Published in final edited form as:
Ophthalmol Glaucoma. 2021 ; 4(4): 390–399. doi:10.1016/j.ogla.2020.12.001.

Predicting Global Test-Retest Variability of Visual Fields in Glaucoma

Eun Young Choi[#1,2], **Dian Li**[#1], **Yuying Fan**[#1], **Louis R. Pasquale**[3], **Lucy Q. Shen**[4], **Michael V. Boland**[4], **Pradeep Ramulu**[5], **Siamak Yousefi**[6], **Carlos G. De Moraes**[7], **Sarah R. Wellik**[8], **Jonathan S. Myers**[9], **Peter J. Bex**[10], **Tobias Elze**[1], **Mengyu Wang**[1]

[1]Schepens Eye Research Institute of Massachusetts Eye and Ear, Harvard Medical School, Boston, MA, USA

[2]Department of Ophthalmology, Duke University, Durham, NC, USA

[3]Eye and Vision Research Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[4]Massachusetts Eye and Ear Infirmary, Harvard Medical School, Boston, MA, USA

[5]Wilmer Eye Institute, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[6]Hamilton Eye Institute, University of Tennessee Health Science Center, Memphis, TN, USA

[7]Edward S. Harkness Eye Institute, Columbia University, New York, NY, USA

[8]Bascom Palmer Eye Institute, University of Miami School of Medicine, Miami, FL, USA

[9]Wills Eye Hospital, Thomas Jefferson University, Philadelphia, PA, USA

[10]Department of Psychology, Northeastern University, Boston, MA, USA

[#] These authors contributed equally to this work.

Abstract

Objective: To model the global test-retest variability of visual fields (VFs) in glaucoma.

Design: Retrospective cohort study.

Participants: 8,088 VFs of 4,044 eyes from 4,044 participants.

Methods: We selected two reliable VFs (SITA 24–2) per eye measured with the Humphrey Field Analyzer within 30 days of each other. Each VF contained fixation losses (FL) ≤ 33%, false-negative rates (FNR) ≤ 20%, and false-positive rates (FPR) ≤ 20%. Stepwise linear regression was applied to select the model that best predicts the global test-retest variability from three categories

Correspondence to: Mengyu Wang, Ph.D., Schepens Eye Research Institute, 20 Staniford Street, Boston, MA 02114, USA, mengyu_wang@meei.harvard.edu.
OTHER CONTRIBUTIONS:
The use of computationally-derived archetype visual field patterns and pointwise total deviation values strongly improved the prediction of the global test-retest variability compared with using the traditional global visual field indices alone.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

of features of the first VF: (1) base parameters (age, mean deviation [MD], pattern standard deviation, glaucoma hemifield test, FPR, FNR, FL); (2) total deviation (TD) at each location; and (3) computationally-derived VF loss patterns (archetypes). The global test-retest variability was defined as root mean square deviation (RMSD) of TD values at all 52 VF locations. Model performance was assessed using adjusted R-squared and Bayesian information criterion (BIC).

Main Outcome Measures: Archetype models to predict the global test-retest variability.

Results: The mean ± standard deviation of RMSD was 4.39 ± 2.55 dB. Between the two VF tests, TD values were more strongly correlated in central than in peripheral VF locations (intraclass coefficient range: 0.66–0.89; $p < 0.001$). Compared with the model using base parameters alone (adjusted R-squared = 0.45), adding TD values improved prediction accuracy of the global variability (adjusted R-$_{\text{squared}}$ = 0.53, $p < 0.001$) and BIC (decreased by 527; a change of > 6 represents strong improvement). Lower TD sensitivity in the outer-most peripheral VF locations was predictive of higher global variability. Adding archetypes to the base model improved model performance with an adjusted R-squared of 0.53 ($p < 0.001$) and lowering of BIC by 583. Greater variability was associated with concentric peripheral defect, temporal hemianopia, inferotemporal defect, near total loss, superior peripheral defect, and central scotoma (listed in order of decreasing statistical significance), and less normal VF and superior paracentral defect.

Conclusions: Inclusion of archetype VF loss patterns and TD values based on first VFs improved the prediction of the global test-retest variability than using traditional global VF indices alone.

Introduction

Visual field (VF) testing with standard automated perimetry (SAP) is an essential tool for diagnosing and monitoring functional progression of glaucoma. However, the performance of VF tests is prone to variability due to short-term and long-term fluctuations.[1–7] Assessing the variability of VF measurements is thus critical for accurate diagnosis and monitoring of disease progression. Previous studies have shown factors such as age, glaucoma severity, visual acuity, and cognitive decline to be associated with test-retest variability.[8–11] In clinical practice, reliability of a test is typically estimated using the false positive rate (FPR), false negative rate (FNR), and fixation losses (FL).[12] However, studies have suggested that these indices may have limitations.[13–16] While the Swedish Interactive Thresholding Algorithm (SITA) for mapping the island of vision has been widely accepted in clinical practice, this strategy does not devote any metrics to measuring test-retest variability. Metrics, comparable to short-term fluctuation used by the prior standard thresholding algorithm, are needed to assess test-retest variability of SITA tests in order to improve the clinical interpretation of VFs.

Previous work has examined differences in test-retest variability by VF location: in general, variability was found to increase with eccentricity and worse VF sensitivity.[1,5,6,10,17–21] While these studies have reached a consensus that peripheral locations generally have greater test-retest variability compared with paracentral locations, no models have been developed to elucidate where exactly, and with how much impact quantitatively, VF loss in peripheral and paracentral regions is associated with the global test-retest variability.

Furthermore, prior studies also reported that point-wise variation was greatest in the superior fields compared with the inferior fields.[10,18] These findings suggest that regional VF defects may affect the global variability depending on their location. However, to our knowledge, there has been no systematic investigation on the relationship between regional VF loss patterns and test-retest variability. An algorithm to assess VF variability based on patterns of VF loss would help clinicians to interpret VF results and to decide when a confirmatory test is needed.

In this study, we utilized a large multicenter dataset to study the impact of VF features on the global test-retest variability, defined as the root mean square deviation (RMSD) of 52 sensitivity values between two repeated tests. We quantify the impact of each test location on the global variability, and further augment our analysis with clinically validated VF loss patterns,[22] termed VF archetypes (Figure 1A). We used an unsupervised artificial intelligence method termed archetypal analysis[23] to determine the VF archetypes. The VF archetypes, which are objectively identified and quantified, have demonstrated ability to improve the assessments of glaucoma diagnosis[24] and progression.[25] In this work, we examine whether the use of archetypal VF loss patterns can enhance the assessment of test-retest variability, compared with using the traditional global VF indices and the entire total deviation (TD) map of the first VF. The purpose of our study is to provide a clinical tool to help clinicians better interpret VF loss under measurement noise based on improved prediction of test-retest variability.

Methods

The VFs used for this study were obtained through the Glaucoma Research Network (GRN), a multi-center consortium of academic institutions listed below. This retrospective study was approved by the institutional review boards (IRB) of each institution and adheres to the Declaration of Helsinki and all federal and state laws. Because of the retrospective nature of this study, the IRB waived the need for informed consent of patients.

Participants and Data

Our dataset consisted of Swedish interactive thresholding algorithm (SITA) standard 24–2 VFs measured with the Humphrey Field Analyzer II (HFA; Carl Zeiss Meditec, Dublin, CA). We included all reliable VF data from the Glaucoma Research Network consortium regardless of patient diagnostics, consisting of Massachusetts Eye and Ear, Wilmer Eye Institute, New York Eye and Ear Infirmary of Mount Sinai, Bascom Palmer Eye Institute, Wills Eye Hospital, Columbia University, and Hamilton Eye Institute. From the entire GRN VF dataset, all eyes with reliable VFs repeated within 30 days of each other were initially included. If both eyes of an individual met the criteria, one eye was selected at random to avoid any potential bias related to inter-eye symmetry. If more than two VF tests conducted within the 30-day period were available, the reliable first and last retest VFs were selected. We used the following reliability criteria to model the test-retest variability of VFs considered to be reliable in clinical practice: fixation losses ≤ 33%, false-negative rates ≤ 20%, and false-positive rates ≤ 20%.[12,26,27] Furthermore, all eyes with MD difference greater than 10 dB between the first and retest VFs were excluded, as these eyes were suspected to have undergone eye procedures, have unexpected damage to the eye, or otherwise considered to be outliers. This exclusion criterion was based on a previous finding that intra-individual test-retest variability of MD was less than 9 dB.[3]

Feature Extraction

All VF features were extracted from the first VFs, as our study aimed to predict the global test-rest variability given a first VF test in clinical practice.

The following VF indices were extracted: MD, PSD, GHT, FPR, FNR, and FL. Additionally, the total deviation (TD) values at each of the 52 locations of the 24–2 VF were extracted and used to derive the VF loss archetype patterns, which were decomposed into 16 computationally-derived VF archetypes with a weighting coefficient for each archetype as previously described (Figure 1A).[23] Briefly, the 16 VF archetypes were identified by an unsupervised artificial intelligence method, termed archetype analysis, based on more than 13,000 reliable VFs. The coefficients for each archetype, which sum to 100%, represent various global and regional VF loss patterns (Figure 1B). Of the 16 archetypes, 9 represent clinically recognizable patterns of glaucomatous field loss, validated by a clinical correlation study[22]: altitudinal VF loss (archetypes 8 and 13), partial arcuate defects (archetypes 9, 10, and 16), nasal step (archetypes 3 and 5), and paracentral defects (archetypes 14 and 16). Archetype 2 was associated with both ptosis and glaucomatous VF loss. Archetype 1 shows the normal VF. All other archetypes typically represent non-glaucomatous defects, such as temporal and nasal hemianopia (archetypes 12 and 15).

Statistical Analysis

All statistical analyses were performed using R 3.6.2.[28] First, the correlation between the first and retest VFs was examined by calculating the intraclass correlation coefficients (ICC) and mean absolute differences between $\mathrm{MD}_{\mathrm{first}}$ and $\mathrm{MD}_{\mathrm{retest}}$, as well as between each pair of $\mathrm{TD}_{\mathrm{first}}$ and $\mathrm{TD}_{\mathrm{retest}}$ at 52 locations. Pearson correlation was used to examine the association between the global test-retest variability and each individual parameter (i.e. age, MD, PSD, GHT, FPR, FNR, FL, and each VF archetype).

The global test-retest variability was assessed by the root mean square deviation (RMSD) of the TD values. We define RMSD as the square root of average square of TD differences over all 52 locations between two repeated measurements, as shown in the following equation. The RMSD reduces skewness of the outcome measure and has been utilized in prior research as a rigorous measure of the global test-retest variability.[9]

$$RMSD = \sqrt{\frac{\sum_{i=1}^{52}\left(TD\,at\,i^{th}\,location\,[firstVF] - TD\,at i^{th}\,location\,[lastVF]\right)^{2}}{52}}$$

Stepwise linear regression[29] was performed to select the optimal combination of variables that predicts the global test-retest variability based on the Bayesian information criterion (BIC).[30] Three predictive models were calculated based on different subsets of independent variables: 1) The “base model,” selected from age, MD, PSD, GHT, FPR, FNR, and FL, 2) the “TD + base model,” selected from the base variables *plus* TD values at each VF location, and 3) the "archetype + base model," selected from the base variables *plus* the 16 VF archetypes. For the purposes of better model graph visualization, age was used in the unit of every decade and MD, PSD, and TD values were used in the unit of every 10 dB. Model performance was quantified using adjusted R^2 and BIC values. Bootstrapping was applied to calculate the confidence intervals of R^2 values. The relative importance of each parameter to its respective model was assessed using the magnitude of BIC increase when that parameter was removed from the optimal model. *P* values were adjusted for multiple comparisons. A *P* value < 0.05 was considered statistically significant.

Results

We selected 4,044 eyes of 4,044 patients (mean ± standard deviation [SD] of age: 61.9 ± 15.0 years) with reliable VF pairs measured within 30 days of each other. Of the selected eyes, 2,142 (53%) were right eyes, and 1,902 (47%) were left eyes. The mean ± SD of time difference between the first and retest VF measurements was 11.8 ± 11.2 days.

Figure 2 shows the distributions of MD for the first and retest VF measurements (mean ± SD: −5.3 ± 6.0 dB and −4.6 ± 5.8 dB, respectively). The mean MD difference and mean absolute difference between the two measurements was 0.7 and 1.7 dB, respectively, which significantly differed from 0 dB ($p < 0.001$ for both). The ICC for the two MD values was 0.92 ($p < 0.001$). Figure 3A shows the average TD values at the 52 VF locations ranging from −8.42 dB to −2.95 dB. The VF loss was more severe in the superonasal region and less severe in the inferotemporal locations. The ICCs for the two TD values at each of the 52 locations are shown in Figure 3B (ICC range: 0.66–0.89; $p < 0.001$ for all), and the mean absolute TD differences are shown in Figure 3C (absolute ΔTD range: 2.28 – 4.59 dB; all $p < 0.001$, differing from zero). In general, peripheral VF locations had greater TD differences and weaker correlations between the two repeated measurements compared to central VF locations.

The mean ± SD of RMSD was 4.39 ± 2.55 dB. We further calculated Pearson correlations between the global test-retest variability (RMSD of TD values) and base and archetype parameters (Figure 4A) and the TD values at the 52 locations (Figure 4B). The correlations between the global test-retest variability and base and archetype parameters ranged from −0.65 to 0.70. Among the base parameters, MD (r = −0.56) had the strongest correlation with the global variability ($p < 0.001$): greater global variability was observed with worsening VF damage. Correlations of FPR and FL were insignificant ($p > 0.05$). Except superonasal step (archetype 3), all archetypes had significant positive correlations with the global variability ($p < 0.01$). The most correlated eight archetypes (p < 0.001 for all) were normal VF(archetypes 1, r = 0.65), concentric peripheral defect (archetype 11, r = 0.38), superior altitudinal defect(archetype 8, r = 0.28), near total loss(archetype 6, r = 0.26), temporal hemianopia (archetype 12, r = 0.20), nasal hemianopia (archetype 15, r = 0.17), inferotemporal defect (archetype 9, r = 0.17) and inferior altitudinal defect (archetype 13, r = 0.16). In comparison, the correlations between the TD values at the 52 locations and the global variability ranged from −0.53 to −0.32 (p < 0.001 for all). VF loss at the peripheral locations was more related to the global test-retest variability than VF loss at the paracentral locations.

Based on these findings, stepwise linear regression was performed to select models that best predict the global variability (Figure 5) by removing redundant parameters. In the "base model," greater global variability was observed ($p < 0.001$) with worse MD, older age, higher PSD, positive GHT, and higher FNR (Figure 5A). FPR and FL did not remain in the optimal combination of parameters. When TD values were included in the model, sensitivities at four VF locations on the 24–2 pattern, primarily in the paracentral zone, were positively ($p < 0.001$) associated with the global variability. In contrast, sensitivities at eight VF locations, primarily in the outer-most peripheral VF zone, were negatively associated ($p < 0.004$) with the global variability (Figure 5B).

When the VF loss archetype patterns were included in the model ("archetype [AT] + base model"), greater weighting coefficients for concentric peripheral defect (AT 11), temporal hemianopia (AT 12), inferotemporal defect (AT 9), near total loss (AT 6), superior peripheral defect (AT 2), and central scotoma (AT 7) were associated ($p < 0.001$) with greater global variability, while greater weighting coefficients for the normal VF (AT 1) and superior paracentral defect (AT 14) were associated ($p < 0.001$) with less global variability (Figure 5C). The archetypes are listed in order of decreasing statistical significance as measured by BIC. For each model, the highest increase in BIC was noted when PSD was removed from the model (Figure 5D–F). Among the archetypes (Figure 5F), concentric peripheral defect (AT 11) was associated with the highest increase in BIC followed by temporal hemianopia (AT 12), inferotemporal defect (AT 9) and near total loss (AT 6). All regression model coefficients and BIC values can be found in Table S1.

Table 1 shows the performance of the selected models using adjusted R^2 and BIC values to predict the global variability. Across all glaucoma severities, the adjusted R^2 of the "base model," "TD + base model," and "AT + base model" was 0.45 (95% CI, 0.43 – 0.48), 0.53 (0.50 – 0.56), and 0.53 (0.50 – 0.56), respectively ($p < 0.001$ for all comparisons). The BIC of the "TD + base model" was lower by 527 compared to the "base model." The BIC of the "AT + base model" was lower by 583 compared to the "base model", and lower by 56 compared to the "TD + base model." BIC lowering more than 6 signifies strong model improvement.[31]

Figure 6 shows example VFs for which our model predicts different levels of variability based on archetypal analysis. In Figure 6A, the normal VF pattern (AT 1, 40%) and superior paracentral loss (AT 14, 10%) comprise the majority of the archetype composition in the first VF. Our model shows ATs 1 and 14 to be negatively associated with variability, so this VF would be expected to have relatively low variability. Indeed, the predicted and actual RMSD were 2.78 dB and 2.36 dB, respectively, which both lie in the lowest 25th percentile of the RMSD distribution. In Figure 6B, superior peripheral defect (AT 2, 48%) and inferotemporal defect (AT 9, 16%) comprise the majority of archetypes in the first VF. Since our model shows ATs 2 and 9 to be positively associated with variability, this VF would be expected to have relatively high variability. The predicted and actual RMSD were 6.54 dB and 6.64 dB, respectively, which are both above the 75th percentile of the RMSD distribution. The base parameters, notably MD and PSD, were also consistent with our model prediction: MD was lower and PSD was higher in Figure 6B compared to 6A.

Discussion

Assessment of test-retest variability is important for VF interpretation, but current reliability metrics have limitations.[13–16] We demonstrate that VF features can be used in a model to accurately predict the short-term global variability in VFs within the accepted range for standard reliability measures. The use of computationally-derived archetype VF patterns and pointwise TD values strongly improved the prediction of the global test-retest variability compared with using the traditional global VF indices alone. VFs with certain defects are subject to greater global variability. Note that, a VF can have several different VF loss archetypes at the same time. Furthermore, some of the VF archetypes that are highly predictive of the global variability are not necessarily glaucomatous (e.g. ATs 11, 2 and 12). Thus the presence of these VF archetypes in a VF measurement (implying greater global variability) could affect whether clinicians should trust the presence of coexisting glaucomatous VF defects in the same VF measurement.

In our study, we used the root mean square deviation (RMSD) of TD values to measure the global variability between two VF tests. Previous works have measured the global variability in different ways, including MD difference or RMSD. Whereas MD difference is a simple and accessible way to conceptualize variability, it may underestimate the global variability because it is weighted heavily in the center. The RMSD reduces skewness and is more heavily weighted in the periphery where variability is known to be higher.

In the base model selected from the global VF indices alone, MD was negatively associated with the global test-retest variability, indicating greater variability with increasing overall field loss. This result is consistent with previous studies showing increased variability as a function of glaucoma severity.[1,20,21] Older age was associated with greater global variability, in line with prior work,[8] though its relative importance in the model was minimal. In terms of reliability indices, FNR was positively associated with the global variability, whereas FPR and FL did not remain in the optimal combination of variables. These findings are in agreement with Matsuura *et al.*[9] and Omodaka *et al.*,[32] who found that FNR was a good predictor of VF reliability, but FPR and FL were not. We also note that FL failed to remain significant in any of our models, consistent with Yohannan *et al.*'s[16] conclusion that FL had no meaningful impact on VF reliability.

Our study expands upon prior research by examining the impact of glaucoma hemifield test (GHT) and pattern standard deviation (PSD) on variability. GHT and PSD were both positively associated with the global variability, which makes sense given that they are markers of VF abnormality, and more damaged points show greater global variability in sensitivity. Interestingly, of the parameters selected in our models, PSD was the most important, as demonstrated by its association with the largest increase in BIC in all three of our models (Figure 5D–F). The relative importance of PSD was greater than that of traditional reliability indices (e.g. FPR, FNR) and MD, suggesting that the irregularity of VF depression, or the degree of focal VF defects, can be highly predictive of the global test-retest variability. The importance of PSD was also highlighted in the Ocular Hypertension Treatment Study, which found PSD to be a risk factor for development of glaucoma from ocular hypertension.[33]

Previous studies have shown that variability tends to be greater in peripheral than in central VF locations.[1,17,19,21] For instance, Heijl *et al.*[1] found that variability increased by distance from fixation in eyes with shallow defects. Similarly, Gardiner *et al.*[21] found that when sensitivity is near normal, variability was lower in central than peripheral locations. Furthermore, Young *et al.*[18] noted that TD differences were greatest in the superior and nasal fields. Our results correlating the pointwise TD values between the two repeat tests confirm these findings with a much larger sample size (Figure 3B, C). Building on this notion, we incorporated location-specific VF loss information to predict the global test-retest variability defined as the RMSD of TD values. Our study significantly improves upon prior work, as we elucidate the exact locations and quantify the relative impact of each location on overall variability (Figure 5B, E). As expected, we found that greater damage in the outer-most peripheral zone (including the superior and inferior peripheral, nasal and temporal VF locations) was predictive of increased global variability. On the other hand, greater damage in the paracentral zone was predictive of decreased global variability (Figure 5B). These findings suggest that VFs with greater peripheral VF loss are likely to have more short-term fluctuations, and should heighten clinical suspicion that the VF loss may represent random variation rather than a true defect. In sum, including the entire TD map significantly improved the model’s ability to predict the global variability.

Most importantly, we assessed whether certain VF loss archetype patterns were predictive of the global test-retest variability and quantified the effect (Figure 5C). One advantage of using archetypes is that they represent recognizable patterns of VF loss[22], which correspond well to retinal nerve fiber topology and therefore are more clinically interpretable. The VF archetypes represent the spatial relationship of VF loss at different VF test locations, which is not addressed by assessing individual TD values. Several archetypes were significantly correlated with the global variability. In order of descending importance, archetypes representing concentric peripheral defect (AT 11), temporal hemianopia (AT 12), inferotemporal defect (AT 9), near total loss (AT 6), superior peripheral defect (AT 2), and central scotoma (AT 7) were positively associated with variability, indicating that VFs that display these patterns of field loss are more prone to variability and may require repeat tests for confirmation. ATs 2, 9, and 11 are associated with glaucomatous field loss, whereas ATs 2 and 11 may also represent eyelid effect (e.g. ptosis) and lens rim effect in hyperopia, respectively. One feature these archetypes share in common is VF loss in the peripheral zone, which is consistent with the TD locations that were related to greater global variability (Figure 5B). In particular, AT 11 with the greatest degree of peripheral field loss was most predictive of variability among all archetypes (Figure 5F). This is worth pointing out, because concentric peripheral defect is a commonly seen VF pattern which can be caused by a lens rim artifact. Therefore, clinicians and technicians should be aware of the importance of proper lens placement during perimetry or be mindful of this artifact when a patient has a high refractive error. Patients with near total loss (AT 6), severe central loss (AT 7), or temporal hemianopia related to strokes (AT 12) may find it difficult to finish the VF test, which could explain the high variability associated with these archetypes. On the other hand, the normal VF (AT 1) and superior paracentral defect (AT 14) were negatively associated with variability. These findings also correspond to the TD model, as AT 14 covers TD locations 21 and 22 which are negatively associated with variability (Figure 5B). The reason why severe central loss (AT 7, likely due to macular diseases) and superior paracentral loss (AT 14) have opposite correlations may be because patients with AT 7 lose almost all central vision and therefore have greater difficulty performing the VF test, whereas patients with AT 14 maintain part of their central vision. In sum, our model identifies distinct VF patterns that suggest how variable a VF is and can be used to recommend clinicians to perform a confirmatory test if the predicted test-retest variability is high. To this end, a summary of VF features affecting test-retest variability is provided in Supplemental Table S3. In comparison to the TD model, using archetypes improved the model prediction even further, which may be attributed to a more precise representation of the spatial information between different test locations than the TD model, which correspond better to the retinal nerve fiber topology.[23,34]

The clinical utility of our models is illustrated in two contrasting pairs of example VFs (Figure 6). In Figure 6A, the base parameters (notably MD and PSD) and the archetypes that comprise the first VF are predictive of decreased variability according to our model. As expected, the retest VF remained stable from the first. On the other hand, in Figure 6B, the base parameters and archetypes are predictive of increased global variability; indeed, the retest VF was highly variable from the first VF. Here, we do not know whether the first or retest VF represents the true VF. However, using these VF features can provide information about the degree of measurement uncertainty. If variability is predicted to be high, clinicians should be cautious to trust the result and consider repeating the test in a few months to confirm the finding. It should be noted that all the VFs included in the study met our reliability criteria based on FPR, FNR, and FL, yet there was still a wide range of variability. Therefore, in situations of clinical uncertainty, VF features can provide an additional layer of objective data that can augment the clinician’s assessment of reliability.

While there are benefits of using a large, de-identified dataset, the lack of clinical information meant that we could not determine the exact reason the VFs were repeated. As we only included VFs that met our reliability criteria based on FPR, FNR, and FL, this suggests there may have been some clinical uncertainty in the remaining VFs that was not captured by the reliability indices alone. We speculate that the most likely reason to repeat a reliable test within 30 days would be suspected progression of VF loss, despite that some of the repeated tests may be from prior clinical trials. The modest improvement in MD may in fact signal a possible learning effect, as observed in other studies.[3,35] Nonetheless, we acknowledge the potential bias introduced by retrospectively analyzing repeated VFs, as doing so may inadvertently select for VFs with inherently high test-retest variability. Therefore, we conducted additional analyses using the same parameters from the retest VF measurement. The “AT + base model” (adjusted $R^2 = 0.32$) still performed significantly ($p < 0.001$) better than the “base model” (adjusted $R^2 = 0.29$, respectively), with an associated decrease in BIC by 153, respectively (Table S2). In the future, the study may be better designed as a prospective trial of early repeat VFs in a random population.

Strengths of our study include accuracy of the models, as indicated by the high R^2 values comparing predicted and measured variability. The large magnitude of ΔBIC also suggest substantial model improvement. Furthermore, our large multicenter dataset includes patients with a wide spectrum of glaucoma severity, while previous studies had relatively small sample sizes with a narrower spectrum of MD.[1,18,20,21] We use the RMSD of TD values, a rigorous and direct measure of overall test-retest variability, in contrast to other studies that used indirect estimates such as MD difference. In addition to using the entire TD map to predict the global variability, we strengthened our analysis by including clinically recognizable patterns of VF loss. Our archetype decomposition method is publicly available[23] and can be widely implemented. Lastly, we only included VFs that met our reliability criteria based on FL, FPR and FNR, focusing on VFs that would be of clinical interest. We show that even if a VF is considered reliable by the traditional reliability indices, certain VF features may suggest otherwise, and therefore caution should be used in interpreting such results.

There are several limitations of our study. First, the lack of clinical diagnoses in our dataset limited our ability to exclude subjects based on prior surgery or ocular disease, such as cataract or other conditions causing non-glaucomatous vision loss. We made our best attempt to exclude patients who may have undergone an eye procedure within the 30-day period or have potentially confounding conditions by excluding those with MD difference greater than 10 dB between the first and retest VFs, as suggested in a previous study.[3] Second, because of the lack of clinical data, we could not determine the exact reason the VF tests were repeated, as discussed in detail above. In addition, while we safely assume that the VF loss in our subjects is mostly due to glaucoma given the origins of our large dataset, it is possible that a small subset of the VFs came from patients with a disease process other than glaucoma or in addition to glaucoma, such as those with hemianopia defects. We believe that this more accurately represents the clinical practice, in which patients with glaucoma can have VF defects from other diseases such as age-related macular degeneration, stroke, or other central nervous system disorders. Furthermore, our results only apply to reliability parameters set in this study. The reliability thresholds may vary among providers, and using different thresholds may yield different results. In particular, we chose a conservative threshold for FN rate so that patients with higher FN rate, which can signal early glaucoma,[13] may be missed. For FP rate, more conservative criteria have been proposed: for instance, the current printout of the HFA marks > 15% FP rate as unreliable. With this criterion, we would exclude 78 more cases out of 4,044. We argue that a FP rate of 20% would increase noise level but would be unlikely to introduce any systematic error. Models using FP rate ≤ 15% as the reliability cutoff are shown in Supplemental Figure S4, which does not differ substantially from Figure 5. We therefore show that our method works with an even higher noise level (20%), which provides evidence for the robustness of our results. In addition, the RMSD does not yield location-specific variability; rather, it is an overall metric of variability, which can be simpler for clinical interpretation. There may have been other determinants of test-retest variability that were not captured in our models. Another important limitation is that gaze tracking was not evaluated in this study. Gaze tracking measures the eye position and fixation status during a VF test. It is associated with factors that can affect the quality of the VF such as dry eyes, and has been shown to be closely related to VF reproducibility.[36] Finally, it would be important to validate our model and determine its generalizability by applying it to independent datasets.

In conclusion, we demonstrate that test-retest variability can be predicted using VF features from the first test with high accuracy. Using the computationally derived VF loss patterns and pointwise TD values significantly improves the prediction of the global VF variability compared with using the traditional global VF indices alone. Clinicians can use this information to help determine which VFs may require further evaluation in situations of clinical uncertainty.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by NIH K99EY028631 (M.W.), NIH R21EY030142 (T.E., S.Y.), NIH R21EY030631 (T.E.), NIH R01EY030575 (T.E.), NIH R01EY015473 (L.R.P.), BrightFocus Foundation (M.W., T.E.), Lions Foundation (M.W., T.E.), Grimshaw-Gudewicz Foundation (M.W., T.E.), Research to Prevent Blindness (M.W., T.E.), Harvard Glaucoma Center of Excellence (M.W., T.E., L.Q.S.), the Eleanor and Miles Shore Fellowship (L.Q.S.), Departmental Grant from Research to Prevent Blindness (C.G.D.M.), NIH R01EY025253 (C.G.D.M.), the Alice Adler Fellowship (T.E.), and NIH NEI Core Grant P30EY003790 (E.Y.C., M.W., T.E.).

References

1. Heijl A, Lindgren A & Lindgren G Test-retest variability in glaucomatous visual fields. Am. J. Ophthalmol 108, 130–135 (1989). [PubMed: 2757094]
2. Flammer J, Drance SM & Zulauf M Differential Light Threshold: Short- and Long-Term Fluctuation in Patients with Glaucoma, Normal Controls, and Patients with Suspected Glaucoma. Arch. Ophthalmol 102, 704–706 (1984). [PubMed: 6721758]
3. Gillespie BW et al. The collaborative initial glaucoma treatment study: Baseline visual field and test-retest variability. Investig. Ophthalmol. Vis. Sci 44, 2613–2620 (2003). [PubMed: 12766064]
4. Katz J, Quigley HA & Sommer A Repeatability of the Glaucoma Hemifield Test in automated perimetry. Invest. Ophthalmol. Vis. Sci 36, 1658–64 (1995). [PubMed: 7601645]
5. Wall M, Woodward KR, Doyle CK & Artes PH Repeatability of automated perimetry: A comparison between standard automated perimetry with stimulus size III and V, matrix, and motion perimetry. Investig. Ophthalmol. Vis. Sci 50, 974–979 (2009). [PubMed: 18952921]
6. Gardiner SK, Swanson WH, Goren D, Mansberger SL & Demirel S Assessment of the reliability of standard automated perimetry in regions of glaucomatous damage. Ophthalmology 121, 1359–1369 (2014). [PubMed: 24629617]
7. Montolio FGJ, Wesselink C, Gordijn M & Jansonius NM Factors that influence standard automated perimetry test results in glaucoma: Test reliability, technician experience, time of day, and season. Investig. Ophthalmol. Vis. Sci 53, 7010–7017 (2012). [PubMed: 22952121]
8. Katz J & Sommer A A longitudinal study of the age-adjusted variability of automated visual fields. Arch. Ophthalmol. (Chicago, Ill. 1960) 105, 1083–6 (1987).
9. Matsuura M, Hirasawa K, Murata H & Asaoka R The relationship between visual acuity and the reproducibility of visual field measurements in glaucoma patients. Investig. Ophthalmol. Vis. Sci 56, 5630–5635 (2015). [PubMed: 26313298]
10. Blumenthal EZ et al. Evaluating several sources of variability for standard and SWAP visual fields in glaucoma patients, suspects, and normals. Ophthalmology 110, 1895–902 (2003). [PubMed: 14522760]
11. Diniz-Filho A, Delano-Wood L, Daga FB, Cronemberger S & Medeiros FA Association Between Neurocognitive Decline and Visual Field Variability in Glaucoma. JAMA Ophthalmol. 135, 734–739 (2017). [PubMed: 28520873]
12. Birt CM et al. Analysis of reliability indices from Humphrey visual field tests in an urban glaucoma population. Ophthalmology 104, 1126–1130 (1997). [PubMed: 9224465]
13. Bengtsson B & Heijl A False-negative responses in glaucoma perimetry: indicators of patient performance or test reliability? Invest. Ophthalmol. Vis. Sci 41, 2201–4 (2000). [PubMed: 10892863]

14. Sanabria O, Feuer WJ & Anderson DR Pseudo-loss of Fixation in Automated Perimetry. Ophthalmology 98, 76–78 (1991). [PubMed: 2023737]
15. Demirel S & Vingrys AJ Eye Movements During Perimetry and the Effect that Fixational Instability Has on Perimetric Outcomes. J. Glaucoma 3, 28–35 (1994). [PubMed: 19920549]
16. Yohannan J et al. Evidence-based Criteria for Assessment of Visual Field Reliability. in Ophthalmology 124, 1612–1620 (Elsevier Inc., 2017). [PubMed: 28676280]
17. Heijl A, Lindgren G & Olsson J Normal Variability Of Static Perimetric Threshold Values Across The Central Visual Field. Arch. Ophthalmol 105, 1544–1549 (1987). [PubMed: 3675288]
18. Young WO, Stewart WC, Hunt H & Crosswell H Static threshold variability in the peripheral visual field in normal subjects. Graefe's Arch. Clin. Exp. Ophthalmol 228, 454–457 (1990). [PubMed: 2227491]
19. Lewis RA, Johnson CA, Keltner JL & Labermeier PK Variability of quantitative automated perimetry in normal observers. Ophthalmology 93, 878–81 (1986). [PubMed: 3763131]
20. Chauhan BC & Johnson CA Test-retest variability of frequency-doubling perimetry and conventional perimetry in glaucoma patients and normal subjects. Invest. Ophthalmol. Vis. Sci 40, 648–656 (1999). [PubMed: 10067968]
21. Gardiner SK Differences in the relation between perimetric sensitivity and variability between locations across the visual field. Investig. Ophthalmol. Vis. Sci 59, 3667–3674 (2018). [PubMed: 30029253]
22. Cai S et al. Clinical Correlates of Computationally Derived Visual Field Defect Archetypes in Patients from a Glaucoma Clinic. Curr. Eye Res 42, 568–574 (2017). [PubMed: 27494512]
23. Elze T et al. Patterns of functional vision loss in glaucoma determined with archetypal analysis. J. R. Soc. Interface 12, (2015).
24. Wang M et al. Reversal of Glaucoma Hemifield Test Results and Visual Field Features in Glaucoma. Ophthalmology 125, 352–360 (2018). [PubMed: 29103791]
25. Wang M et al. An Artificial Intelligence Approach to Detect Visual Field Progression in Glaucoma Based on Spatial Pattern Analysis. Invest. Ophthalmol. Vis. Sci 60, 365–375 (2019). [PubMed: 30682206]
26. Newkirk MR, Gardiner SK, Demirel S & Johnson CA Assessment of False Positives with the Humphrey Field Analyzer II Perimeter with the SITA Algorithm. Investig. Opthalmology Vis. Sci 47, 4632 (2006).
27. Pasquale LR et al. Prospective Study of Type 2 Diabetes Mellitus and Risk of Primary Open-Angle Glaucoma in Women. Ophthalmology 113, 1081–1086 (2006). [PubMed: 16757028]
28. R Core Team. R: A Language and Environment for Statistical Computing. (2014).
29. Jennrich RI Stepwise regression. Stat. Methods Digit. Comput 3, 58–75 (1977).
30. Schwarz G Estimating the Dimension of a Model. Ann. Stat 6, 461–464 (1978).
31. Kass RE & Raftery AE Bayes Factors. J. Am. Stat. Assoc 90, 773 (1995).
32. Omodaka K et al. 3D evaluation of the lamina cribrosa with swept-source optical coherence tomography in normal tension glaucoma. PLoS One 10, e0122347 (2015). [PubMed: 25875096]
33. Gordon MO et al. The Ocular Hypertension Treatment Study: Baseline factors that predict the onset of primary open-angle glaucoma. Arch. Ophthalmol 120, 714–720 (2002). [PubMed: 12049575]
34. Keltner JL et al. Classification of visual field abnormalities in the ocular hypertension treatment study. Arch. Ophthalmol 121, 643–650 (2003). [PubMed: 12742841]
35. Heijl A & Bengtsson B The effect of perimetric experience in patients with glaucoma. Arch. Ophthalmol 114, 19–22 (1996). [PubMed: 8540846]
36. Ishiyama Y, Murata H, Mayama C & Asaoka R An objective evaluation of gaze tracking in humphrey perimetry and the relation with the reproducibility of visual fields: A pilot study in glaucoma. Investig. Ophthalmol. Vis. Sci 55, 8149–8152 (2014). [PubMed: 25389198]

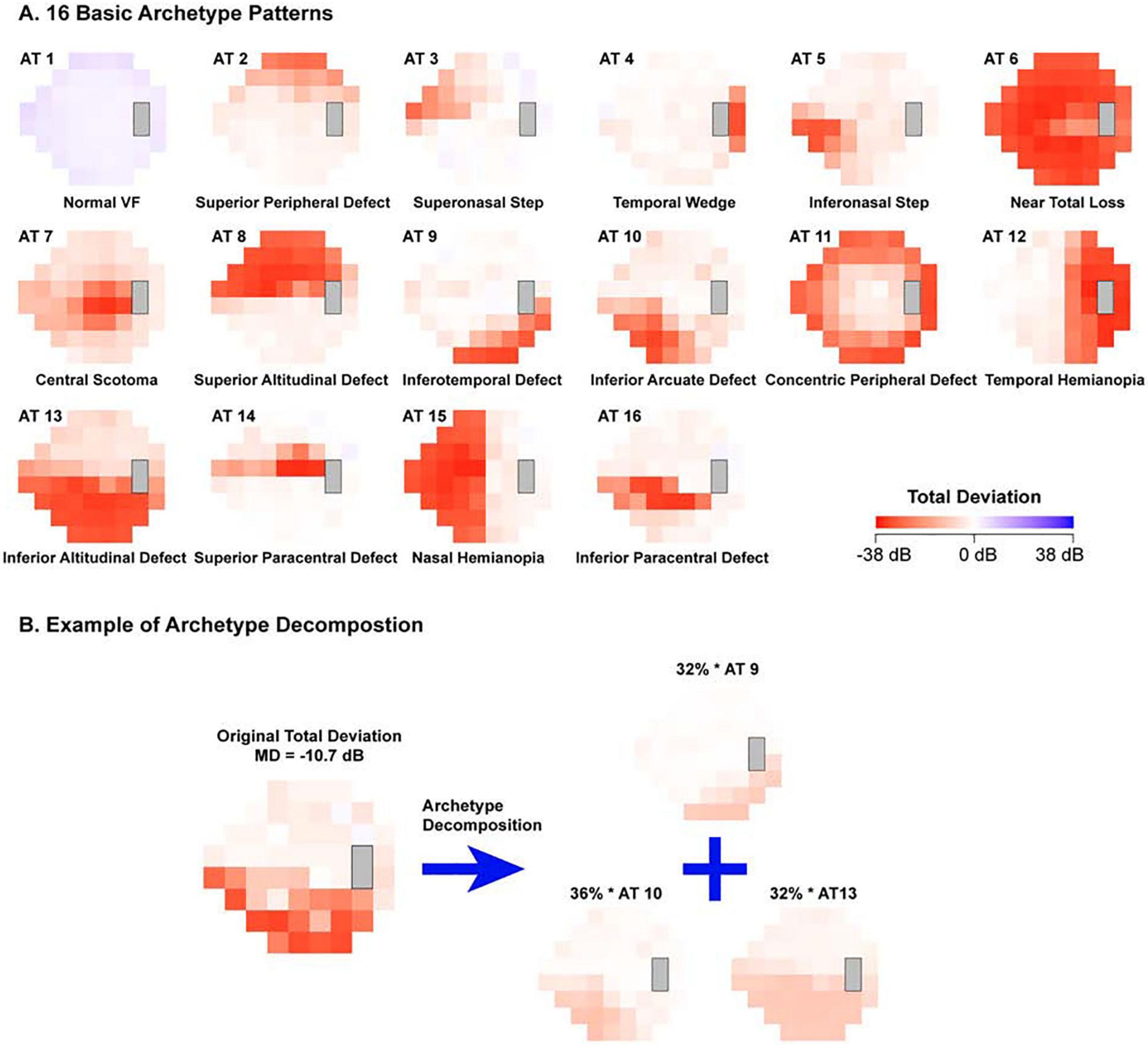

Figure 1.

Illustration of visual field (VF) loss patterns with archetypes (ATs): (**A**) the 16 computationally derived archetypes and (**B**) an example of VF decomposition into its corresponding archetypes.

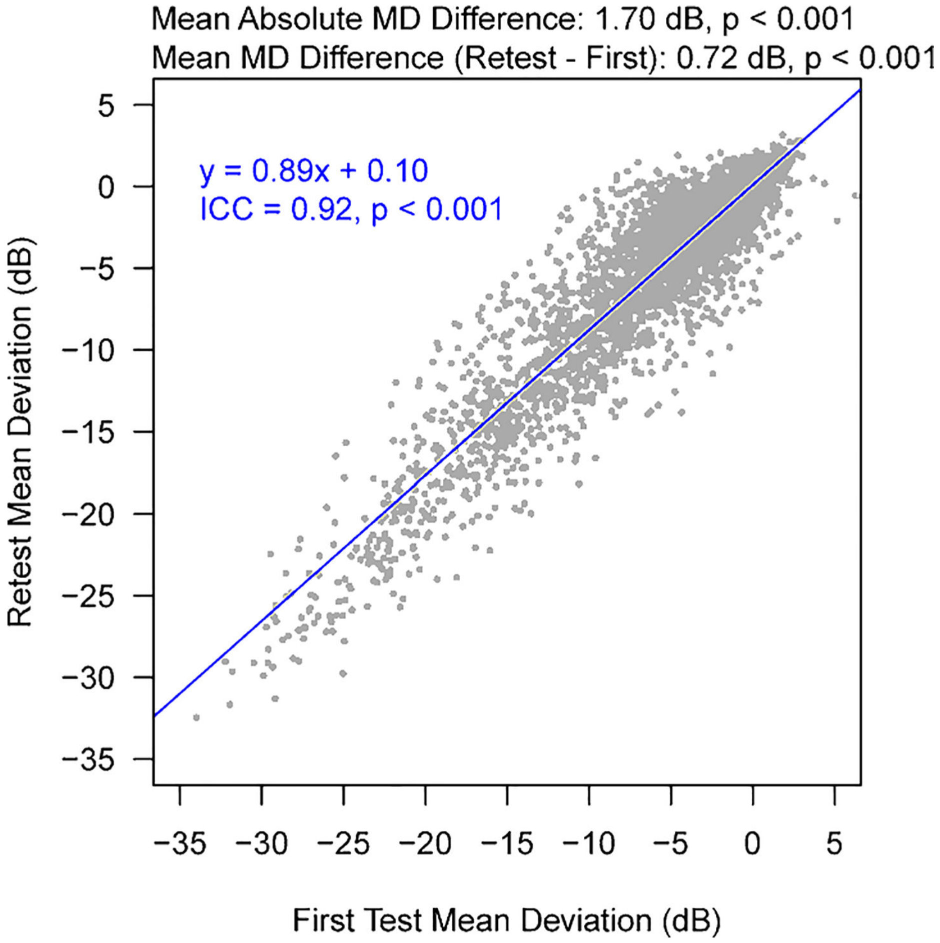

Figure 2.

The distribution of the mean deviations for the first and retest VFs. ICC = intraclass correlation.

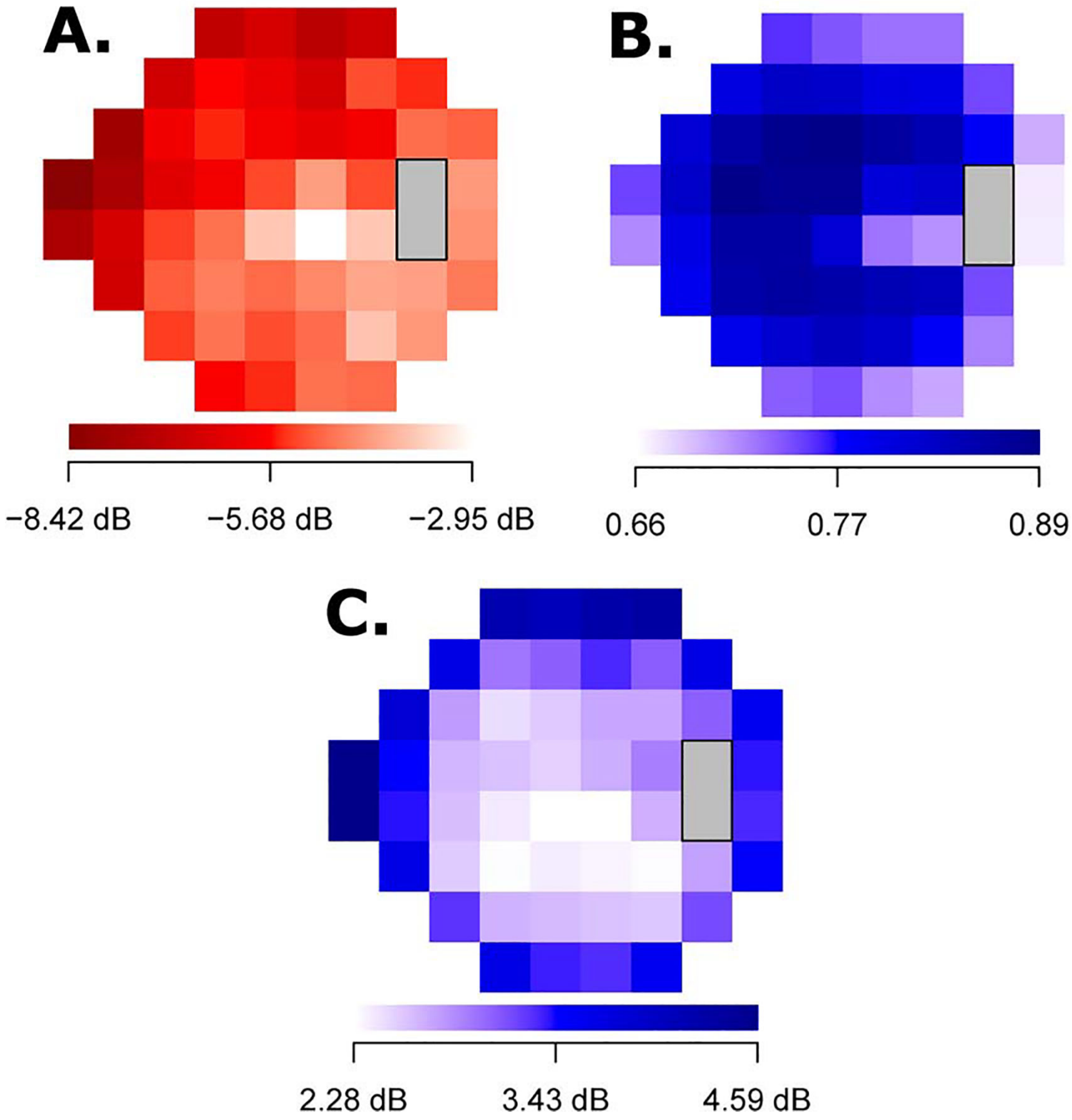

Figure 3.

(**A**)The average TD values of the first VF at 52 test location, (**B**) the intraclass correlations of the TD values at the 52 locations between the first and retest VFs ($p < 0.001$ at all locations), (**C**) the pointwise test-retest variability measured by absolute total deviation differences at 52 test locations ($p < 0.001$ at all locations, differing from zero). TD = total deviation; VF = visual field. P values were corrected for multiple comparisons.

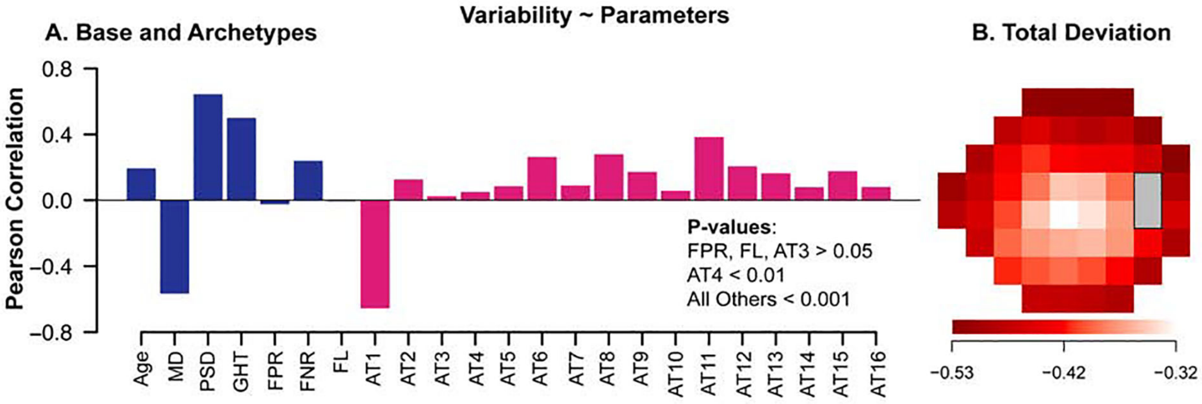

Figure 4.

The Pearson correlations between the global test-retest variability and (**A**) base and archetype parameters, and (**B**) TD values at the 52 locations. TD = total deviation.

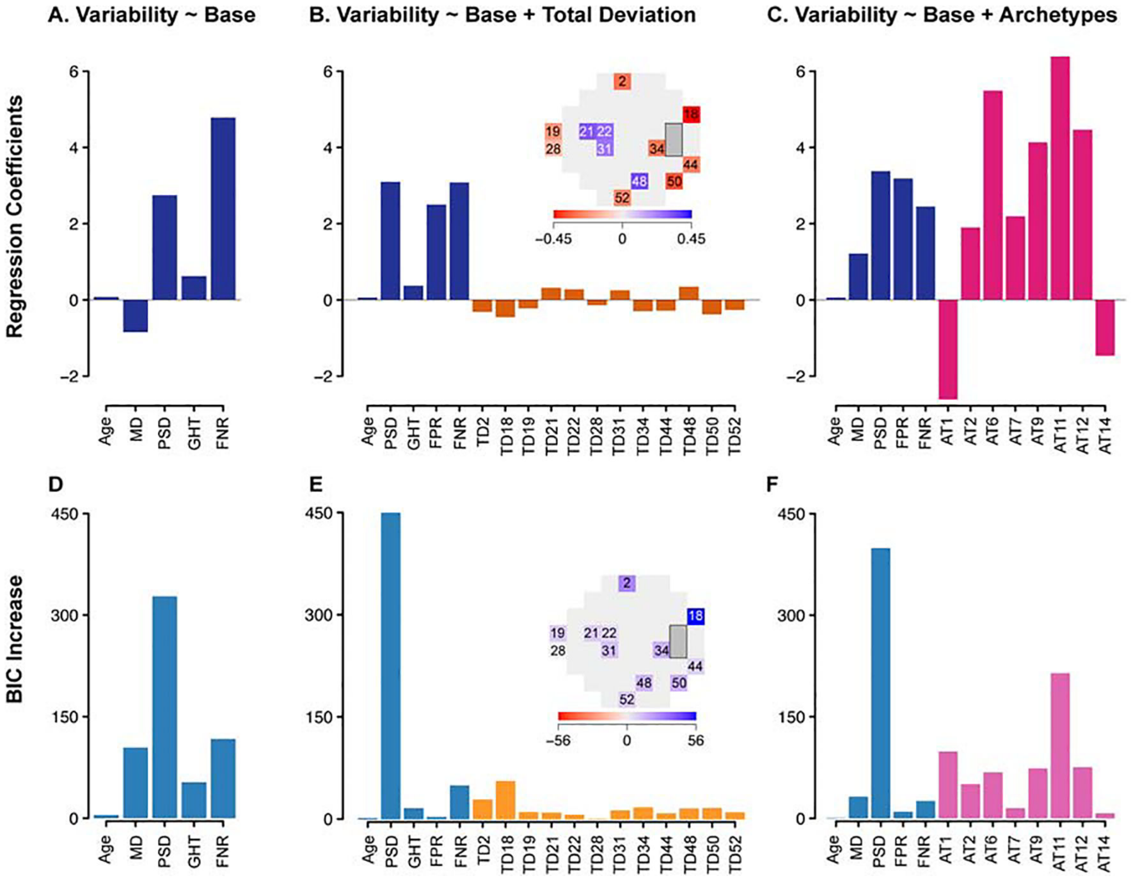

Figure 5.

Best predictive models for test-retest variability using stepwise linear regression. Top panel (**A-C**): regression coefficients for (**A**) "base model" selected from global and reliability indices only, (**B**) "base + total deviation (TD) model" selected from base parameters as well as TD values at 52 locations, and (**C**) "base + archetype (AT) model" selected from base parameters as well as archetypes. Bottom panel (**D-F**): increase in Bayesian information criterion (BIC) when each parameter is removed from the respective models. *Blue* = base parameters; *orange* = TD values; *red* = archetypes. MD = mean deviation; PSD = pattern standard deviation; GHT = glaucoma hemifield test; FNR = false negative rate; FPR = false positive rate.

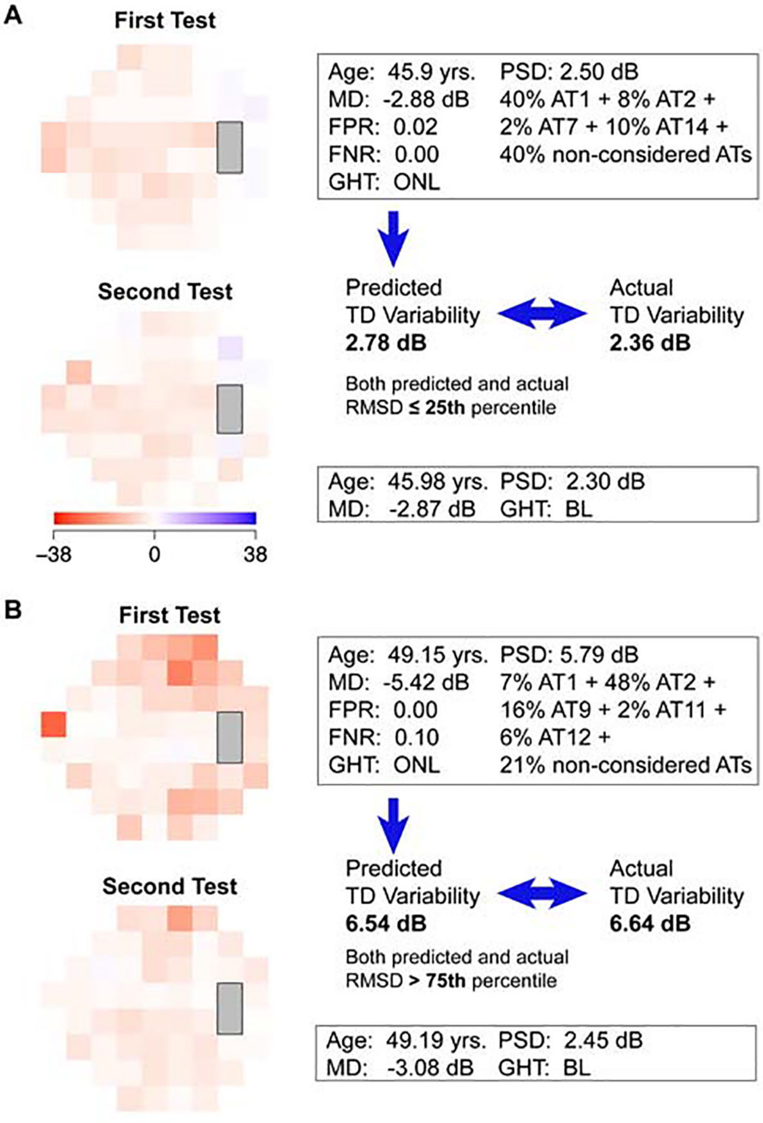

Figure 6.

Example pairs of visual fields (VFs) which display (**A**) relatively low variability (predicted and actual RMSD ≤ 25th percentile) and (**B**) relatively high variability (predicted and actual RMSD > 75th percentile) between the first and retest VFs. VF features, including the archetype (AT) composition of the first VF, are shown in the box. Non-considered ATs are those that are not selected in the best “AT + base model.” The color bar represents total deviation (TD) values in dB. RMSD = root mean square deviation; MD = mean deviation; FPR = false positive rate; FNR = false negative rate; PSD = pattern standard deviation.

Table 1.

Model performance by adjusted R^2 and BIC

Model	Adjusted R^2 (95% CIs)	Absolute BIC	ΔBIC
Base model	0.45 (0.43 – 0.48)	16653.6	BIC_{TD} – BIC_{base} = −527
TD + base model	0.53 (0.50 – 0.56)	16126.6	BIC_{AT} – BIC_{base} = −583
AT + base model	0.53 (0.50 – 0.56)	16070.3	BIC_{AT} – BIC_{TD} = −56

CI = confidence intervals; BIC = Bayesian information criterion; MD = mean deviation; AT = archetype; TD = total deviation.

BIC_{base} = BIC of “base model”; BIC_{AT} = BIC of “AT + base model”; BIC_{TD} = BIC of “TD + base model.”