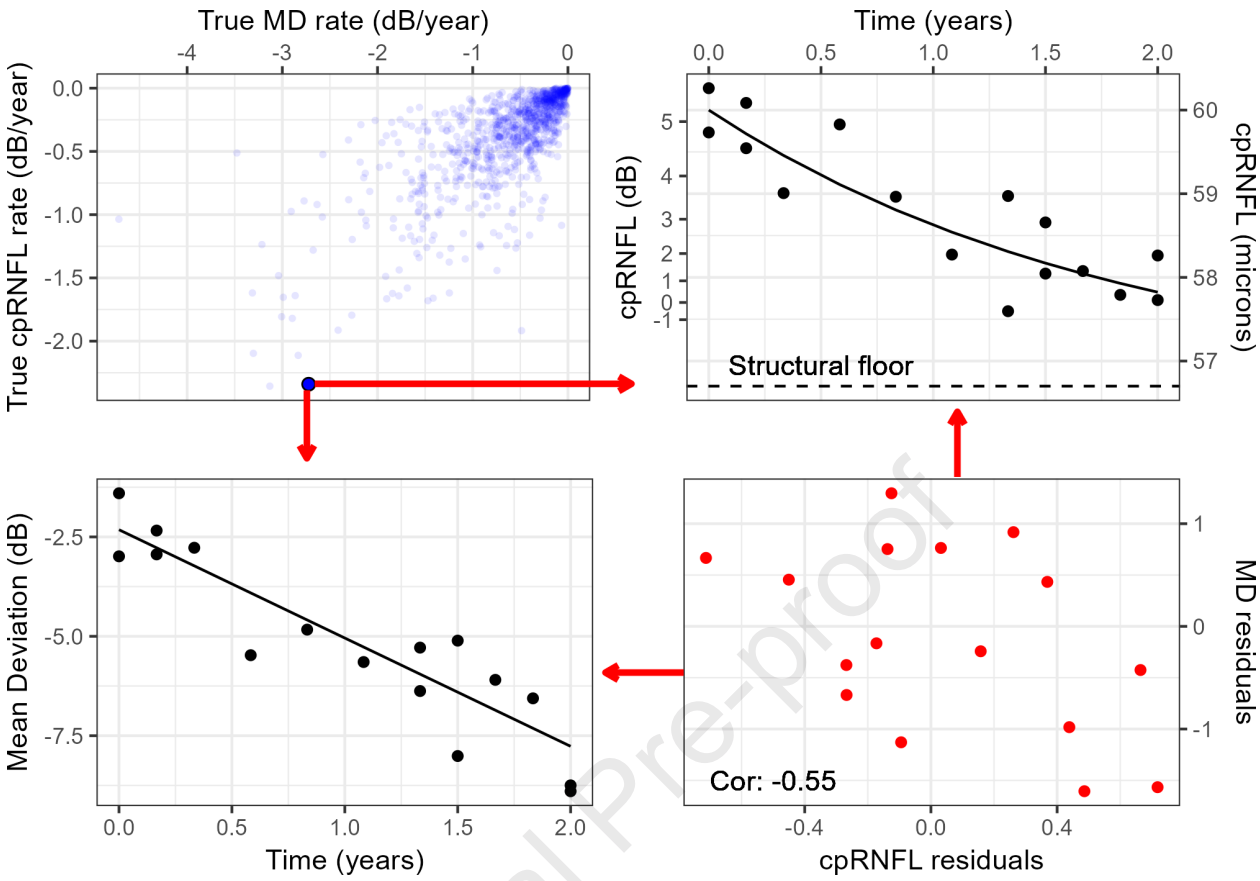
Characteristic	**Visual Field (MD, dB)**, N = 107[1]	**cpRNFL Thickness (µm)**, N = 107[1]
Average	-2.34 (-5.47, -1.02)	72.85 (63.84, 84.19)
Variability, SD	0.66 (0.49, 0.95)	0.62 (0.47, 0.82)
False positive rate, %	1.43 (0.50, 2.56)	
Quality, dB		28.22 (27.17, 29.35)
Number of tests	9 (7, 10)	9 (7, 10)

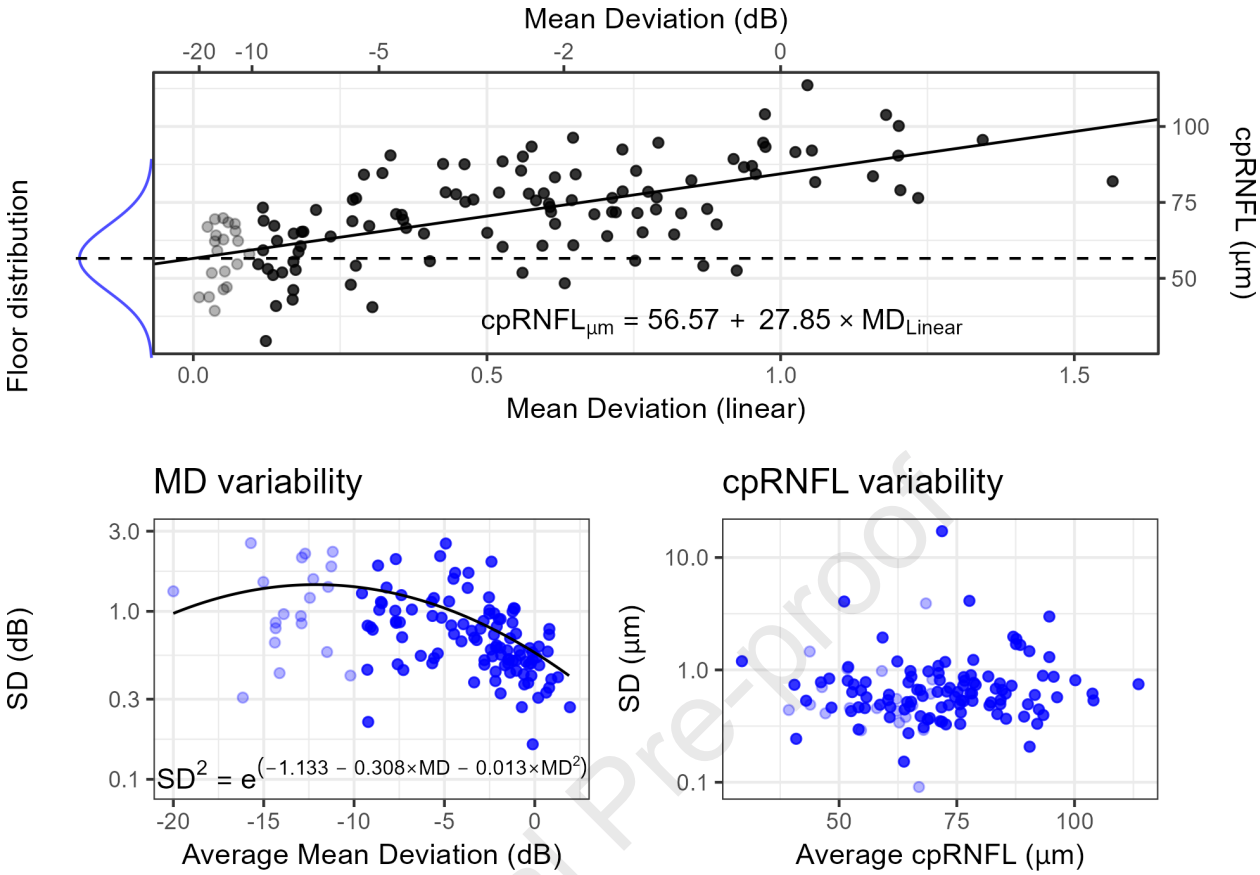
[1]Median (IQR)

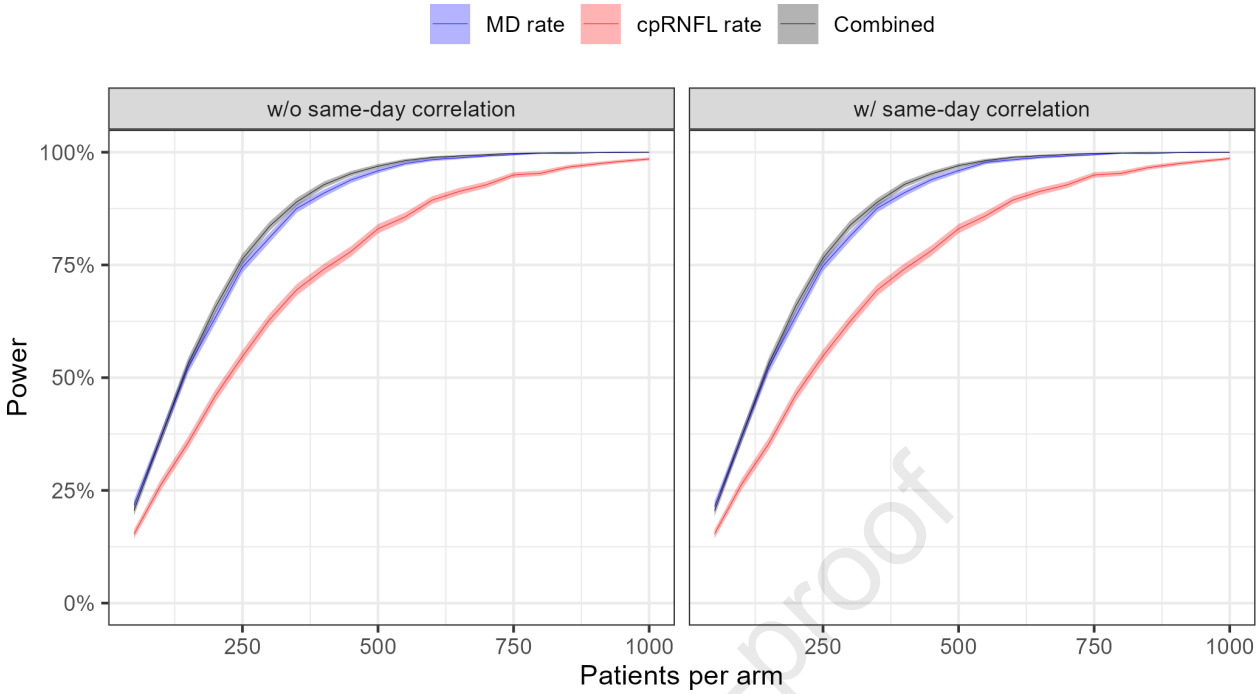
Table 1. Descriptive statistics of the sample selected for the simulations (MD ⩾ -10 dB). SD = Standard deviation; RNFL = Retinal Nerve Fibre Layer; MD = Mean Deviation; IQR = Interquartile Range.

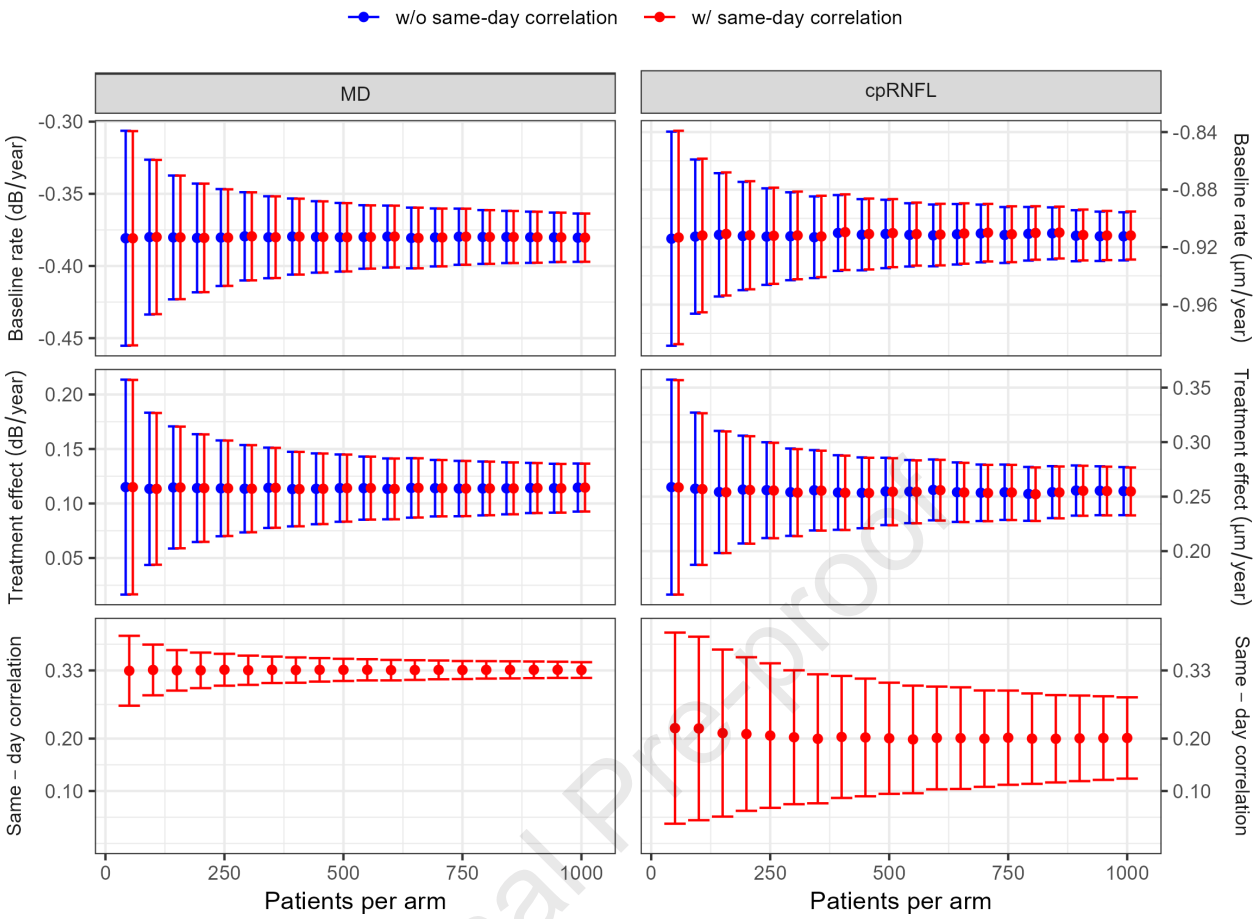
		Sample size [95% CIs]	
Model	Outcome	80% power	90% power
w/o same-day correlation	MD rate	292 [300, 283]	386 [398, 374]
	cpRNFL rate	470 [481, 459]	616 [636, 597]
	Combined	275 [283, 268]	363 [374, 352]
w/ same-day correlation	MD rate	289 [298, 281]	385 [397, 373]
	cpRNFL rate	469 [480, 457]	616 [636, 596]
	Combined	274 [281, 266]	364 [374, 353]

Table 2. Sample size calculations to detect a 30% effect ($p < 0.05$) for the different outcomes. Results obtained with two types of linear mixed effect models are reported, one neglecting and one modelling the correlations among test results obtained on the same visit. MD = Mean Deviation; cpRNFL = circumpapillary retinal nerve fiber layer; CI = confidence intervals.









Using a realistic simulation model, the rate of progression of visual field mean deviation showed higher statistical power than the rate of the average retinal nerve fiber thickness as an outcome for neuroprotection trials.