OCT-based Visual Field Estimation Using Segmentation-free 3D CNN Shows Lower Longitudinal Variability than Standard Automated Perimetry

Makoto Koyama, MD[1]; Hidenori Takahashi, MD, PhD[23]; Satoru Inoda, MD, PhD[3]; Yuta Ueno, MD, PhD[4]; Yoshikazu Ito, MD[4]; Tetsuro Oshika, MD, PhD[4]; Masaki Tanito, MD, PhD[5]*

[1] Minamikoyasu Eye Clinic, 2-8-30 Minamikoyasu, Kimitsu-shi, Chiba, Japan

[2] Center for Cyber Medicine Research, Institute of Medicine, University of Tsukuba, Ibaraki, Japan

[3] Department of Ophthalmology, Jichi Medical University, Shimotsuke, Tochigi, Japan

[4] Department of Ophthalmology, Faculty of Medicine, University of Tsukuba, Ibaraki, Japan

[5] Department of Ophthalmology, Shimane University Faculty of Medicine, Shimane, Japan

***Correspondence:**

Makoto Koyama, MD

Minamikoyasu Eye Clinic, 2-8-30, Minamikoyasu, Kimitsu-shi, Chiba, 299-1162, Japan

E-mail: minamikoyasuganka@gmail.com

Keywords: Visual field; Optical coherence tomography; Artificial intelligence; Glaucoma; Perimetry

Abstract

Purpose: To train and evaluate a segmentation-free 3D convolutional neural network (3DCNN) model for estimating visual field (VF) from optical coherence tomography (OCT) images and to independently assess the longitudinal variability and progression detection capabilities of Humphrey Field Analyzer (HFA) measurements and OCT-based estimated VF (OCT-VF) in a diverse clinical population.

Design: Retrospective multicenter study.

Participants: 13,366 patients (24,313 eyes) underwent HFA tests (24-2 or 10-2 test patterns; 30-2 data were trimmed to 24-2) and macular OCT imaging at five ophthalmic institutions. The dataset included 129,007 paired OCT-VF data points representing various ocular conditions.

Methods: We trained a segmentation-free 3DCNN based on the EfficientNet3D-b0 architecture using a comprehensive OCT dataset without disease-specific exclusions to estimate VF thresholds and mean deviation (MD), employing 10-fold cross-validation. Performance was evaluated using correlation coefficients and Bland-Altman analysis. We assessed longitudinal variability using jackknife resampling and compared residual variability between OCT-VF and HFA measurements using generalized estimating equations, adjusting for follow-up duration, age, and clustering by eye and patient.

Main Outcome Measures: Correlations between OCT-VF and HFA parameters, residual variability, minimum measurements to detect significant progression, and progression rate correlations.

Results: OCT-VF and HFA parameters showed correlations (Pearson's r: 24-2 thresholds 0.863, MD 0.924; 10-2 thresholds 0.881, MD 0.939; all $p < 0.001$). OCT-VF demonstrated significantly lower residual variability than HFA for all parameters (OCT-VF vs. HFA: 0.59 vs. 1.21 dB for 24-2 MD; 0.72 vs. 1.16 dB for 10-2 MD; $p < 0.001$). This advantage persisted across all test points (mean variability reduction: 61.6% for 24-2; 55.6% for 10-2), severity levels, and age groups. Despite shorter follow-up durations, OCT-VF required fewer measurements to detect significant progression and identified more progression events (24-2 MD: 28.4% more). MD slopes showed

correlations between OCT-VF and HFA (Pearson's r: 24-2 MD 0.705, 10-2 MD 0.817; all $p < 0.001$).

Conclusions: A segmentation-free 3DCNN model objectively estimated VF from OCT images with significantly lower longitudinal variability than performance-dependent HFA measurements across diverse ocular conditions. The lower variability of OCT-VF enhances statistical power for progression detection, suggesting its clinical potential for reducing test burden, decreasing measurement noise, and enabling more timely therapeutic interventions.

Introduction

Visual field (VF) testing is crucial for diagnosing and monitoring various ocular conditions, particularly glaucoma, a leading cause of irreversible blindness worldwide.[1–3] Although the Humphrey Field Analyzer (HFA; Carl Zeiss Meditec, Jena, Germany) remains the gold standard for VF assessment, HFA testing has recognized limitations, including its performance-dependent nature, high test-retest variability, and time-consuming process.[2,4,5]

Optical coherence tomography (OCT) has revolutionized ophthalmic imaging, providing high-resolution, objective assessments of ocular structures.[6] The established structure-function relationship between OCT-measured retinal parameters and functional deficits observed in VF testing has created opportunities for more objective VF assessment. Recent advances in artificial intelligence (AI) have enabled the development of approaches that estimate VF directly from OCT images, potentially addressing many limitations of conventional perimetry.[7–15]

Various OCT-based VF estimation models exist, including 2D and 3D approaches. While 2D methods offer computational efficiency, they have limitations: segmentation-based methods reduce generalizability,[7,8,15] and B-scan-based methods have partial imaging coverage.[9–11,15] We adopted a 3DCNN model because 3D approaches leverage complete volumetric data to capture the full depth and spatial complexity of the retinal architecture, allowing for potentially more accurate estimations, as suggested by previous studies.[12–14] This volumetric approach preserves the intricate 3D relationships between retinal layers critical for understanding structure-function relationships, which may substantially reduce variability. Additionally, this approach is segmentation-free, similar to B-scan-based methods, allowing the model to learn directly from 3D OCT data across diverse disease statuses, enhancing generalizability and minimizing selection bias.

Prior research has typically approached VF estimation by treating HFA as a reference, aiming to predict HFA measurements from OCT data.[7–15] However, this approach has limitations due to HFA's inherent response biases and increased test-retest variability with disease severity.[2,4] Unlike previous studies, we evaluated our model by independently analyzing both methods'

internal consistency and variability rather than merely assessing concordance with HFA measurements. This balanced approach acknowledged the variability inherent in both assessment methods and compared their residual variability over time to determine which provided more consistent and clinically useful information for disease monitoring.

The primary aims of this study were to (1) train and evaluate a segmentation-free 3DCNN model for objectively estimating VF from OCT images in a diverse clinical population, (2) compare the residual variability of OCT-based estimated VF (OCT-VF) with that of HFA measurements, and (3) assess the comparative ability of both methods to detect VF progression, potentially offering a more reliable approach for VF assessment and progression monitoring in clinical practice. This research may provide insights into the potential of objective OCT-based VF estimation to complement or enhance conventional performance-dependent perimetry, ultimately improving disease monitoring and patient outcomes in clinical settings. The reduced measurement variability could enable earlier detection of true progression, potentially allowing for more timely therapeutic interventions that preserve visual function.

Methods

Study Design and Participants

The Institutional Review Boards of Shimane University Hospital (IRB No. KS20230719-3, approved on Aug 10, 2023) and Jichi Medical University (IRB No. Rindai24-130, approved on Feb 4, 2025) approved this retrospective multicenter study. Participating institutions included Shimane University Hospital, Shimane, Japan; Jichi Medical University, Tochigi, Japan; JCHO Tokyo Shinjuku Medical Center, Tokyo, Japan; Omiyahamada Eye Clinic, Saitama, Japan; and Minamikoyasu Eye Clinic, Chiba, Japan. This study adhered to the tenets of the Declaration of Helsinki. Due to the retrospective nature of the study, the requirement for informed consent was waived by the respective IRBs. An opt-out approach was employed, and study information was publicly available on each institution's website and premises, allowing patients to decline participation.

Inclusion Criteria

We included eyes that met the following criteria: (1) availability of at least one macular OCT scan with a signal strength index (SSI) ≥ 7 and (2) completion of at least one VF test using the HFA with 30-2, 24-2, or 10-2 test patterns using Swedish Interactive Threshold Algorithm (SITA) strategies. To maintain sample size and reflect real-world clinical practice, we included both SITA-Standard and SITA-Fast strategies in model training and evaluation. Since SITA-Standard is generally considered to have lower test-retest variability,[5] selected key analyses restricted to SITA-Standard only are available in the Supplementary Materials. We trimmed the peripheral points for 30-2 VF tests to match the 24-2 pattern. We applied a ≥33% threshold for all three reliability indices (false-positive, false-negative, and fixation loss rates). This threshold was selected based on prior experimentation showing optimal estimation performance and is consistent with previous studies.[7,9] OCT images were acquired using RS-3000, RS-3000 Advance, RS-3000 Advance2, or Mirante OCT devices (Nidek, Gamagori, Japan) with a 9 mm × 9 mm macular scan protocol.

Data Acquisition and Preprocessing

We paired OCT scans with HFA tests performed within 90 days of the OCT acquisition date, selecting the HFA test with the smallest absolute time difference when multiple tests were available. For eyes with both 24-2 and 10-2 test patterns, we included all available data. For eyes with only one test pattern (e.g., only 24-2 but not 10-2) within the 90-day window, we used mask values to indicate missing data points. We included all eligible paired data from eyes with multiple OCT scans and VF tests in the model development dataset, using the raw thresholds and mean deviation (MD) values without any clipping.

Model Architecture and Configuration

We adopted a segmentation-free 3DCNN model based on EfficientNet3D-b0

architecture[16] with a 30% dropout layer to reduce overfitting (Fig. 1). The model was trained from scratch using a comprehensive OCT dataset including various ocular conditions without disease-based exclusions. Input OCT images were standardized to 224 × 224 × 128 resolution with z-score normalization applied to both images and VF data. The model outputs estimated VF thresholds (52 points for 24-2 and 68 points for 10-2 patterns) and their respective MDs. Following a previous report, we employed horizontal flipping for left eye data and vertical flipping as a data augmentation technique.[8] Model training utilized the Adam optimizer with a mini-batch size of 4 and a variable learning rate schedule, minimizing mean squared error while accounting for missing data points through masked backpropagation.

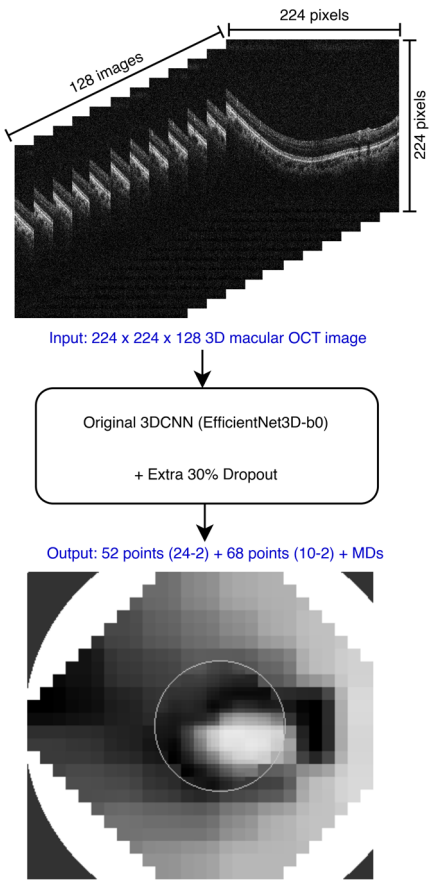

Figure 1. Schematic representation of the segmentation-free 3DCNN model. We based the model on the EfficientNet3D-b0 architecture with an additional 30% dropout layer, which directly connects to the output layer and produces estimated VF thresholds and MD for both 24-2 and 10-2 test patterns.
3DCNN = 3D convolutional neural network; OCT = optical coherence tomography; VF = visual field; MD = mean deviation.

Model Training and Evaluation

We employed a 10-fold cross-validation approach, randomly dividing patients into

training, validation, and test sets at an 8:1:1 ratio. This patient-wise split ensured that data from the same patient did not appear in more than one set, preventing data leakage and allowing unbiased model evaluation. We selected the epoch that showed the best performance on the validation set to evaluate the model's performance on the test set, ensuring the use of the model's optimal weights for the final assessment.

We calculated the mean absolute error (MAE), Pearson's and Spearman's correlation coefficients, and the intraclass correlation coefficient (ICC(2,1)) to assess the relationships between OCT-VF and HFA parameters. We used Bland-Altman plots to evaluate the agreement between OCT-VF and HFA MDs. Additionally, we analyzed the relationships between the MAE and VF severity, refractive errors, and individual test points to assess the model's performance across different clinical scenarios and VF regions.

Independent Longitudinal Analysis for Comparative Reliability Assessment

Following the model performance evaluation, we conducted a separate analysis to compare the longitudinal reliability of OCT-VF and HFA measurements by creating independent datasets for both methods. For the OCT-VF dataset, we generated estimates from all available macular OCT images (SSI $\geq$ 7) using our trained 3DCNN models. This dataset was distinct from the original paired OCT-HFA dataset used for model training and initial evaluation. We assigned each patient to the appropriate model based on their test fold allocation in the cross-validation process, with patients not included in any fold assigned to a randomly selected model. For the HFA dataset, we included all HFA tests, regardless of their false-positive, false-negative, or fixation loss rates, to reflect real-world clinical conditions. We used the thresholds and MD values without any clipping for both groups.

To minimize within-subject variability, we averaged multiple VF measurements taken on the same day at each test point in each group separately. Then, all thresholds in both groups were rounded to integers to maintain consistency with the original HFA measurement scale. We included only eyes with at least six measurements in both groups, excluding any eyes that did not

meet this threshold. Furthermore, we identified which group had fewer measurements for each eye and randomly eliminated excess measurements from the other group until both contained the same number of data points.

Statistical Analysis

We assessed residual variability using jackknife resampling by iteratively excluding data points, fitting regression lines, and calculating residuals. We compared OCT-VF and HFA datasets using generalized estimating equations (GEE), adjusting for follow-up duration, age, clustering by eye and patient, and measurement point location for threshold analyses only. We independently evaluated OCT-VF and HFA datasets for longitudinal analysis to avoid biases from performance-dependent testing.

To address multiple comparisons, we applied Bonferroni correction ($p=0.0125$) for the four primary outcomes and Holm's method for cluster-specific analyses. We visualized relationships between variables using boxplots with regression lines and heat maps. We performed all analyses using Python 3.12.9 with scikit-learn (1.6.1) and statsmodels (0.14.4), and implemented deep learning models with PyTorch (2.6.0).

Results

Model Performance in Estimating Visual Field

Table 1 provides an overview of the population characteristics for the dataset used to train the 3DCNN model, including the number of patients, eyes, and paired OCT-VF data points used during model development. Pearson's correlation coefficients (r) for VF thresholds and MDs were 24-2 thresholds $r=0.863$, MD $r=0.924$; 10-2 thresholds $r=0.881$, MD $r=0.939$ (all $p < 0.001$, Table 2 and Fig. 2). The Bland-Altman analysis showed mean differences between OCT-VF and HFA values less than -0.12 dB for all analyses, with correlation coefficients ranging from -0.28 to -0.13 (Fig. 3).

The model's performance showed slight increases in error for severe cases (Fig. 4). The

model's estimation accuracy remained relatively stable across various OCT focus values, indicating the minimal impact of refractive errors on performance (Fig. 5). The spatial distribution of MAE across the 24-2 (Fig. 6) and 10-2 (Fig. 7) test patterns indicated that points in the nasal VF, where nerve fiber layer pathways are more extensively captured within the OCT scan area, tended to show lower error rates in severe cases. Conversely, points in the temporal VF closer to the optic disc tended to exhibit higher error rates in severe cases.

Table 1. Training Dataset Characteristics

Patient characteristics	Value
Total number of patients	13366
Number of patients in HFA24-2	12995
Number of patients in HFA10-2	4659
Total number of eyes	24313
Number of eyes in HFA24-2	23462
Number of eyes in HFA10-2	8085
Total number of sets of paired data	129007
Number of sets of paired data in HFA24-2	89711
Number of sets of paired data in HFA10-2	39296
Number of HFA24-2 tests per eye (mean ± SD)	9.69 ± 7.61
Number of HFA10-2 tests per eye (mean ± SD)	8.09 ± 5.14
Proportion of SITA-Standard strategy in HFA24-2 (%)	46.5
Proportion of SITA-Standard strategy in HFA10-2 (%)	47.2
Number of OCT tests per eye (mean ± SD)	5.31 ± 7.27
Mean deviation of the HFA24-2 (dB, mean ± SD)	-5.93 ± 7.65
Mean deviation of the HFA10-2 (dB, mean ± SD)	-8.03 ± 8.69
Age (years, mean ± SD)	65.0 ± 14.2
SSI (mean ± SD)	8.38 ± 1.03
Focus (D, mean ± SD)	-2.60 ± 3.56

HFA, Humphrey Field Analyzer; SD, standard deviation; SITA, Swedish Interactive Threshold Algorithm; OCT, optical coherence tomography; SSI, signal strength index of OCT image.

Table 2. OCT-VF and HFA Parameter Correlations

Parameter	Pearson's r	Spearman's ρ	MAE (dB)	ICC (2,1)
24-2 Thresholds	0.863*	0.819*	3.15 ± 2.12	0.853
10-2 Thresholds	0.881*	0.832*	3.45 ± 2.24	0.875
24-2 MD	0.924*	0.832*	2.01 ± 2.13	0.922
10-2 MD	0.939*	0.909*	2.10 ± 2.14	0.938

HFA = Humphrey Field Analyzer; OCT = optical coherence tomography; OCT-VF = OCT-based estimated visual field; MAE = mean absolute error; ICC = intraclass correlation coefficient; MD = mean deviation.
*All correlations significant at $p < 0.001$

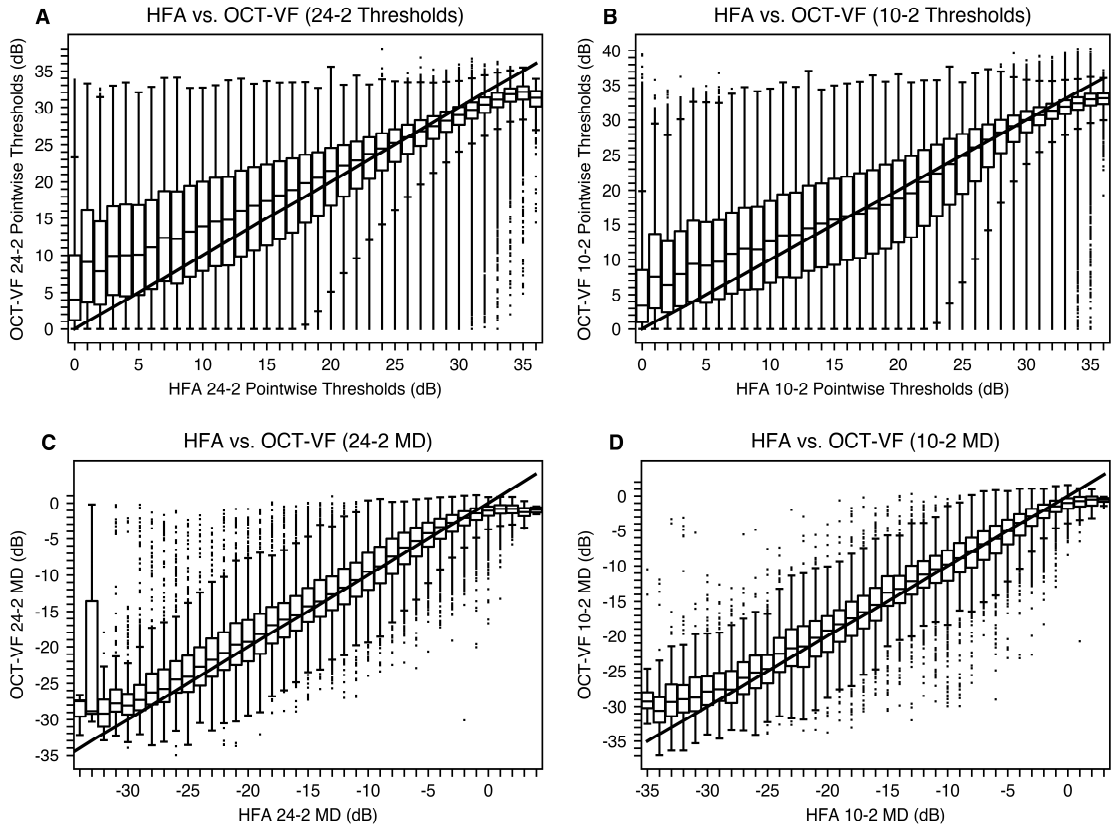

Figure 2. Correlations between HFA and OCT-VF parameters. (A) 24-2 VF threshold (pointwise values, Pearson's r = 0.863), (B) 10-2 VF threshold (pointwise values, Pearson's r = 0.881), (C) 24-2 MD (Pearson's r = 0.924), and (D) 10-2 MD (Pearson's r = 0.939). All correlations are significant at $p < 0.001$. Each panel displays boxplots of estimated values for each measured value and an identity line (x=y). Note the increased error in panel (C) at the MD value of -33 dB. Preliminary investigations from a single-center subset suggested that these outliers were primarily associated with non-organic VF loss, testing errors, or intracranial pathologies such as bitemporal hemianopia.

OCT = optical coherence tomography; VF = visual field; OCT-VF = OCT-based estimated visual field; MD = mean deviation.

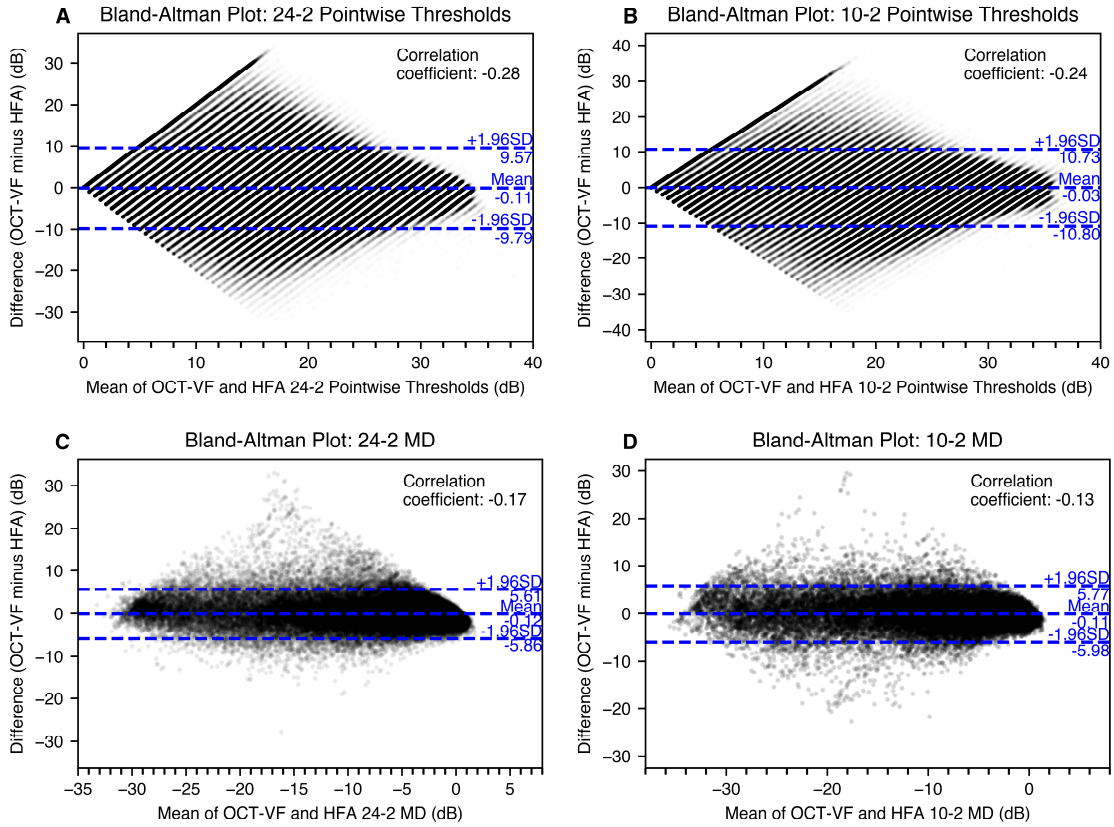

Figure 3. Bland-Altman plots for agreement between OCT-VF and HFA parameters. (A) 24-2 VF threshold (pointwise values), (B) 10-2 VF threshold (pointwise values), (C) 24-2 MD, and (D) 10-2 MD. The horizontal axis represents the mean of the OCT-VF and HFA values, whereas the vertical axis represents the difference between the OCT-VF and HFA values. The central dotted line indicates the mean difference, and the upper and lower dotted lines represent the 95% limits of agreement (±1.96 SD). The striped pattern observed for thresholds (A, B) arises from the discrete integer values of HFA measurements. The plots show the agreement between the OCT-VF and HFA values, with mean differences approaching zero for all analyses. Slight proportional biases were observed, with correlation coefficients ranging from -0.28 to -0.13. The magnitude of the error varied across different levels of VF loss.
VF = visual field; MD = mean deviation; OCT = optical coherence tomography; OCT-VF = OCT-based estimated visual field; HFA = Humphrey Field Analyzer; SD = standard deviation.

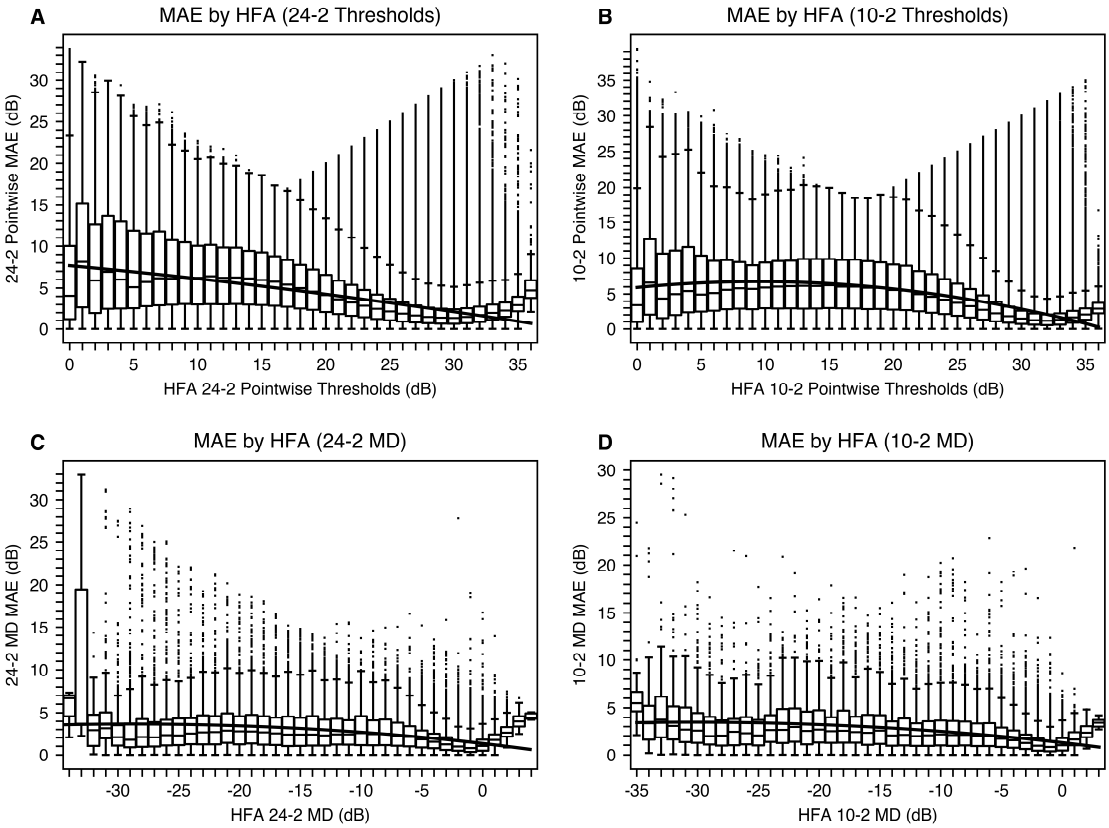

Figure 4. Relationships between VF severity and MAEs. (A) 24-2 VF threshold, (B) 10-2 VF threshold, (C) 24-2 MD, and (D) 10-2 MD. The horizontal axis represents VF severity, and the vertical axis represents MAE (dB). Each panel shows boxplots of MAEs along with a quadratic regression curve. For thresholds (A and B), there are slight increases in MAE in more severe cases. For MD (C and D), MAEs remain relatively stable across severity levels, with a notable exception at -33 dB in panel (C). Prior single-center analysis indicated that cases with extreme MD values and high estimation errors were often related to non-organic VF loss, technical measurement errors, or neurological visual pathway disorders rather than primary ocular pathologies.
VF = visual field; MAE = mean absolute error; MD = mean deviation.

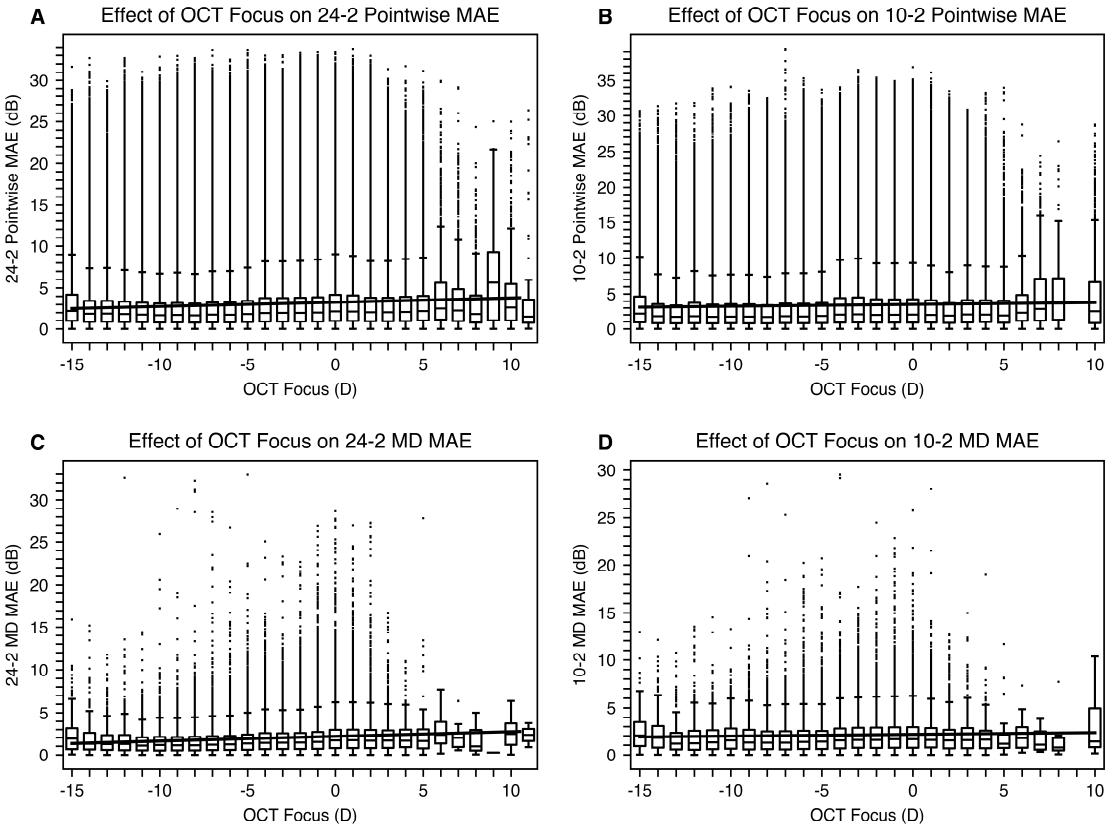

Figure 5. Impact of refractive error on model performance. (A) 24-2 VF thresholds, (B) 10-2 VF thresholds, (C) 24-2 MD, and (D) 10-2 MD. The horizontal axis represents OCT focus values (which closely correspond to refractive error), and the vertical axis represents MAE. Each panel shows boxplots of MAEs for each refractive error, with linear regression lines. Slight increases in MAEs are observed for hyperopic eyes across all measures. Linear regression analysis revealed slopes of 0.052 dB/D (24-2 VF thresholds), 0.027 dB/D (10-2 VF thresholds), 0.056 dB/D (24-2 MD), and 0.019 dB/D (10-2 MD). These minimal slopes quantify the negligible impact of refractive errors on model performance.

OCT = optical coherence tomography; VF = visual field; MD = mean deviation; D = diopter; MAE = mean absolute error.

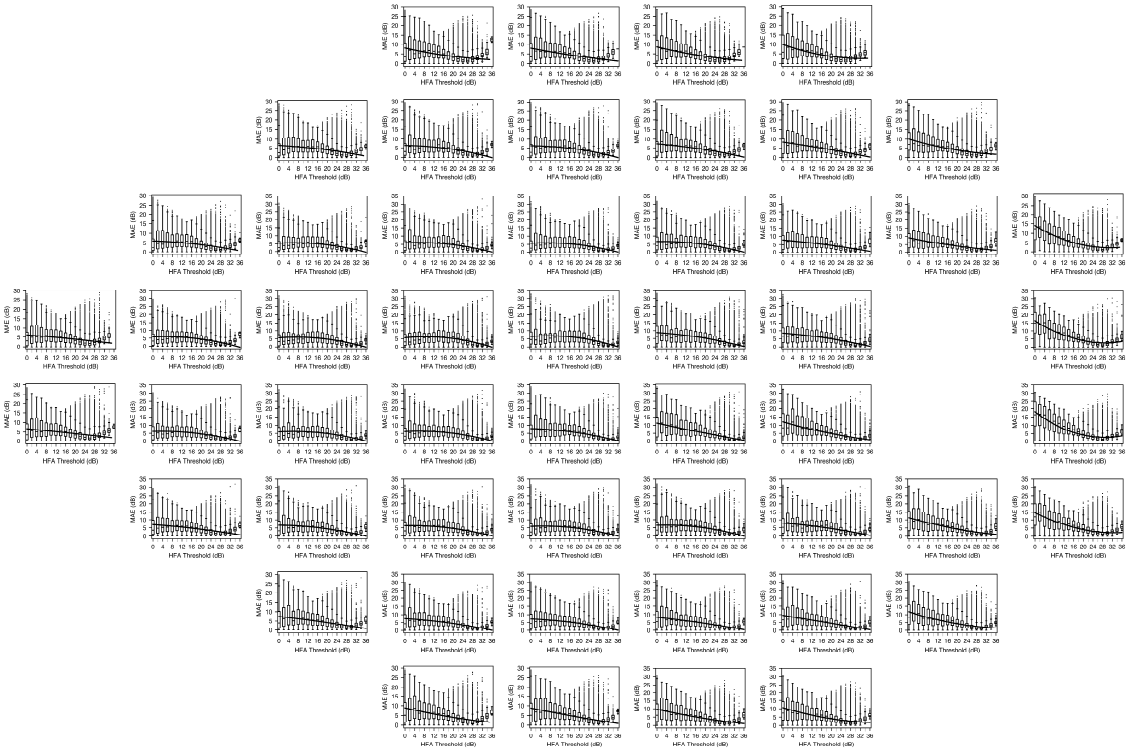

Figure 6. Spatial distribution of estimation accuracy across the 24-2 VF. The boxplots show MAE values for each test point, with quadratic regression curves showing the relationship between severity (horizontal axis) and MAE (vertical axis). Since the OCT imaging covered a 9 mm × 9 mm square centered on the macula (approximately extending to the optic disc), estimation accuracy varies across the VF. In severe cases, points in the nasal VF, where nerve fiber layer information is captured within the scan area, tended to show lower error rates. In contrast, points closer to the optic disc in the temporal VF tended to show higher error rates in severe cases, particularly in regions beyond the area covered by the OCT scan. This pattern suggests a relationship between OCT scan coverage and estimation accuracy across different VF regions. Data from left eyes were horizontally flipped and converted to right eye position for standardized analysis.

MAE = mean absolute error; VF = visual field; OCT = optical coherence tomography.

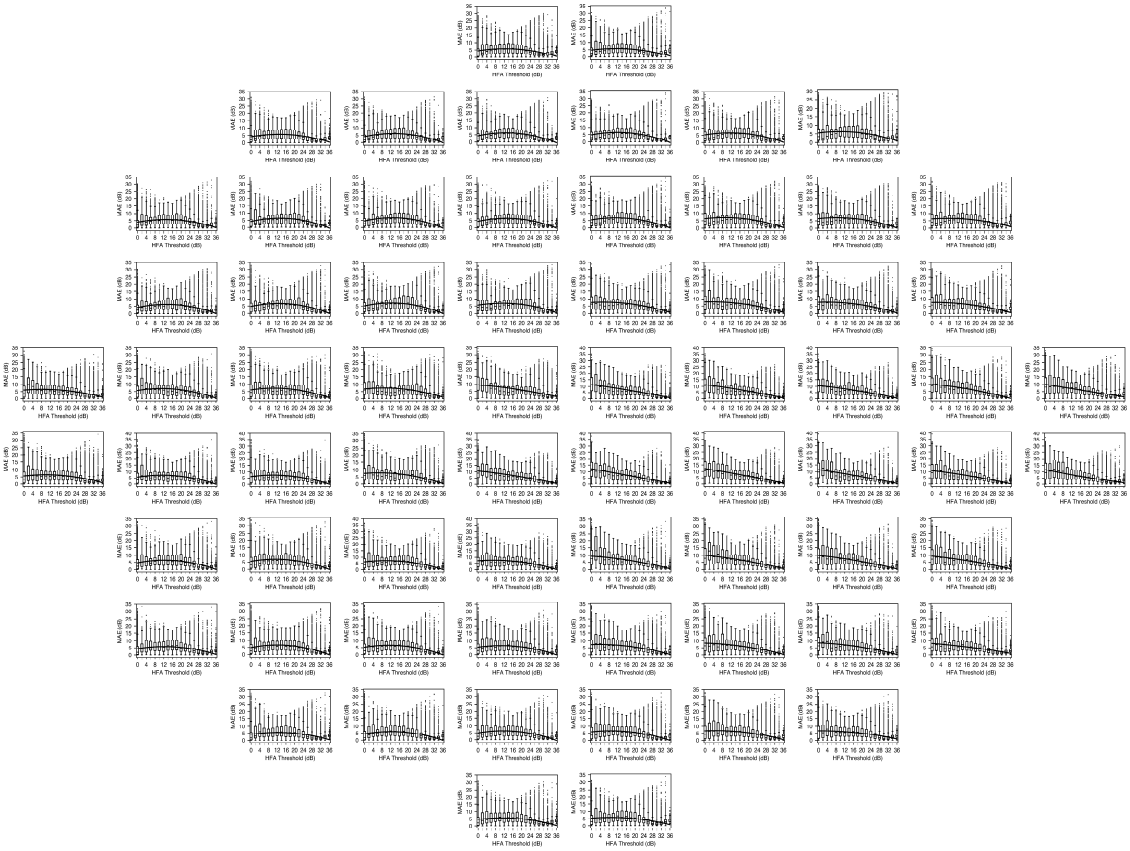

Figure 7. Spatial distribution of estimation accuracy across the 10-2 VF. Similar to Figure 6, these boxplots illustrate how MAE varies with threshold values across the central VF, with quadratic regression curves showing the relationship between severity (horizontal axis) and MAE (vertical axis). In severe cases, points in the nasal VF, where nerve fiber layer pathways are captured within the OCT scan area, tended to show lower error rates. Conversely, points in the temporal VF closer to the optic disc tended to show higher error rates in severe cases. This pattern suggests a relationship between neural pathway coverage in the OCT scan and estimation accuracy. Data from left eyes were horizontally flipped and converted to right eye position to ensure consistent spatial representation.

MAE = mean absolute error; VF = visual field; OCT = optical coherence tomography.

Longitudinal Dataset Characteristics for Comparative Analysis

For the longitudinal analysis comparing OCT-VF and HFA variability, we created independent time-series datasets for each method. Using jackknife resampling, we analyzed the deviations from regression lines for both OCT-VF and HFA measurements separately to determine which method demonstrated less variability and better progression detection capability. Table 3 summarizes the characteristics of these time-series datasets used for comparative analysis. After standardizing the number of measurements between groups, we observed that the mean observation period for the HFA group (24-2: 3179 ± 1410 days, 10-2: 2497 ± 1081 days) was significantly longer than for the OCT-VF group (24-2: 2410 ± 1258 days, 10-2: 2410 ± 1188 days; all $p < 0.001$, Wilcoxon signed-rank test). This difference creates a more stringent comparison condition for the OCT-VF group, as longer observation periods typically provide greater statistical power for detecting progression. A subanalysis including only SITA-Standard tests showed a higher proportion of more severe cases compared to the full dataset (Supplementary Table 4).

Table 3. OCT-VF and HFA Longitudinal Dataset Characteristics

Characteristics	**OCT-VF 24-2**	**HFA 24-2**	**OCT-VF 10-2**	**HFA 10-2**
Number of patients	3519	3519	1557	1557
Number of eyes	6569	6569	2721	2721
Age (years)	66.3 ± 13.4	64.6 ± 13.2	66.2 ± 13.2	65.9 ± 13.1
Mean deviation (dB)	-6.77 ± 7.44*	-5.80 ± 7.15	-8.58 ± 8.48*	-8.32 ± 8.59
Number of tests	10.1 ± 3.69	10.1 ± 3.69	9.48 ± 3.14	9.48 ± 3.14
SITA-Standard (%)	-	45.8	-	50.0
Follow-up duration (days)	2410 ± 1258	3179 ± 1410	2410 ± 1188	2497 ± 1081

OCT = optical coherence tomography; OCT-VF = OCT-based estimated visual field; HFA = Humphrey Field Analyzer; SITA, Swedish Interactive Threshold Algorithm. The values are

presented as the means ± standard deviations.
*Mean deviations for OCT-VF are estimates derived from the OCT-VF models.

Analysis of Visual Field Measurements at the Lower Threshold Limit

Analysis of measurement distributions between OCT-VF and HFA (Fig. 8) revealed differences in handling severely damaged VF regions, particularly at 0 dB. Further longitudinal investigation of measurements following initial 0 dB readings (Fig. 9) showed that HFA measurements had a higher probability of remaining at 0 dB in subsequent tests compared to OCT-VF. When recovering from 0 dB, HFA measurements showed substantial recovery with high variability (24-2: 12.4 ± 8.2 dB, 10-2: 12.2 ± 8.8 dB), while OCT-VF values remained concentrated near the threshold floor (24-2: 1.7 ± 1.4 dB, 10-2: 1.8 ± 1.5 dB). SITA-Standard only subanalyses yielded similar results with consistent trends in the distribution patterns at the 0 dB threshold (Supplementary Figs. 10-11).

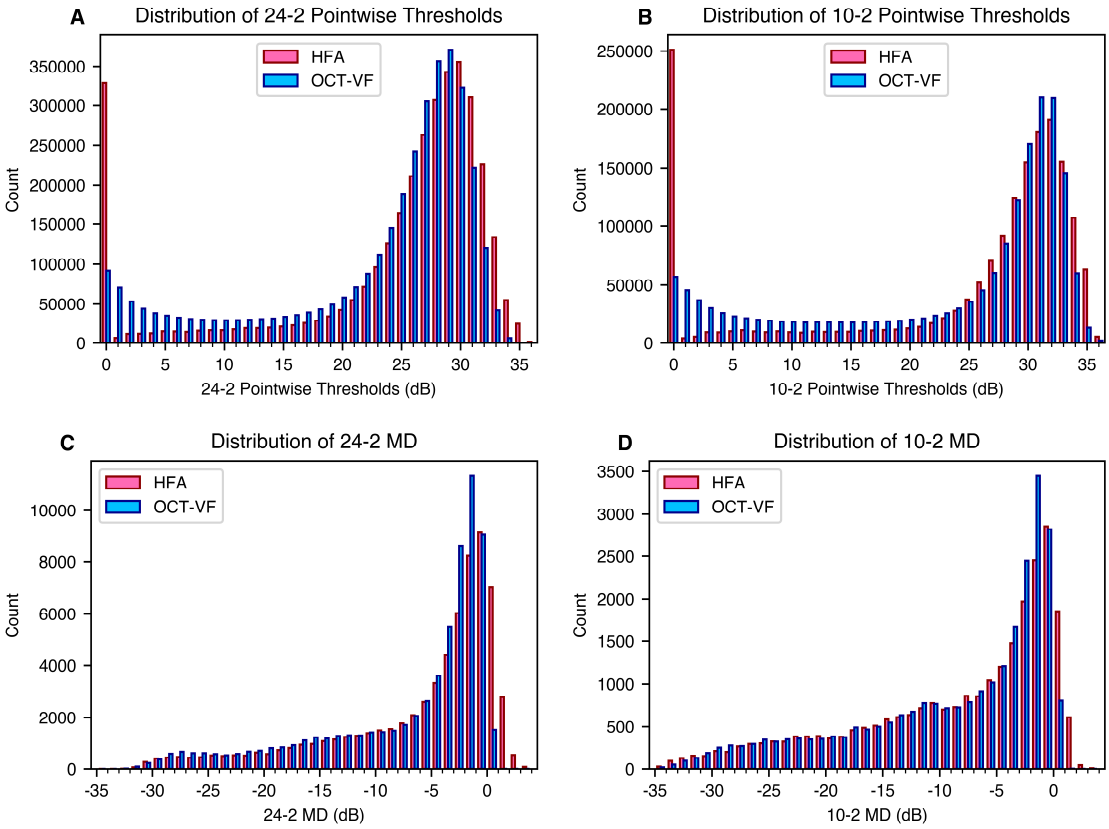

Figure 8. Population distribution by severity level. (A, B) show threshold distributions for 24-2 and 10-2 test patterns, respectively, while (C, D) show MD distributions for 24-2 and 10-2 patterns. In the threshold distributions (A, B), there is a notable discrepancy at 0 dB, with OCT-VF showing fewer measurements at this floor value compared to HFA. This difference prompted the further longitudinal analysis shown in Figure 9 to investigate the behavior of measurements following initial 0 dB readings.
HFA = Humphrey Field Analyzer; OCT = optical coherence tomography; OCT-VF = OCT-based estimated visual field; MD = mean deviation.

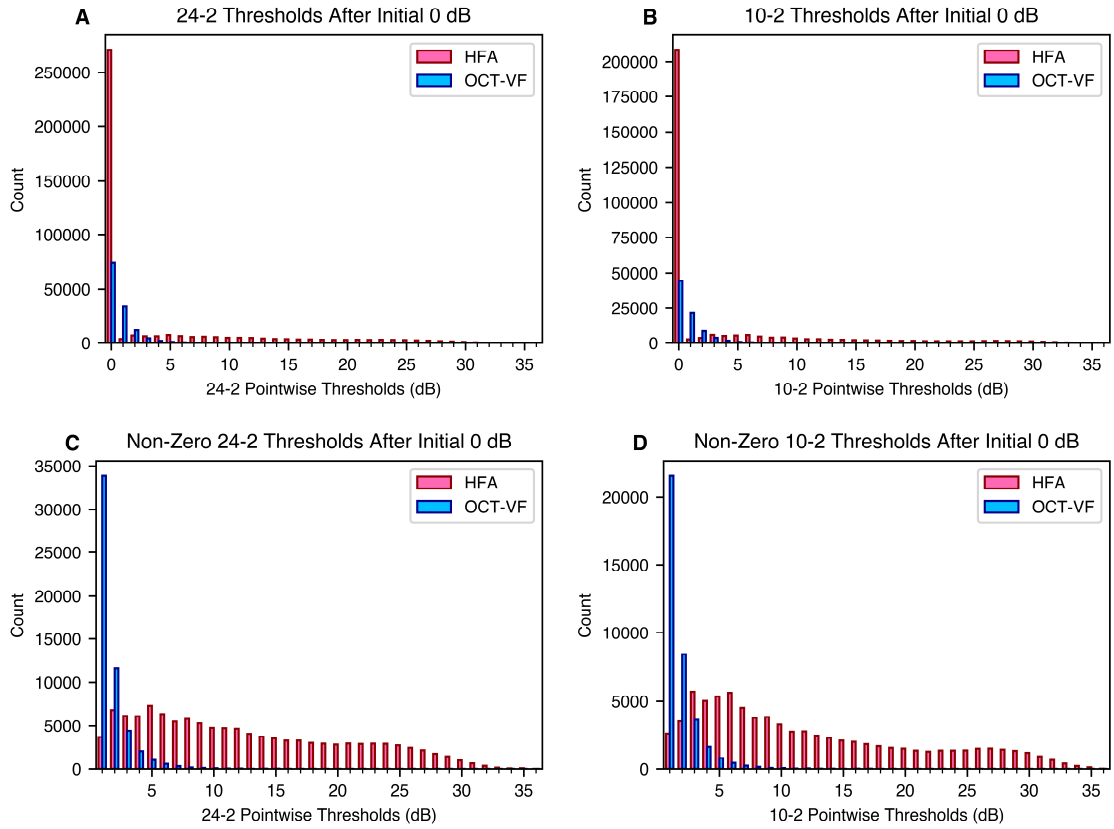

Figure 9. Longitudinal Analysis of Visual Field Thresholds Following Initial 0 dB Measurements. (A, B) show distributions of all pointwise thresholds following initial 0 dB readings in 24-2 and 10-2 tests. HFA measurements remain at 0 dB in 69.7% (24-2) and 72.3% (10-2) of cases, compared to 57.6% (24-2) and 54.5% (10-2) for OCT-VF.
(C, D) show distributions of non-zero thresholds after initial 0 dB readings in 24-2 and 10-2 tests. HFA measurements showed substantial recovery (24-2: 12.4 ± 8.2 dB, 10-2: 12.2 ± 8.8 dB), while OCT-VF concentrate at lower values (24-2: 1.7 ± 1.4 dB, 10-2: 1.8 ± 1.5 dB).
HFA = Humphrey Field Analyzer; OCT = optical coherence tomography; OCT-VF = OCT-based estimated visual field.

Residual Variability Analysis using Generalized Estimating Equations (GEE)

Table 5 presents the residual variability and standard deviations for OCT-VF and HFA datasets. For both 24-2 and 10-2 test patterns, the OCT-VF group showed lower residual variability than the HFA group. For thresholds, OCT-VF vs. HFA residual variability were 0.90 vs. 2.35 dB for 24-2, and 1.07 vs. 2.41 dB for 10-2. For MDs, OCT-VF vs. HFA residual variability were 0.59 vs. 1.21 dB for 24-2, and 0.72 vs. 1.16 dB for 10-2. We compared the residual variability between the OCT-VF and HFA datasets using the GEE models described in the Methods section. OCT-VF exhibited significantly lower residual variability than HFA across

all four analyses ($p < 0.001$), with all results remaining significant after Bonferroni correction ($p < 0.0125$). The SITA-Standard only subanalysis confirmed these findings, with OCT-VF showing lower residual variability than HFA across all parameters (Supplementary Table 6).

Table 5. Residual Variability Comparison: OCT-VF vs. HFA

Variability Metric	OCT-VF Mean ± SD (dB)*	HFA Mean ± SD (dB)*	Difference† (95% CI, dB)	P-value	Working Correlation‡
24-2 Threshold	0.90 ± 0.76	2.35 ± 2.12	-1.42 (-1.45, -1.39)	<0.001	0.44
10-2 Threshold	1.07 ± 0.96	2.41 ± 2.47	-1.33 (-1.37, -1.29)	<0.001	0.38
24-2 MD	0.59 ± 0.54	1.21 ± 1.00	-0.65 (-0.67, -0.62)	<0.001	0.29
10-2 MD	0.72 ± 0.62	1.16 ± 1.03	-0.45 (-0.49, -0.42)	<0.001	0.33

MD = mean deviation; OCT = optical coherence tomography; OCT-VF = OCT-based estimated visual field; HFA = Humphrey Field Analyzer; CI = Confidence Interval.

*The values for OCT-VF and HFA represent the residual variability calculated using jackknife resampling, measuring deviations from the regression line for each method independently; lower values indicate less variability and better measurement consistency.

†The Difference column represents the OCT-VF minus the HFA value, with negative values indicating lower variability in OCT-VF.

‡Working Correlation represents the correlation between measurements from the same eye in the generalized estimating equations model.

Factors Influencing Residual Variability in OCT-VF and HFA Measurements

When examining the relationship between VF severity and residual variability (Fig. 12), cubic regression analysis demonstrated that OCT-VF exhibited lower variability than HFA

measurements across most severity levels for both 24-2 and 10-2 test patterns. This advantage was observed in both threshold measurements and mean deviation values.

The spatial distribution of residual variability, visualized as heat maps, further confirmed this finding (Fig. 13). OCT-VF demonstrated substantially lower variability across all test points compared to HFA, with values ranging from 0.74 to 1.24 dB for OCT-VF versus 1.83 to 3.09 dB for HFA, representing a mean variability reduction of 61.6% for 24-2 and 55.6% for 10-2 test patterns.

Age-related analysis (Fig. 14) revealed that while both methods showed increased variability with advancing age, OCT-VF maintained its advantage across all age groups. The difference in variability between methods widened in older patients, with OCT-VF showing a less pronounced age-related increase in variability.

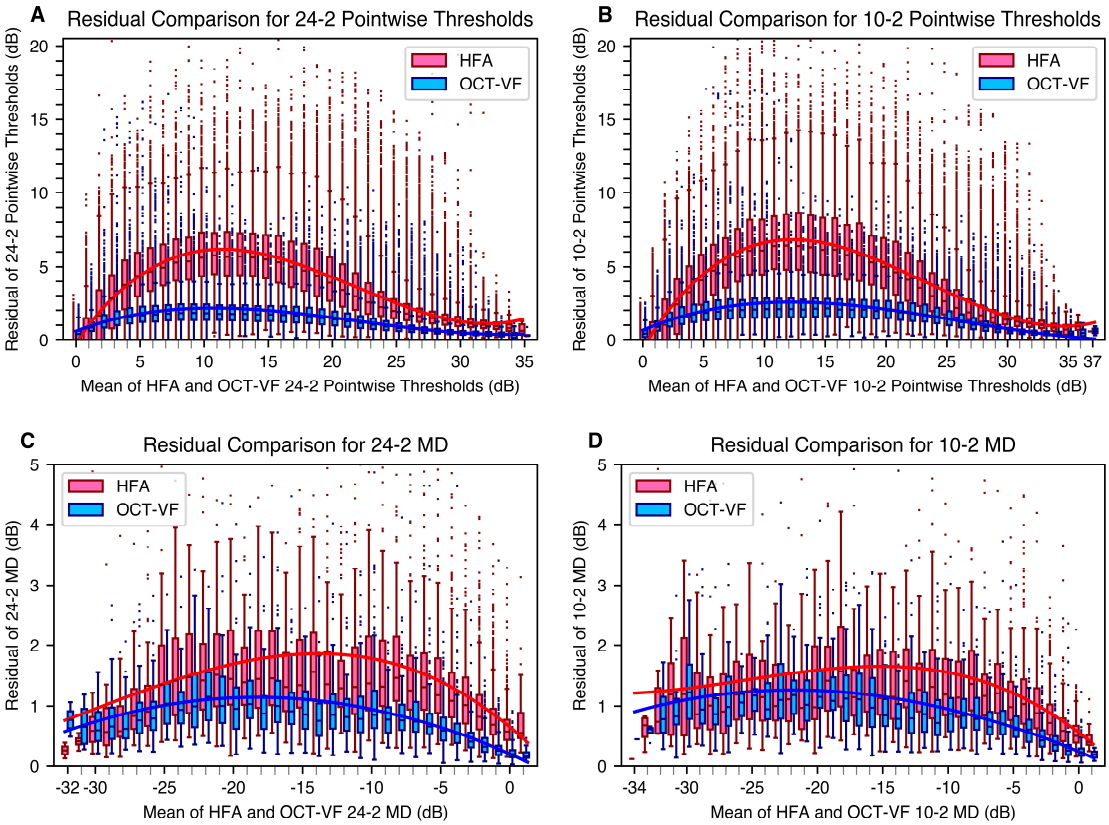

Figure 12. Comparison of residual variability between OCT-VF and HFA measurements across VF severity levels. (A) 24-2 thresholds (pointwise values), (B) 10-2 thresholds (pointwise values), (C) 24-2 MD, and (D) 10-2 MD. The horizontal axis represents VF severity, and the vertical axis

represents residual variability. Each panel shows boxplots of residuals for each severity and cubic regression lines. In all panels, the cubic regression lines for the OCT-VF dataset lie below those of the HFA dataset, indicating lower residual variability across most levels of VF severity for OCT-VF.

HFA = Humphrey Field Analyzer; OCT = optical coherence tomography; VF = visual field; OCT-VF = OCT-based estimated visual field; MD = mean deviation.

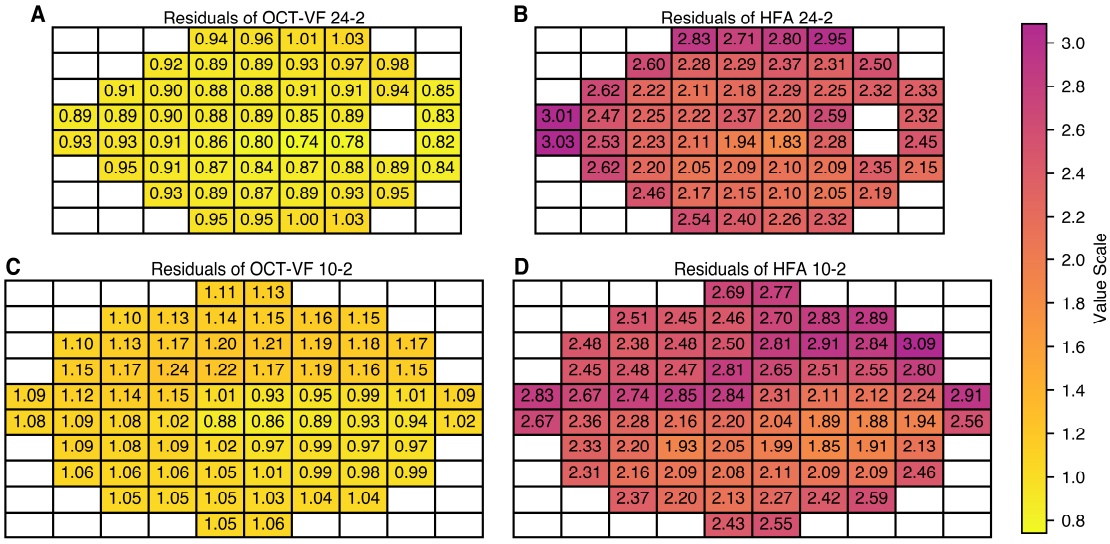

Figure 13. Heat maps of residual variability for each test point. (A) OCT-VF 24-2, (B) HFA 24-2, (C) OCT-VF 10-2, and (D) HFA 10-2. The color scale represents the magnitude of residual variability at each test point. OCT-VF shows consistently lower residual variability than HFA measurements across all test points, with values ranging from 0.74 to 1.24 dB for OCT-VF compared to 1.83 to 3.09 dB for HFA measurements. We horizontally flipped the left eye data and integrated them with the right eye data.

OCT = optical coherence tomography; OCT-VF = OCT-based estimated visual field; HFA = Humphrey Field Analyzer.

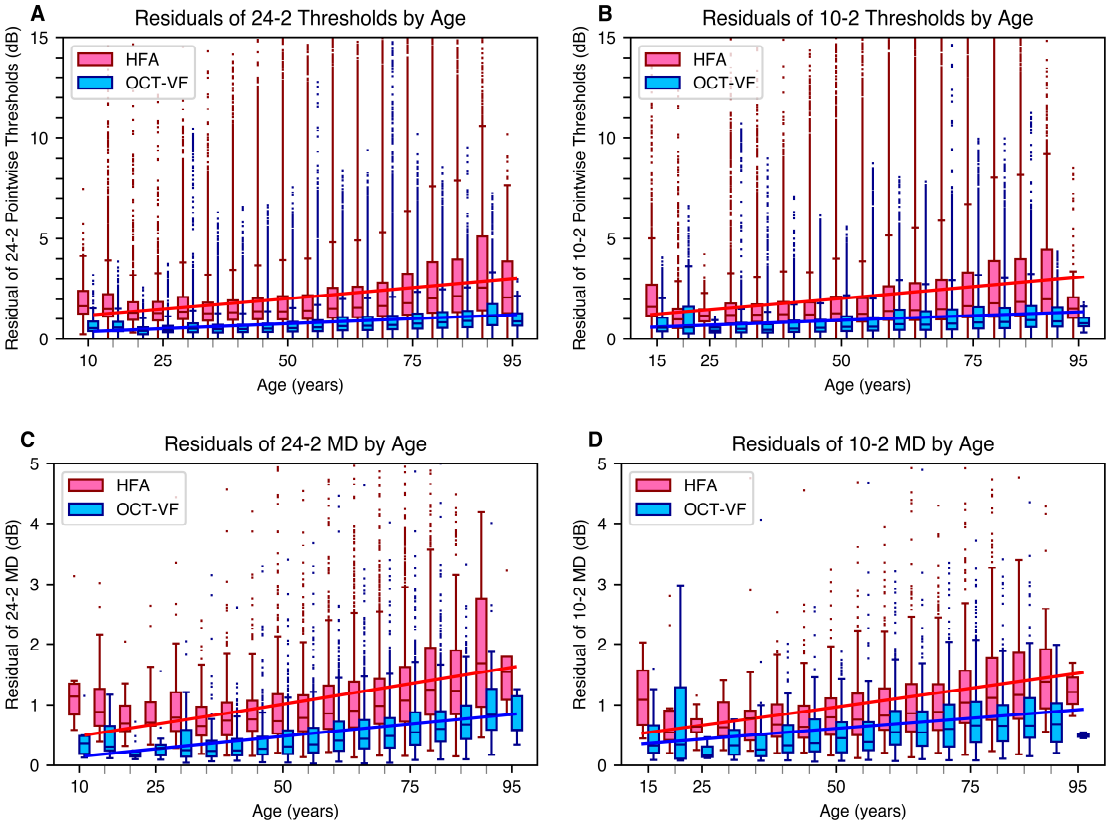

Figure 14. Comparison of residual variability between HFA measurements and OCT-VF across age groups. (A) 24-2 thresholds, (B) 10-2 thresholds, (C) 24-2 MD, and (D) 10-2 MD. The horizontal axis represents age in 5-year increments, and the vertical axis represents residual variability. Each panel shows boxplots of residuals for each age group, along with linear regression lines for HFA (red) and OCT-VF (blue). The regression lines for OCT-VF consistently lie below those for HFA, indicating lower residual variability across all age groups. Both methods show increased variability with age, but HFA demonstrates steeper slopes, particularly for threshold measurements. The slopes of the regression lines (in dB/year) for HFA and OCT-VF are (A) 0.0214 vs. 0.0096, (B) 0.0230 vs. 0.0093, (C) 0.0135 vs. 0.0083, and (D) 0.0123 vs. 0.0071, respectively. These values suggest that, compared with HFA, OCT-VF is less affected by age-related increases in variability.
HFA = Humphrey Field Analyzer; OCT = optical coherence tomography; OCT-VF = OCT-based estimated visual field; MD = mean deviation.

Progression Rate Analysis

To assess progression detection capabilities, we divided VF into anatomically relevant clusters (Fig. 15). For the 24-2 pattern, we used 10 clusters based on the Glaucoma Hemifield Test,[17] while for the 10-2 pattern, we employed 8 clusters according to Suzumura et al.'s classification.[18] We analyzed longitudinal correlations using Spearman's rank correlation

coefficient (Table 7), finding OCT-VF demonstrated stronger correlations and detected more significant progression events than HFA across all parameters despite having shorter follow-up durations: 47.2% more for 24-2 clusters, 63.4% more for 10-2 clusters, 28.4% more for 24-2 MD, and 32.6% more for 10-2 MD. For cases showing significant progression in both methods (Table 8), OCT-VF consistently required fewer measurements than HFA across all parameters. Bland-Altman analysis (Fig. 16) showed agreement in progression rates between methods, with mean differences (all ≤0.20 dB/year) and correlation coefficients (all ≤0.22). Boxplot analysis (Fig. 17) showed correlations between OCT-VF and HFA progression rates (Pearson's r: 24-2 clusters 0.702, 10-2 clusters 0.772, 24-2 MD 0.705, 10-2 MD 0.817; all $p < 0.001$).

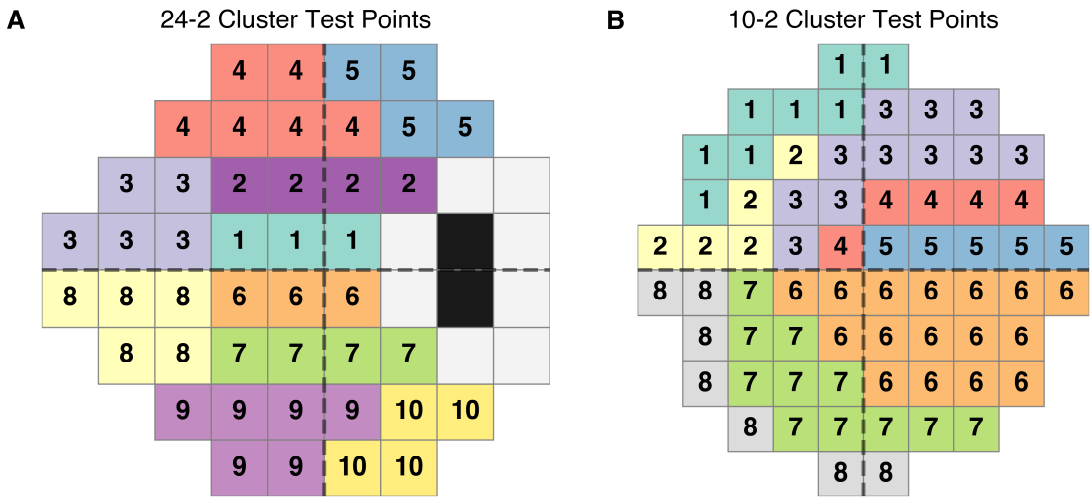

Figure 15. Visual field test point clusters used for progression analysis. (A) 24-2 test pattern with 10 clusters based on the Glaucoma Hemifield Test regions. (B) 10-2 test pattern with 8 clusters based on Suzumura et al.'s classification. Different colors represent different cluster regions used for statistical analysis of progression.

Table 7. Longitudinal Correlation Metrics: OCT-VF vs. HFA

Metric	Test Type	Spearman's ρ (Mean ± SD)	Significant Events (n)*
24-2 Clusters	OCT-VF	-0.353 ± 0.441	8390 / 65690
24-2 Clusters	HFA	-0.222 ± 0.471	5698 / 65690
10-2 Clusters	OCT-VF	-0.373 ± 0.451	3051 / 21768
10-2 Clusters	HFA	-0.264 ± 0.466	1867 / 21768
24-2 MD	OCT-VF	-0.438 ± 0.430	1879 / 6569
24-2 MD	HFA	-0.198 ± 0.535	1463 / 6569
10-2 MD	OCT-VF	-0.487 ± 0.428	863 / 2721
10-2 MD	HFA	-0.306 ± 0.517	651 / 2721

HFA = Humphrey Field Analyzer; OCT = optical coherence tomography; MD = mean deviation; OCT-VF = OCT-based estimated visual field; SD = standard deviation.

*Significant Events (n) are the "number of significant events / total number of observations." The numerator represents cases with statistically significant progression after Bonferroni correction ($p < 0.0125$) for the four main parameters and additional Holm correction for clusters, using Spearman's rank correlation coefficient. The denominator represents the total number of cluster regions or eyes analyzed.

Table 8. Required Test Count for Progression Detection

Metric	OCT-VF (Mean ± SD)	HFA (Mean ± SD)	Progression Events (n)*
24-2 Clusters	11.1 ± 3.4	12.5 ± 3.9	2134
10-2 Clusters	9.8 ± 2.9	10.6 ± 3.0	849
24-2 MD	9.6 ± 3.1	10.2 ± 3.4	682
10-2 MD	8.5 ± 2.2	9.0 ± 2.6	366

HFA = Humphrey Field Analyzer; OCT-VF = optical coherence tomography-based estimated visual field; MD = mean deviation; SD = standard deviation; CI = confidence interval.
*Progression events (n) represent the number of cases with statistically significant progression detected in both HFA and OCT-VF methods after Bonferroni correction ($p < 0.0125$) for the four main parameters and additional Holm correction for clusters, using Spearman's rank correlation coefficient.

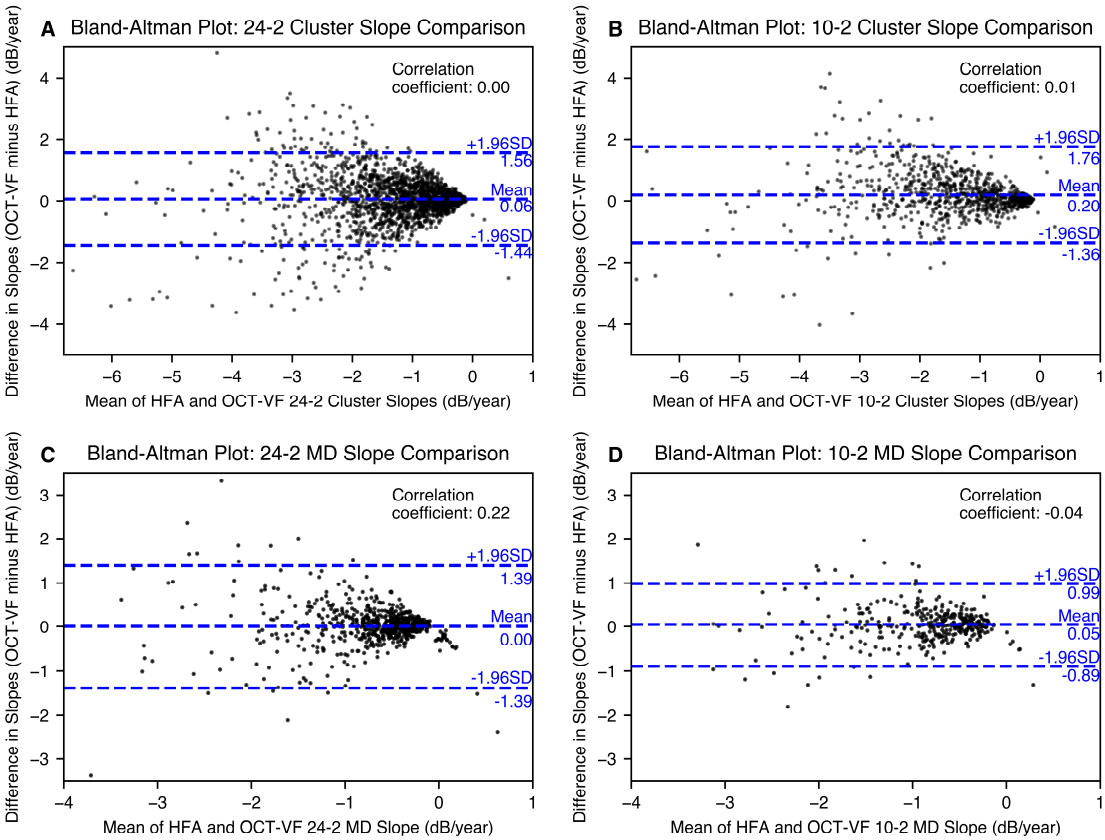

Figure 16. Bland-Altman plots comparing progression rates between OCT-VF and HFA. (A) 24-2 clusters, (B) 10-2 clusters, (C) 24-2 MD, and (D) 10-2 MD. The analysis includes only clusters or MDs that showed significant progression (after Bonferroni and Holm corrections) in both

methods. The total number of progression events was 2134 in (A), 849 in (B), 682 in (C), and 366 in (D). Mean differences across all analyses were 0.20 dB/year or less, with correlation coefficients 0.22 or less.
HFA = Humphrey Field Analyzer; OCT = optical coherence tomography; OCT-VF = OCT-based estimated visual field; MD = mean deviation.

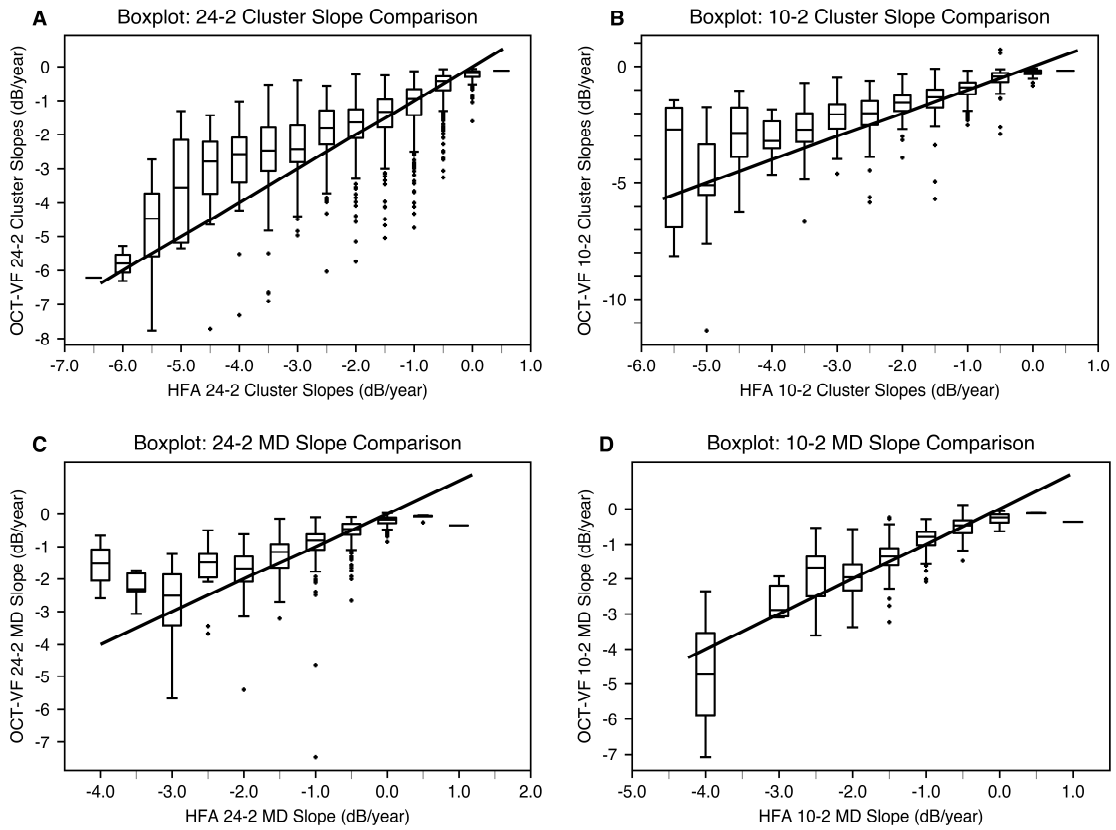

Figure 17. Boxplots comparing progression rates between HFA and OCT-VF. (A) 24-2 clusters, (B) 10-2 clusters, (C) 24-2 MD, and (D) 10-2 MD. The analysis includes only clusters or MDs that showed significant progression in both methods (2134, 849, 682, and 366 events, respectively). Each boxplot represents OCT-VF progression rates for the corresponding HFA progression rate. The solid line represents the identity line (x=y). Pearson's correlation coefficients were 0.702 (A), 0.772 (B), 0.705 (C), and 0.817 (D), all $p < 0.001$.
HFA = Humphrey Field Analyzer; OCT = optical coherence tomography; OCT-VF = OCT-based estimated visual field; MD = mean deviation.

Discussion

This study demonstrated that our segmentation-free 3DCNN model objectively estimated VF with significantly lower residual variability than HFA measurements. The neural

network likely converges toward the underlying "true" VF by averaging out random variability in individual HFA measurements—a key advantage of data-driven approaches in extracting stable patterns from noisy training data.

Previous VF research typically used HFA measurements as the reference standard due to challenges in determining the true underlying VF. Our analysis at the lower threshold limit (0 dB) revealed critical differences between methods. When HFA measurements changed from 0 dB, they showed substantial fluctuations (mean: ~12 dB). This aligns with Gardiner et al.'s findings on reduced reliability of perimetric sensitivities below 19 dB.[19] In contrast, OCT-VF values showed higher stability, with changes confined to a narrow range (mean: ~2 dB). One possible explanation for the higher prevalence of 0 dB measurements in HFA may be threshold clipping effects. When actual VF values are near 0 dB but subject to variability, HFA measurements can only range from 0 dB upward, potentially creating an asymmetric distribution with increased frequency at 0 dB. Conversely, the OCT-VF model may estimate values above 0 dB when uncertainty exists.

We employed a balanced methodological approach in two stages. First, we used HFA as a reference to establish baseline performance metrics for our OCT-VF model. Then, we independently analyzed longitudinal data for each method for our primary analysis of variability and progression detection. This approach allowed us to evaluate the intrinsic variability of both methods without assuming either as a gold standard.

OCT-based VF estimations demonstrated significantly lower residual variability than HFA measurements across all parameters. GEE analysis revealed OCT-VF's superior consistency for 24-2 thresholds, 10-2 thresholds, 24-2 MD, and 10-2 MD, even after adjusting for confounding factors. This enhanced reliability persisted across severity levels, regions, and age groups, with pronounced advantages in older populations. The progression analysis showed that OCT-VF required fewer measurements to detect significant progression and identified more significant progression events despite the HFA group having longer observational periods.

Our findings have significant clinical implications. OCT-VF's lower variability

addresses a key clinical challenge: distinguishing true disease progression from test fluctuation. This reduced measurement noise may facilitate the detection of functional changes with increased statistical power, potentially advancing therapeutic intervention timelines. Requiring fewer measurements offers practical benefits, including reduced patient burden, lower healthcare costs, and more efficient resource allocation. This objective approach may benefit patients with conditions that affect traditional VF testing reliability, such as limited test proficiency, advanced age, or cognitive impairments.[20]

While our OCT-VF approach offers advantages, this study has several limitations. The model's architecture and training process may introduce inherent biases, particularly in cases where structural OCT findings and functional visual field results are discordant. Additionally, our study was limited by using OCT devices from a single manufacturer with a specific imaging protocol. Future research should validate these findings through external studies with diverse populations, assess model performance across different disease subtypes, and evaluate compatibility with various OCT platforms.

Conclusions

Our findings demonstrate that a segmentation-free, OCT-based 3DCNN model offers less variable visual field estimations than standard automated perimetry. This approach addresses a key limitation in visual field testing by providing more consistent measurements, particularly valuable in monitoring disease progression and in vulnerable populations. By recognizing the complementary nature of structural and functional assessments, OCT-VF enhances the detection of true functional changes with fewer measurements, potentially enabling earlier interventions and better preservation of vision.

References

1. Tham YC, Li X, Wong TY, Quigley HA, Aung T, Cheng CY. Global prevalence of glaucoma and projections of glaucoma burden through 2040: A systematic review and meta-analysis. *Ophthalmology*. 2014;121(11):2081-2090.

2. Delgado MF, Nguyen NTA, Cox TA, et al. Automated perimetry: a report by the American Academy of Ophthalmology. *Ophthalmology*. 2002;109(12):2362-2374.

3. Weinreb RN, Aung T, Medeiros FA. The pathophysiology and treatment of glaucoma: A review. *JAMA*. 2014;311(18):1901-1911.

4. Heijl A, Lindgren A, Lindgren G. Test-retest variability in glaucomatous visual fields. *Am J Ophthalmol*. 1989;108(2):130-135.

5. Artes PH, Iwase A, Ohno Y, Kitazawa Y, Chauhan BC. Properties of perimetric threshold estimates from Full Threshold, SITA Standard, and SITA Fast strategies. *Invest Ophthalmol Vis Sci*. 2002;43(8):2654-2659.

6. Huang D, Swanson EA, Lin CP, et al. Optical coherence tomography. *Science*. 1991;254(5035):1178-1181.

7. Asaoka R, Xu L, Murata H, et al. A joint multitask learning model for cross-sectional and longitudinal predictions of visual field using OCT. *Ophthalmol Sci*. 2021;1(4):100055.

8. Hashimoto Y, Kiwaki T, Sugiura H, et al. Predicting 10-2 visual field from optical coherence tomography in glaucoma using deep learning corrected with 24-2/30-2 visual field. *Transl Vis Sci Technol*. 2021;10(13):28.

9. Kamalipour A, Moghimi S, Khosravi P, et al. Deep learning estimation of 10-2 visual field map based on circumpapillary retinal nerve fiber layer thickness measurements. *Am J Ophthalmol*. 2023;246:163-173.

10. Kihara Y, Montesano G, Chen A, et al. Policy-driven, multimodal deep learning for predicting visual fields from the optic disc and OCT imaging. *Ophthalmology*. 2022;129(7):781-791.

11. Lazaridis G, Montesano G, Afgeh SS, et al. Predicting visual fields from optical coherence tomography via an ensemble of deep representation learners. *Am J Ophthalmol*. 2022;238:52-65.

12. Mohammadzadeh V, Vepa A, Li C, et al. Prediction of central visual field measures from macular OCT volume scans with deep learning. *Transl Vis Sci Technol*. 2023;12(11):5.

13. Chen Z, Shemuelian E, Wollstein G, Wang Y, Ishikawa H, Schuman JS. Segmentation-free OCT-volume-based deep learning model improves pointwise visual field sensitivity estimation. *Transl Vis Sci Technol*. 2023;12(6):28.

14. Yu HH, Maetschke SR, Antony BJ, et al. Estimating global visual field indices in glaucoma by combining macula and optic disc OCT scans using 3-dimensional convolutional neural networks. *Ophthalmol Glaucoma*. 2021;4(1):102-112.

15. Mohammadzadeh V, Wu S, Besharati S, et al. Prediction of visual field progression with baseline and longitudinal structural measurements using deep learning. *Am J Ophthalmol*. 2024;262:141-152.

16. Shi J. EfficientNet-PyTorch-3D. github.com. Updated Mar 26, 2023. Accessed Feb 1, 2025. https://github.com/shijianjian/efficientnet-pytorch-3d.

17. Asman P, Heijl A. Glaucoma Hemifield Test. Automated visual field evaluation. *Arch Ophthalmol*. 1992;110(6):812-819.

18. Suzumura H, Yoshikawa K, Kimura T, Nanno M, Tsumura T. Cluster formation for analyses of glaucomatous visual field defects in central 10-2 visual field in normal tension glaucoma eyes. *Clin Ophthalmol*. 2022;16:611-618.

19. Gardiner SK, Swanson WH, Goren D, Mansberger SL, Demirel S. Assessment of the reliability of standard automated perimetry in regions of glaucomatous damage. *Ophthalmology*. 2014;121(7):1359-1369.

20. Ichitani A, Takao E, Tanito M. Roles of cognitive function on visual field reliability indices among glaucoma patients. *J Clin Med*. 2023;12(22):7119.

Data availability statement:

We are unable to make the datasets publicly available due to privacy and ethical considerations related to patient data. We conducted the study on an opt-out basis, without obtaining explicit consent from all participants to release their raw data.

Financial Disclosures / Conflicts of Interest:

Makoto Koyama: Royalties or licenses: DeepEyeVision Inc. (potential future royalties if the product is commercialized); Stock or stock options: DeepEyeVision Inc. (purchased stock in the company); Other financial or non-financial interests: DeepEyeVision Inc. (ongoing discussions about potential future product development and commercialization).

Hidenori Takahashi: Company: DeepEyeVision Inc.; Relationship: Equity interest (major shareholder); Leadership role: (Founder and Board Member); Stock or stock options: DeepEyeVision Inc. (significant equity ownership as founder); Category: Equity Interest, Leadership Role

Satoru Inoda: None

Yuta Ueno: None

Yoshikazu Ito: None

Tetsuro Oshika: None

Masaki Tanito: None

Financial Support:

The Ministry of Education, Culture, Sports, Science and Technology of Japan, Grant-in-Aid for Scientific Research (21K16903). The sponsor or funding organization had no role in the design or conduct of this research.

Author's Contributions:

The first author, MK, was involved in all aspects of the study from conception to manuscript

writing. HT, TO and MT supervised the overall research. MK, HT and MT were responsible for data collection. MK developed and implemented the machine learning model-related code. MK, HT, SI, and YU conducted data analysis and interpretation. MK drafted the initial manuscript, and all authors actively participated in discussions to revise and improve the paper. Finally, all authors reviewed and approved the final version of the manuscript.