Article

The Frontloading Fields Study: The Impact of False Positives and Seeding Point Errors on Visual Field Reliability When Using SITA-Faster

Jack Phu[1,2] and Michael Kalloniatis[1,2]

[1] Centre for Eye Health, University of New South Wales, Kensington, New South Wales, Australia
[2] School of Optometry and Vision Science, University of New South Wales, Kensington, New South Wales, Australia

Correspondence: Jack Phu, Centre for Eye Health, Gate 14, Barker Street, Rupert Myers Building, South Wing, University of New South Wales, Sydney 2052, New South Wales, Australia.
e-mail: jack.phu@unsw.edu.au

Purpose: The purpose of this study was to evaluate the impact of two conventional reliability criteria (false positives [FPs] and seeding point errors [SPEs]) and the concurrent effect of low sensitivity points ($\leq$19 dB) on intrasession SITA-Faster visual field (VF) result correlations.

Methods: There were 2320 intrasession SITA-Faster VF results from 1160 eyes of healthy, glaucoma suspects, and subjects with glaucoma that were separated into "both reliable" or "reliable-unreliable" pairs. VF results (mean deviation and pointwise sensitivity) were analyzed against the spectrum of FP rates and SPE, with and without censorship of sensitivity results $\leq$19 dB. Segmental linear regression was used to identify critical points where visual field results were significantly different between tests due to FP levels.

Results: There was a significant, but small (0.09 dB per 1% exceeding 12%) increase in mean deviation, and an increase in the number of points showing a >3 dB sensitivity increase (0.25–0.28 locations per 1% exceeding 12%). SPEs were almost exclusively related to a decrease in sensitivity at the primary seeding points but did not result in significant differences in other indices. Censoring sensitivity results $\leq$19 dB significantly improved the correlation between reliable and unreliable results.

Conclusions: Current criteria for judging an unreliable VF result (FP rate >15% and SPE) can lead to data being erroneously excluded, as many results do not show significant differences compared to those deemed "reliable." Censoring of sensitivity results $\leq$19 dB improves intrasession correlations in VF results.

Translational Relevance: We provide guidelines for assessing the impact of FP, SPE, and low sensitivity results on VF interpretation.

Introduction

Visual field testing remains an integral part of the clinical assessment of glaucoma, as it provides a means to diagnose, prognosticate, and determine the impact of the disease.[1,2] Recommendations for visual field testing have highlighted the need to obtain at least 6 results within the first 2 years to obtain a robust impression of disease progression or stability,[3] This recommendation reflects the inherent variability of results obtained during testing, on the part of patient- and instrument-related factors.[4]

Although historical methods of perimetric testing have not been conducive to obtaining this many results in routine clinical practice due to the long test duration, recent algorithmic changes leading to faster testing protocols, such as SITA-Faster, have provided an opportunity to meet these recommendations.[5–7] The frontloading approach, doing more than one visual field test per clinical visit, has been proposed as a practical method for capturing sufficient clinical data to facilitate more confident clinical diagnosis and patient management.[8,9]

SITA-Faster has been found to have a slightly greater propensity to return results that do not meet commonly used "reliability" criteria in comparison to

its predecessor SITA-Standard.[6] Such results are often manually disregarded by the clinician or automatically by computer-based progression analysis techniques. Due to this limitation of apparent unreliability, clinicians may therefore question the value of SITA-Faster and its application in the frontloading approach, instead, potentially favoring SITA-Standard.

However, the impact of failing clinically used indicators of reliability, such as excessive false positives and seeding point errors, on the repeatability of frontloaded field tests is not well understood.[8,10] The question raised is whether all results that fail to meet reliability thresholds should indeed be discarded, or if there is still potential usefulness in at least part of the data. It is possible that such criteria, which are often regarded in a binarized pass/fail fashion, do not provide clinicians with the opportunity to retain, at least in part, some useful clinical data. For example, Heijl and colleagues[11] have recently presented findings suggesting that false positive metrics are not strongly associated with output perimetric indices using the SITA family of algorithms, recommending that historical cutoffs such as the 15% false positive limit should be revised. Overall, such questions raise the possibility that conventionally used parameters for assessing reliability may not truly reflect the usefulness of the test result.

The purpose of the present study was to examine the impact of two reliability metrics (false positive errors and seeding point errors) on pairs of frontloaded visual field tests performed within the same clinical visit using SITA-Faster. The central hypothesis was that there is a threshold at which reportedly unreliable clinical data remain comparable to that of its reportedly reliable counterpart. We used three approaches to test this hypothesis. First, we compared mean deviation and mean sensitivity found on reliable-unreliable pairs of visual field tests. We would then be able to determine if there was a threshold at which the difference in mean deviation or mean sensitivity began changing significantly. Second, we analyzed pointwise sensitivity changes across the visual field. Further to this, we also applied cluster analysis to determine if there were areas that were particularly likely to show greater differences in sensitivity in reliable-unreliable pairs. Third, we examined the effect of "correcting" test locations exhibiting low test reliability on the correlation between reliable-unreliable pairs. Alongside these approaches, we also examined the contribution of low sensitivity values (at or below 19 dB) on the correlations between results. Thus, in combination, these results would allow us to provide recommendations on extracting useful information from apparently unreliable visual field results obtained using SITA-Faster.

Methods

Ethics Statement

This was a cross-sectional study using prospectively acquired data from the files of patients seen within the Centre for Eye Health, University of New South Wales. Ethics approval was provided by the Human Research Ethics Committee of the University of New South Wales (HC210563). The study adhered to the tenets of the Declaration of Helsinki. All subjects provided their written informed consent for use of their de-identified clinical data for research purposes.

Subjects

Subject data were acquired from consecutive patients seen within the general and glaucoma service of the Centre for Eye Health, University of New South Wales between September 1, 2020, and March 31, 2021. The clinic is referral-only, optometry and ophthalmology service, providing assessment and management of patients with diseases of the visual pathways, including glaucoma.[12,13] The subjects were part of the Frontloading Fields Study (FFS), an ongoing study at the Centre for Eye Health examining the deployment of frontloaded SITA-Faster visual fields in clinical decision making and patient management.[8,9]

Subjects were divided into three possible diagnostic categories based on their clinical assessment outcome: no evidence of diseases of the visual pathway (healthy), glaucoma suspect, or manifest glaucoma.

Glaucoma was defined as per current clinical guidelines and our previous studies.[6,8,10,14] In brief, glaucoma was defined as the presence of glaucomatous structural defects (for example, optic disc cupping, diffuse or focal rim thinning of the neuroretinal rim, and adjacent retinal nerve fiber layer defects) with or without accompanying reproducible concordant visual field defects on the 24-2 test grid, in the absence of other retinal or neurological pathologies. Glaucoma suspect subjects were those in whom there were one or more signs of glaucoma (on disc or visual field examination) that were present or patients diagnosed with ocular hypertension or narrow anterior chamber angles, but where their combination was insufficient for a diagnosis of glaucoma requiring medical or surgical intervention. A healthy subject was one in whom there was none of the above signs. The diagnoses were extracted from the patient's medical record. As per the clinical protocols of the Centre for Eye Health,[12] a diagnosis was made by an examining clinician, with remote review by a senior clinician working within the
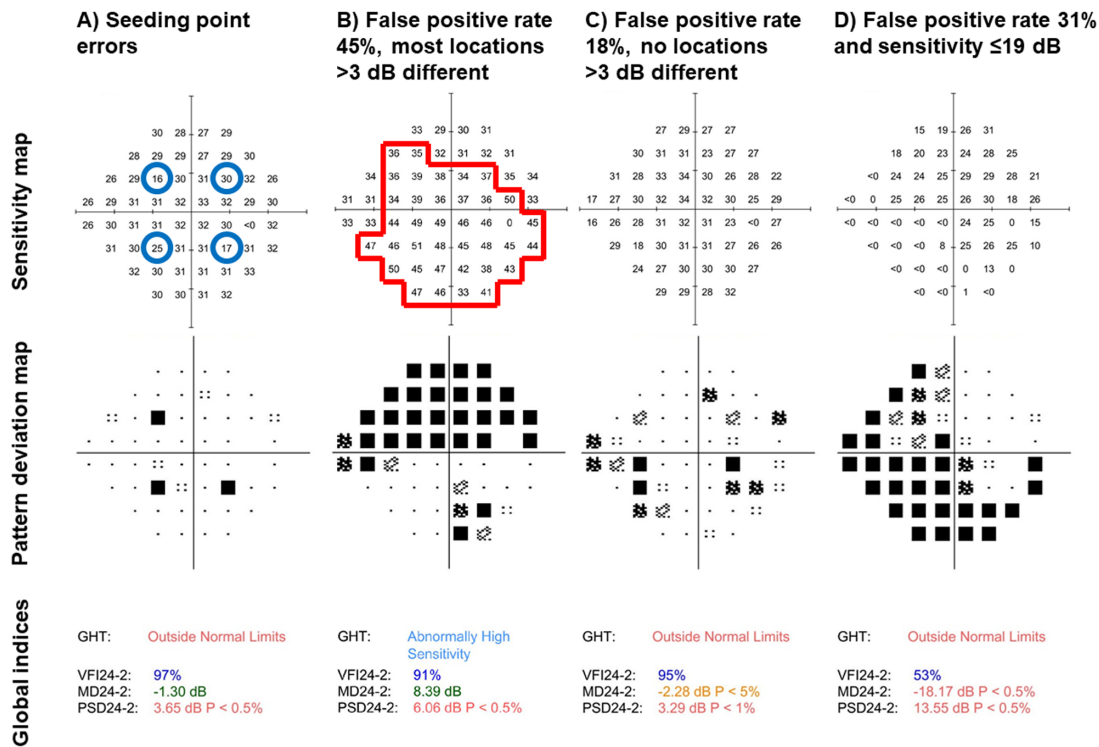
Figure 1. Examples of visual field results not meeting "reliability" criteria examined in the present study. Sensitivity maps (dB), pattern deviation maps, and select global indices are shown. (**A**) Seeding point error, where three of the four primary seeding locations are markedly reduced in isolation (*blue circles*). (**B**) False positive rate 45%, with most locations showing a sensitivity increase of >3 dB above the age-expected value (*red bordered area*). (**C**) False positive rate 18%, with no locations showing a sensitivity increase of >3 dB. (**D**) False positive 31% with a glaucomatous arcuate defect with sensitivity results less than or equal to 19 dB. See Methods for additional detail.

clinic. A third expert further examined the record for inclusion in the present study.

Although a previous classification scheme[15] has been proposed for standardizing the definition of glaucoma in prevalence surveys, its categorizations refer to cup-disc ratios, which are not utilized in more current clinical guidelines or practice patterns.[14] Specifically, as the scheme is focused on defining glaucoma in prevalence surveys (at the population level), it does not refer to longitudinal data or disease progression, both of which are used in clinical practice and applied at a granular, individual patient level.

Visual Field Test Reliability and Parametrization

For the present study, we utilized data from consecutive patients who returned the following combinations of visual field results: both tests reliable, first reliable and second unreliable, or first unreliable and second reliable.

We specifically focused on two criteria that are used to identify results as unreliable in clinical practice, which might lead to a result being excluded from analy-

sis. We note that our use of "reliable" or "unreliable" nomenclature specifically refers to the current clinical perception of these metrics, but not an objective measure of their ability to identify an uninterpretable visual field result. When performing our analysis, we instead use the terms "passed criteria" (as a surrogate for "reliable") and "failed criteria" (as a surrogate for "unreliable") with reference to the criteria set out below.

The first criterion was the presence of seeding point errors,[10] which is a relatively common occurrence arising due to the features of the SITA-Faster algorithm (Fig. 1A). In this error, at least one of the four primary seeding points initially tested in the grid is abnormally low in sensitivity, in the absence of pathology. The result of this error is one or more isolated points of artificial sensitivity reduction that may consequently affect calculation of the hill of vision and global indices. We used the following definition: at least one at the $P < 0.0001$, or the product of normative significance values of two or more points equal to or less than $P < 0.0001$.[8]

Second, false positive rates were extracted as a percentage reported by the instrument. A cutoff value

of 15% or greater is typically used as a criterion for an unreliable result, as such, a cutoff was uncommonly seen in reliable perimetry (see Fig. 1B–D).[16] Notably, this cutoff value is more stringent (lower) compared to the 33% limit used in older clinical trials.[17] We extracted the absolute false positive rate for an individual test, as well as the difference in false positive rate between unreliable (>15%) and reliable tests.

Notably, Figure 1 shows 3 examples of elevated false positive rates that do not meet the 15% cutoff (Figs. 1B, 45%; 1C, 18%; 1D, 31%). In Figure 1B, most test locations had clearly elevated sensitivity results, whereas there were no instances where the sensitivity result was more than 3 dB above age-expected limits in Figure 1C. In Figure 1D, we show an example where false positive rates are elevated (31%), but with an arcuate pattern of loss that has sensitivity values less than or equal to 19 dB (which we defined as an alternate measurement floor[18,19] – see more below) where measurements are not expected to be highly repeatable.

We analyzed files that met only one of the above criteria, as multiple sources may confound each other. In addition, we did not analyze gaze tracker errors[20] in the present study, as these add another confounding layer of uncertainty, and also because it is a scalar, but not vector, measurement, rendering it difficult to elucidate its effect on the visual field measurement. Although current recommendations for assessing gaze tracker errors are largely qualitative,[21] for the purposes of the present study, we excluded visual field results where over 20% of gaze tracker deviations exceeded 6 degrees, as per our previous studies.[6,22]

The role of the group of patients exhibiting visual field results that did not meet either criterion (i.e. deemed "passed criteria") in both instances was twofold. First, it would serve as the reference group for intra-session retest variability of global and pointwise measurements of interest to which the unreliable results could be compared (see more below). Second, given that the false positive unreliability criterion uses a cutoff of 15%, this would enable the analysis of a spectrum of lower false positive rates up until the cutoff point.

Visual Field Data Extraction

As per current clinical protocols at the Centre for Eye Health, all patients underwent visual field testing twice for each eye within the same test session. The order of testing was at the discretion of the administering technician, with rest breaks between each test as requested by the patient. All testing was performed using the Humphrey Field Analyzer 3 instrument, using the 24-2 test grid and SITA-Faster algorithm (Carl Zeiss Meditec, Dublin, CA).

Visual field data of interest were the right and left eyes (or only one eye in cases where the patient was monocular) results collected within the same clinical visit. A custom written Matlab program (The Mathworks, Natick, CA) was used to extract the following parameters of interest from each visual field printout: pointwise visual field sensitivity, mean deviation, pattern standard deviation, test duration, and false positive rate.

We examined the role of reliability metrics on visual field outputs using the following three approaches.

Approach 1: Analysis of Mean Deviation

Mean deviation is a commonly used global index of visual field integrity for glaucoma assessment and staging,[23] effectively representing the average sensitivity reduction in an individual's result relative to the age normative reference. We analyzed mean deviation differences between reliable and unreliable visual field results as a function of false positive rates, and binarized present/absent seeding point errors. Because there is a possible spectrum of false positive responses, we plotted false positive rates as both a difference (between reliable and unreliable pairs) and as the highest absolute value (between the two results) as the independent variable to examine its effect on the difference in mean deviation. Using a segmental linear regression analysis (slope 1 set to 0, and the inflection point X_0 and slope 2 were calculated), we could determine if a level of false positive rate exists where the difference in mean deviation becomes clinically significant. The purpose of a segmental linear regression, in comparison to an exponential function, was to identify an inflection point (not set a priori) at which no significant effect (the flat portion denoted by slope 1) demonstrates an increase in the dependent variable.

Approach 2: Analysis of Pointwise Sensitivity Results: Difference in Sensitivity and Cluster Analysis

The 24-2 test grid returns 52 sensitivity values, and we examined the pointwise differences between the "passed criteria" and "failed criteria" visual field test results. We identified the proportion of locations exhibiting a difference greater than 3 dB, as an illustrative threshold for variability,[6] noting that this threshold can be adjusted depending on the severity of sensitivity loss.

In addition to the above criterion, we also identified locations that were at or below an alternate measurement floor. Nineteen dB has been shown by others to represent the approximate level below which perimetric data becomes less reliably measured.[18,19] Thus, we performed a secondary analysis using the above approach after excluding all test locations that were returned a sensitivity value of 19 dB or less.

Aside from counting the proportion of locations showing a difference exceeding expected retest variability, we also applied a cluster analysis approach to determine if there were systematic regions of interest sharing the same difference characteristics (see Supplementary Material for further methodological details).[24–26] A secondary goal of the study was the identification of locations showing systematic differences between "criteria passed" and "criteria failed" visual field results that may lead to development of a correction factor to overcome visual field artifacts.

Approach 3: "Correcting" for Anomalous Sensitivity Results

Aside from examining the extent to which current reliability metrics can be used to identify altered visual field test results, we also sought to determine if methods for correcting for erroneous sensitivity measurements might provide more useful clinical information.

The relationships identified in approaches 1 and 2 might return methods for obtaining a correction factor which could be applied to unreliable results. To analyze this, we used mean sensitivity and measured changes in the model using the coefficient of determination and the width of the 95% prediction interval.

Because the calculation of mean deviation involves an instrument-specific, proprietary modulation on top of the individual's sensitivity result and is scaled across different visual field locations, we also calculated and compared mean sensitivity. Calculation of mean sensitivity has been previously detailed in other papers.[27–29] First, the conversion of the decibel value (dB) returned by the instrument to linear luminance threshold (in cd.m^{-2}; Equation 1). Then, the linear contrast values were averaged to represent the linearized sensitivity. Finally, the average linear sensitivity was then converted back into a decibel value (Equation 2).

$$\Delta L = \frac{3183}{10^{\frac{dB}{10}}} \qquad (1)$$

$$Mean\ sensitivity = 10\ \times log_{10}\left(\frac{3183}{Average\ luminance}\right) \qquad (2)$$

Results

Of the 1575 eyes of 779 patients seen within the data extraction period, 913 (57.3%) had both fields which were reliable, 183 (11.5%) had first only "passed," 293 (18.4%) had second only "passed," and 186 had both "failed." Out of the 476 eyes with a pair of "passed" – "failed" fields, 138 (29.0% within the total "failed" group) had false positive rate >15% and 109 (22.9% within the unreliable group) had seeding point errors alone, which were used for analysis (1826 both "passed," 276 with false positive rate >15%, 218 with seeding point errors totaling 2320 visual field results analyzed).

The characteristics of the 1160 eyes analyzed in the present study are shown in the Table. There were no differences in the distributions of age and diagnoses between groups. There were more women in the seeding point error group, and more left eyes showing both results having "passed criteria." The latter result was expected, most likely due to a combination of learning, practice, and instruction effects, as we have previously discussed.[8] In brief, this was because whereas the perimetrists were permitted to test the eyes at their discretion, the right eye was tested first >90% of the time (724/779 patients contributing to "pass" and "fail" pairs noted above). After initial experience with the test, the left eye, tested second, returned more instances of "passing" the reliability criteria. The overall mean deviation value was lower in the seeding point error group, but the range was narrow, with no instances of patients with more advanced loss. This was likely due to the definition of seeding point errors, in which a prominent defect in the seeding point may be attributable to pathological loss. There were also differences in the distributions of ethnicities across the reliability categories.

Approach 1: Analysis of Mean Deviation

The difference in mean deviation as a function of false positive rate is shown in Figure 2A (difference in false positive rates: "failed" – "passed") and Figure 2B (highest false positive rate). With an increase in false positive rates, there was a tendency for the "failed" visual field result with elevated false positive rates to show a "better" mean deviation score, as expected. The inflection point differed between Figures 2A (12.7%) and 2B (22.7%) as they represented the difference in false positive rate and absolute highest false positive within a pair, respectively. There were also no significant differences between the seeding point error group and the group with "passed" results on both tests

Table. Demographic and Diagnostic Parameters of the Patients Whose Eyes Were Used for the Present Study, Categorized by Their Reliability Output

|		Both Reliable (n = 913)	"Passed" – "Failed" Pair (False Positive Rate >15%) (n = 138)	"Passed" – "Failed" Pair (Seeding Point Error) (n = 109)	P Value
Age, years, mean (standard deviation)		58.7 (13.5)	57.8 (13.5)	61.5 (13.2)	0.2266
Gender, n (%)	Male	486 (53.2)	73 (52.9)	44 (40.4)	0.0386
	Female	427 (46.8)	65 (47.1)	65 (59.6)	
Ethnicity, n (%)	Caucasian	477 (52.2)	59 (42.8)	57 (52.3)	<0.0001
	East Asian	320 (35.0)	51 (37.0)	43 (39.4)	
	South Asian	74 (8.1)	17 (12.3)	5 (4.6)	
	South American	20 (2.2)	1 (0.7)	3 (2.8)	
	Aboriginal Australian	13 (1.4)	4 (2.9)	0 (0)	
	African	7 (0.8)	0 (0)	1 (0.9)	
	Pacific Islander	1 (0.1)	2 (1.4)	0 (0)	
	Mixed	1 (0.1)	4 (2.9)	0 (0)	
Diagnosis, n (%)	Normal	227 (24.9)	41 (29.7)	23 (21.1)	0.4012
	Glaucoma suspect	497 (54.4)	72 (52.2)	58 (53.2)	
	Glaucoma	189 (20.7)	25 (18.1)	28 (25.7)	
Eye examined, n (%)[1]	Right eye	423 (46.3)	80 (58.0)	58 (53.2)	0.0220
	Left eye	490 (53.7)	58 (42.0)	51 (46.8)	
Mean deviation	Median, IQR and full range	−0.92, IQR −2.38 to +0.04, full range −26.31 to +2.53	−0.01, IQR −0.91 to +0.74, full range −23.19 to +3.79	−1.54, IQR −2.54 to −0.71, full range −7.52 to +1.27	<0.0001
	Number with mean deviation lower than −5 dB	78 (8.6)	2 (1.4)	5 (4.6)	0.0061

[1]Note that these proportions do not reflect a specific test order.

The *P* values indicate the comparison between groups (one-way ANOVA for age and Fisher's exact test for the categorical values).

IQR, interquartile range.
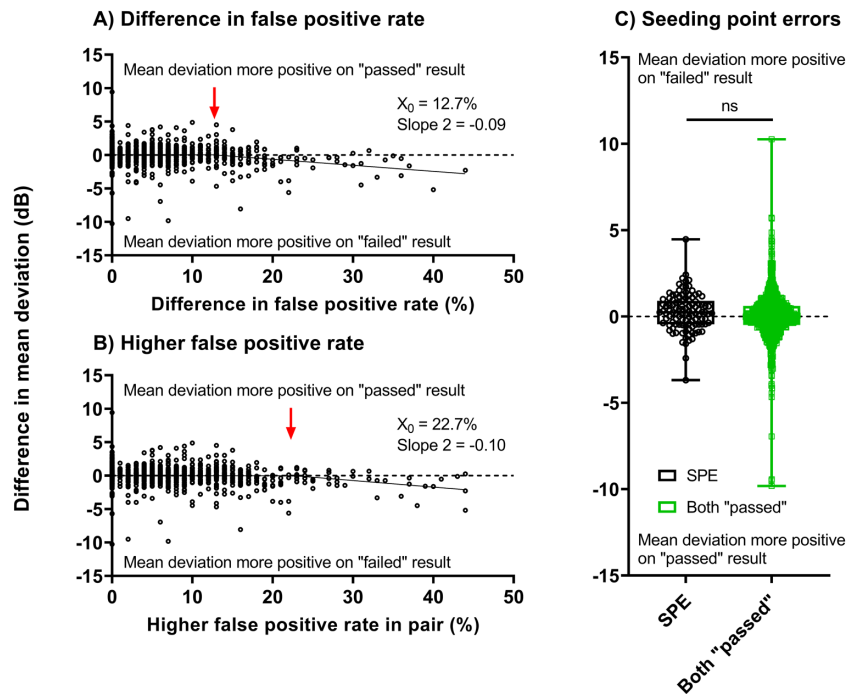
Figure 2. Difference in mean deviation (dB) between "passed" and "failed" visual field results by criteria. A positive y-axis value indicates that the mean deviation was better (more positive) on the "passed" result, and a negative value indicates that the mean deviation was better on the "failed" result. (**A**) Difference in mean deviation as a function of difference in false positive rate ("failed" – "passed" result). (**B**) Difference in mean deviation as a function of the higher false positive rate within the pair of results. For **A** and **B**, a segmental linear regression was performed, indicated by the *black solid line*, with the point of inflection (X_0) and second slope shown in the inset. The point of inflection is also identified by the *red arrow*. (**C**) Distribution of difference in mean deviation found in the seeding point error (SPE) and groups with both results "passed." The box and whiskers indicate the median, interquartile range, and full range. Each datum point indicates the result from one eye. The black dashed line indicates y = 0 (no difference in mean deviation).
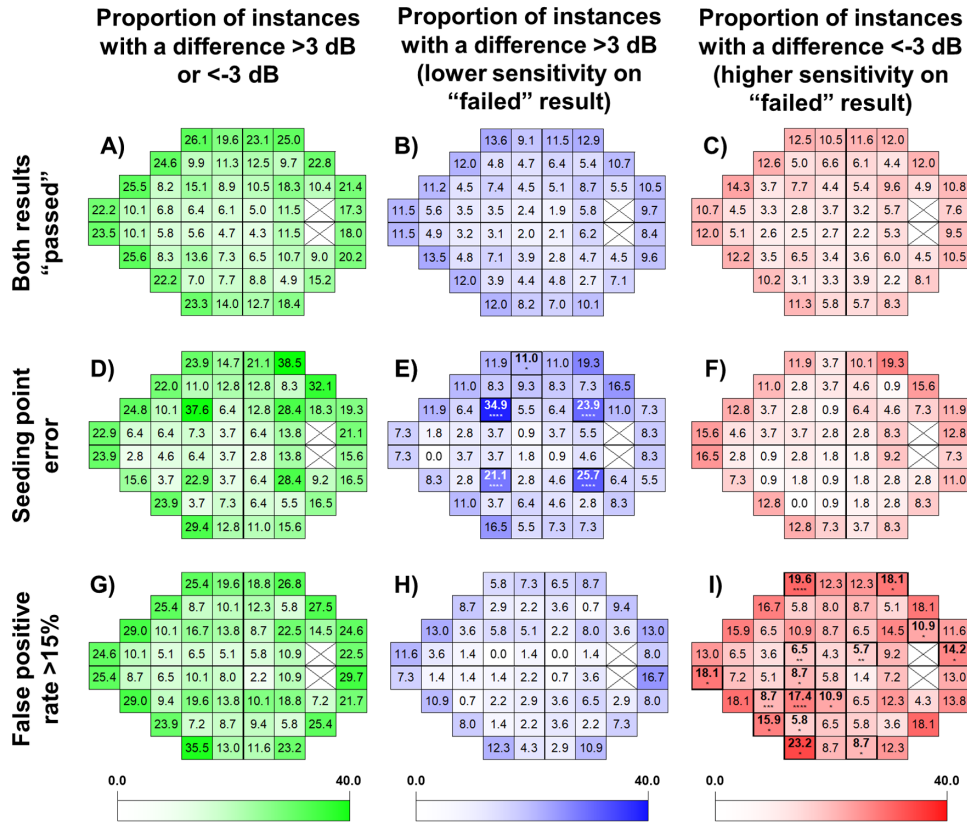
Figure 3. Heat maps showing the proportion of instances with a difference exceeding 3 dB between "passed" and "failed" results (*left column, green color code*), difference greater than 3 dB (lower sensitivity on the "failed" result; *middle column, blue color code*), and difference less than −3 dB (higher sensitivity on the "failed" result; *right column, red color code*). Numerical proportions are shown within each cell, indicating the position within the 24-2 test grid. The crosses indicate the two locations next to the physiological blind spot, which were excluded from analysis. The cells with dark borders, bolded text, and asterisks (* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; **** $P < 0.0001$) in seeding point errors (*middle row*) and false positive rates greater than 15% (*bottom row*) indicate locations where the proportion was significantly different to the distribution of differences seen when both results were reliable (*top row*). The key to the color code is shown below each column. Note that at some locations the sum of proportions with greater than 3 dB difference (*blue*) and less than −3 dB difference (*red*) did not exactly equal the total proportion (*green*) due to decimal rounding.

($P = 0.1002$; Fig. 2C). Upon inspecting the data and the relatively narrow band of mean deviation differences at the upper limit of x-values, an exponential function was not considered further.

Approach 2A: Analysis of Pointwise Sensitivity Results: Difference in Sensitivity

We mapped the distributions of differences between "passed" and "failed" visual field test results (or in the case where both "passed," the difference between randomized tests) across the 24-2 test grid. From these distributions, we counted the proportion of subjects who exhibited differences between test results exceeding 3 dB (where < -3 dB indicates that the "failed" result returned a higher sensitivity result, and where > 3 dB indicates that the "failed" result returned a lower sensitivity result). The pairs where both results

"passed" (Figs. 3A–C) were used as the reference proportions to which the "passed" – "failed" pairs were compared using Fishers exact test. This was to identify whether specific error types were associated with greater proportions of either increased or decreased sensitivity at certain locations. We describe the results shown in Figure 3 in greater detail below.

As expected, the seeding point error group showed the four primary seeding points returning a high proportion of instances with significantly lower sensitivity, indicated by the darker blue points (Fig. 3E). There were three other locations that showed a greater likelihood of having lower sensitivity, but the overall proportion of instances where the difference exceeded 3 dB at those locations was similar to all the other locations.

The elevated false positive rate group showed several locations that had a statistically elevated proportion
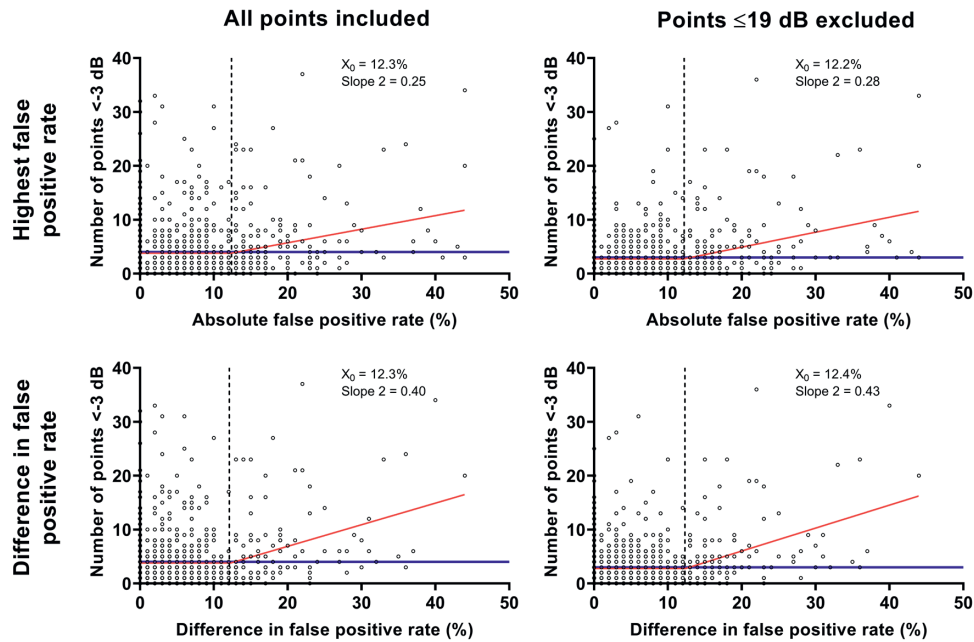
Figure 4. The number of points where sensitivity values were more than 3 dB greater found as a function of false positive rate. With "passed" – "failed" pairs, this indicated the number of occasions where the "failed" result was more than 3 dB compared to the "passed" result. When both results were "passed," we compared the number found on the result with the relatively higher false positive rate, or if both were the same, in random order. A higher value on the y-axis indicates more points showing elevated sensitivity. Each datum point indicates the result from one eye. The *blue solid line* indicates the average number of points where there was a 3 dB increase in sensitivity when both results were "passed." The *red solid line* indicates the segmental linear regression with the point of inflection (X_0) and second slope shown in the inset. The *left column* indicates the results when all test locations were included, and the *right column* indicates the results when points reaching an alternate measurement floor (19 dB) were excluded. The *top row* indicates results as a function of absolute false positive rate and the bottom row indicates results as a function of difference between the higher and lower false positive rates. For each regression analysis, a *vertical black dashed line* indicates the point of inflection.

of instances where sensitivity was higher on the "failed" result. The locations were spread across all four quadrants and were mainly located at the edges of the test field, as indicated by the darker red cells (Fig. 3I).

Specifically for the elevated false positive criterion, we plotted the number of instances over which the difference exceeded 3 dB as a function of false positive rate (Fig. 4). In this analysis, we plotted the number of instances for cases where both results were "passed" and pairs where one of the results had a false positive rate greater than 15%. The cases where both results were "passed" served as a reference point for when the number of instances where the difference exceeded 3 dB changed significantly (when the slope was fixed at 0, indicating no effect of the independent variable, false positive rate). The inflection points across all conditions were identical at approximately 12%. Every percentage increase in absolute false positive rate increased the number of instances by 0.25 to 0.28, and every percentage increase in relative false positive rate ("passed" to "failed" at >15%) increased the number of instances by 0.37 to 0.40. Similar to the distribution of mean deviation data, we did not proceed to fit the

sensitivity data with an exponential function. Unlike the distribution of mean deviation data, there was no difference in the inflection point between absolute and relative difference in false positive rates, likely due to the greater amount of variability in data. Instead, the differences manifested as a change in the slope parameter.

Approach 2B: Cluster Analysis Applied to Pointwise Sensitivity Differences Across the Test Grid

We applied cluster analysis to the distributions described above to determine groups of test locations that had similar differences between pairs of tests. When both were "passed," there were two separable clusters, but the average difference between clusters was small at 0.3 dB (Supplementary Fig. S1A). With elevated false positive rates, there were also 2 clusters, with 51 of 52 locations sharing a common distribution where the "failed" result was, on average, 0.6 dB higher than the "passed" result (see Supplementary Fig. S1B). The seeding point errors, as expected,
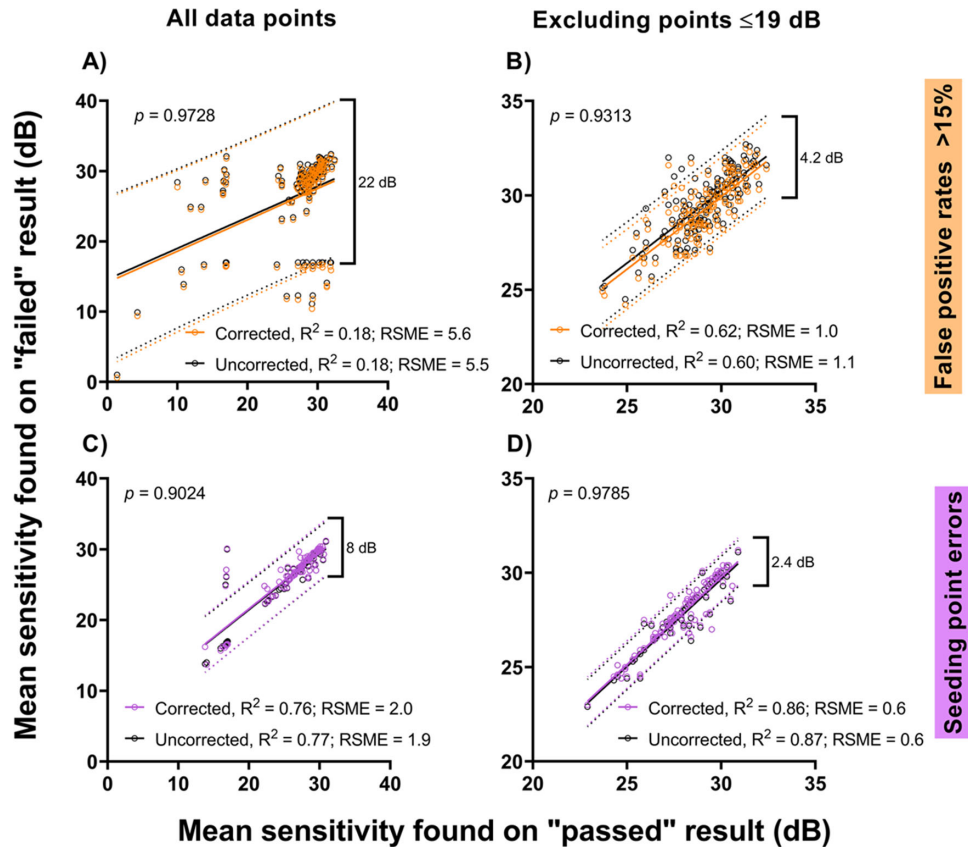
Figure 5. Correlation between "failed" result mean sensitivity (dB) and "passed" result mean sensitivity (dB) pairs for false positive rates >15% (*top row*) and seeding point error (*bottom row*) groups. The results from Figures 3 and 4 and Supplementary Figure S1 were used to create a model used to correct the "failed" visual fields, excluding test locations that were statistically likely to be unreliably elevated or depressed (*orange* for false positive rates >15% and *purple* for seeding point errors). The corrected mean sensitivity was compared with the uncorrected visual field result (*black*). Linear regression analysis is shown by the *solid lines* (R^2 values and root mean squared error [RSME] for corrected and uncorrected data are shown in the *inset*), and the *dotted lines* indicate the 95% prediction intervals (the width of the interval is shown by the *brackets*). The *left column panels* indicate the results when all data points were included, and the *right column panels* indicate the results when points reaching the alternate measurement floor (19 dB) were excluded. The 95% prediction intervals were notably narrower when using comparing all points and the condition where points reaching the measurement floor were excluded.

showed the 4 primary seeding locations as belonging to separate distributions (cluster 2 showing 1.9 dB lower sensitivity and cluster 3 showing 2.7 dB lower sensitivity found on the "failed" result), whereas all other locations showed an average difference of 0 dB (see Supplementary Fig. S1C), supporting the results shown in Figure 2C.

Approach 3: "Correcting" for Anomalous Sensitivity Results

Approaches 1 and 2 showed that although statistically significant, the overall magnitude of sensitivity difference caused by "failed" visual field results by elevated false positives and seeding point errors was, on average, small across the entirety of the cohort. The prediction for approach 3 was that mean sensitivity

would be similar when considering the ground truth (the "passed" result of the pair serving as the reference standard), corrected and uncorrected "failed" results.

The model in Supplementary Figure S1 was used to identify the number of points that were possibly erroneously elevated due to the high false positive rate in the "failed" result. For each "failed" result with an elevated false positive rate, the highest sensitivity results across this number of visual field test locations were removed, and a new mean sensitivity was calculated. Application of the correction showed no significant difference in fit quality (using the coefficient of determination and the root squared mean error) compared to the uncorrected data (Figs. 5A, 5B).

Given the well-defined locations affected in seeding point errors, the approach we used was the censorship of the primary seeding points. Removal of the seeding point locations returned no improvement in the

correlation between "passed" and "failed" visual field mean sensitivity (Figs. 5C, 5D).

In both instances, the main driver of improving the relationship between "passed" and reportedly "failed" results was the censorship of sensitivity results at or below 19 dB (representing 1.8–2.2% of results), with significant reduction in the magnitude of the root mean squared error. Furthermore, the 95% prediction bands were narrower following censorship: from 22 dB to 4.2 dB for elevated false positive rates, and from 8 dB to 2.4 dB for seeding point errors, improving potential clinical interpretation through consistency of the results. Overall, the results in Figure 5 support the results shown in Figure 2, which show small magnitudes of global visual field differences that would unlikely be clinically significant.

Discussion

The present study sought to systematically determine the impact of two clinically used "reliability" parameters found in SITA-Faster visual field results, the elevated false positive rates and seeding point errors. An absolute and relatively higher false positive rate of 12% to 13% was associated with significantly higher sensitivity measurements and was also associated with a greater number of test locations with sensitivity values greater than 3 dB from the reference result. Seeding point errors predictably led to lower sensitivity measurements at the 4 primary seeding locations by approximately 2 to 3 dB. Despite the systematic characteristics of sensitivity changes arising from these two parameters, the differences were small in magnitude and thus would not be clinically significant. Thus, these results raise the question of whether perimetric results that fail these manufacturer-defined reliability criteria need to be excluded from clinical interpretation or progression analysis, and whether these criteria are antiquated. Irrespective of the error type, censorship of sensitivity results at or below 19 dB improved correlation between intrasession results.

"Reliability" Parameters and the Perimetric Algorithm

A goal of clinical perimetry is to obtain useful threshold measurements across the visual field within a practical test duration. Development of adaptive and fast techniques for achieving this goal has been a research focus, and SITA-Faster is an example of commercially available modern test algorithms. The modifications made to old SITA paradigms leading to SITA-Faster have been previously described.[5] Psychophysically, three specific modifications have potentially resulted in a greater propensity of higher false positive rates and seeding point errors compared to SITA-Standard. In brief, these were: the nearer to threshold starting stimuli at the seeding points, only one staircase reversal at the seeding points, and removal of delay after non-seen stimuli. Attentional factors and the lack of a "second chance" at obtaining thresholds at test initiation have been described as the reasons for seeding point errors.[10] Similarly, shortened intervals between stimuli, particularly in older patients, may increase the likelihood of elevated false positive results if assessed using a historical measurement technique and threshold.

Another potential contributor to apparently elevated false positive rates and seeding point errors is the learning effect.[30–32] Part of this effect may be mitigated by the frontloading approach, which is in part supported by the greater proportion of patients showing test 1 being unreliable, compared to test 2 or later.[8] Patients in the present study were nearly all perimetrically experienced. Thus, the effects of false positive rates and seeding point errors seen within the present study are more likely reflective of algorithmic characteristics.

Practical Recommendation 1 for Interpreting Reliability Metrics: False Positive Rates

Recently, Heijl and colleagues[11] examined the effect of false positive rates on intrasession perimetric results using the SITA algorithms, including SITA-Faster, in a cross-sectional setting. As expected, our results were similar to that of Heijl and colleagues,[11] where higher false positive rates were associated with higher mean deviation scores. Our rate of increase of mean deviation score per percentage point of false positive rates was slightly higher than Heijl and colleagues[11] (0.9 dB per 10% increase in relative false positive rates above 13%, or 1 dB per 10% increase in absolute false positive rates above 23%, compared to 0.3–0.6 dB per 10%). We note that we used a segmental linear regression which identified a larger change in mean deviation in excess of a false positive rate of 12% to 13%. Our "cutoff" point for when the difference in mean deviation increased significantly was at an absolute false positive rate of 23%, with the rate of increase in mean deviation difference remaining small. We postulate that this may also be because of the different ranges of false positive ranges, where our study had a larger range of false positive values (up to 45%), leading to a potentially more pronounced sensitivity elevation.

The rate of increase in mean deviation was slightly lower than that reported by Yohannan and colleagues.[33] However, we note that our methods were slightly different in several ways. We did not set the inflection point for a significant change a priori, allowing us to identify the point at which the slope change in the score as a function of false positive rate became statistically significantly different to 0. We also compared intrasession results using SITA-Faster (which has been documented to return higher rates of false positive rates[6,8]), and our sample had a higher proportion of results with false positive rate >20% (approximately 5.5%).

Additionally, we found a statistically significant, but overall small and poorly explained relationship between number of test locations that returned 3 dB higher sensitivity on the high false positive results, again at approximately 13%. Similar to the effect on mean deviation, the effect of elevated false positive rates on the number of test locations with elevated sensitivity was small, with 0.25 to 0.28 locations per percentage point of the false positive rate in excess of an absolute false positive rate of 13%. When comparing intrasession visual field tests, the slope describing the number of test locations and difference in false positive rates was steeper (0.40–0.43 per 1%). Although this has implications for assessing the repeatability of cluster criteria in intrasession visual field tests, again, the difference in the false positive rate needs to be in excess of 12% between tests. Both mean deviation and sensitivity change data were fit using a segmental linear regression, rather than an exponential function. This was due to the relatively sparse sample of subjects exhibiting large magnitudes of mean deviation and sensitivity differences at the upper limits of false positive rates. A sample with more diverse visual field artifacts may serve further to explore this potential exponential relationship.

The slight tendency for more peripheral test locations exhibiting elevated sensitivity found on cluster analysis was unsurprising given the known greater effects of spatial uncertainty in those test regions.[34,35] However, the overall frequency of the elevated sensitivity results was low. At the individual level, this small magnitude of effect of false positives was seen through the lack of improvement in correlation following correction of sensitivity results.

In combination, the small effect size on the commonly used mean deviation and pointwise sensitivity across most of the visual field suggests that false positive rates should be regarded along a continuum. Therefore, based on ours and Heijl and colleagues' results (who notably had a different study design),[11] there is evidence across different testing modalities that

the historical precedent of 15% false positive rate as a cutoff for reliability should be reconsidered, as useful information can still be obtained for results with apparently high false positive rates. For example, if we expect that the difference in the mean deviation score between frontloaded tests to be within an illustrative magnitude of ±2 dB 95% of the time, a false positive rate of up to 32% (20% above the 12% level) might still be within the range of the expected variability. When combined with the relationship found between mean deviation difference and absolute false positive rate, the results suggest that one of the results would need to have a false positive rate of approximately 43% for a 2 dB difference in mean deviation (20% above the 23% level). At this false positive rate, one might also expect to see 8 to 10 test locations with falsely elevated sensitivity readings. Careful clinical examination of the sensitivity maps would be further revelatory.

However, the converse may also occur, whereby low false positive rates that are "within" the current cutoff of 15% may also be accompanied by artificial elevations in mean deviation. Such false positive rates also do not preclude falsely elevated sensitivity measurements. Thus, when assessing visual field results, false positive rates should be used to as a guide, rather than act as a dogmatic binarized pass-fail criterion.

Practical Recommendation 2 for Interpreting Reliability Metrics: Seeding Point Errors

Seeding point errors present as an obvious 1 to 4 point artifact, especially markedly within the probability map.[10] These may consequently affect the thresholding of adjacent points and their analyses. Sensitivity reductions at these locations are likely to affect pointwise progression analysis, but seem unlikely to affect global metrics such as mean deviation in a clinically meaningful way. It should be noted that seeding point errors are more challenging to identify in cases with visual field defects, as the definition requires a reduction in sensitivity with typically normal or near-normal sensitivity at adjacent locations. This was reflected by the small number of overall cases with low mean deviation values in the seeding point error cohort.

We have previously proposed methods for identifying the presence of seeding point errors to mitigate their effects on the final result.[10] In the present work, comparisons of frontloaded pairs of visual fields suggests that the impact of seeding point errors are, as expected, generally localized to the seeding points, with minimal effect on global scores and correlations with reliable results. Disregarding these artificially depressed points would potentially mitigate errors in interpreting pointwise progression analysis at the relevant locations.

The recommendation from these results is to continue to disregard the locations presenting with seeding point errors when assessing perimetry results, but to repeat the test if the locations are relevant to the assessment of scotomata with a pathological etiology.

Practical Recommendation 3 for Interpreting Reliability Metrics: Censorship of Points Reaching the Measurement Floor

An analysis of global metrics and pointwise sensitivity revealed no significant, systematic difference between reliable and unreliable visual field result pairs in most instances. Instead, the main contributor to discordance between results appeared to be situations in which an alternate visual field measurement floor (19 dB) was reached. Previously explored in depth, this sensitivity measurement level represents the lower limit of reliable perimetric measurements using current standard automated perimetry techniques.[18,19]

In the present study, censorship of points reaching 19 dB or below continued to improve the correlation between "passed" and "failed" results. This was especially pronounced when examining results that had elevated false positive results. An implication is that techniques and metrics for measuring low test reliability are confounded by the worsening visual field, which been previously demonstrated.[36] Given the algorithmic changes that have contributed to the increased frequency of these criteria being "failed" in SITA-Faster, in combination, this suggests the role of uncertainty in returning "lower" test reliability. This theory is supported by previous work showing increases in uncertainty in regions of worse visual field loss.[19,34] This represents an ongoing clinical issue as patients with worsening visual field result are also those who require close monitoring of their functional status. These results therefore suggest the need to develop better strategies for monitoring sensitivity changes within these patients after accounting for worsened correlations between tests.

Practical Implications for Frontloading and Obtaining "Reliable" Visual Field Data

Our previous study showed that a large proportion of SITA-Faster visual field tests return false positive rates exceeding 15%.[6] From this, estimates of the proportion of test results that may therefore meet traditional "reliability" criteria tempered the potential time-saving benefits of performing SITA-Faster in lieu of SITA-Standard. The frontloading approach was designed to overcome this potential limitation. When two visual field tests were performed per clinical visit, it mitigated instances of high false positive rates, thereby providing clinicians with at least one visual field result meeting traditional "reliability" metrics over 90% of the time.[8] Although most patients had more than one clinically useful visual field result from the frontloading approach, the "trade-off" was that in approximately 30% of cases, there was time spent performing additional visual field tests to overcome the elevated false positive rate (amongst other potential sources of low test "reliability").

The implications of the present study and the above practical recommendations instead suggest that the application of these traditional "reliability" metrics results in an overestimate of low test reliability, and injudicious discarding of potentially useful clinical results. Thus, the frontloading approach would be expected to return a significantly greater proportion of useful results compared to those previously estimated, increasing the time efficiency per useful visual field result. Importantly, this means that more data can be included for the purposes of automated (or manual) progression analysis.

We recently modeled the application of the frontloading approach for detecting glaucomatous mean deviation change.[22] In situations where more data are obtained using a frontloading approach, the time to detect significant mean deviation change is significantly reduced compared to the current clinical standard of performing one test per visit.[22] Thus, although each clinical visit may be longer, there are potentially significant benefits for case detection of visual field change, and the time-saving benefits may occur by requiring fewer future clinical visits required to detect this change.

Limitations

Our consecutive sampling approach was performed to minimize selection bias, and to reflect the real-world probability of error identification in visual field testing. As such, our sample did not have a high prevalence of more advanced cases of glaucoma and vision loss. In relation to the cohort tested, there were some differences in the distributions of self-reported gender and ethnicity in the reliability outcomes, but the reasons for this were not explored in the present study. For this reason, our comments and recommendations are not directly applicable to cases of advanced visual field loss, nor do we provide information on the impact of these metrics as a function of the magnitude of defect. Importantly, significant false positive errors may mask specific scotomata, confounding interpretation and progression analysis. Patterns of falsely elevated

sensitivity results that are inconsistent with age-expected normative values, historical data, or structural parameters – even independent of the false positive metric – should alert clinicians to the possibility of a false positive result.

We also focused on two distinct reliability metrics: elevated false positive rates and seeding point errors. These error types were predicted to return opposing effects on sensitivity: false positive rates cause higher sensitivity and seeding point errors leading to lower sensitivity. There may be interactions between these metrics, as well as contributions from other metrics, such as gaze deviations. This analysis would require a more complex approach and would benefit from future study.

Conclusions

Elevated false positive rates and seeding point errors are common occurrences in SITA-Faster, with worsened correlations between results driven significantly by sensitivity readings less than or equal to 19 dB. However, injudicious application of historical elevated false positive rates and seeding point errors may erroneously exclude useable clinical data. We therefore provide the following three recommendations:

Recommendation 1: Reconsider the historical precedent of 15% false positive rate as a cutoff for reliability, as higher rates of false positives may still produce useful results. Because the converse may also be true (low false positive rates may be accompanied by falsely high mean deviation results), a dogmatic approach to false positive cutoffs is not recommended.

Recommendation 2: The presence of seeding point errors should prompt clinicians to disregard erroneously seeded points, but not necessarily to disregard results at other test locations. Repeating the test is specifically recommended when erroneously seeded points are relevant to cross-sectional or longitudinal interpretation of potential scotomata.

Recommendation 3: Censorship of sensitivity values at or below 19 dB improves intrasession pointwise sensitivity correlations between visual field results. However, results at or below 19 dB should still be integrated into pointwise progression analysis and global metrics like mean deviation.

In situations where metrics, such as elevated false positive rates or seeding point errors, are found to confound accurate visual field interpretation cross-sectionally or longitudinally, the recommendation remains to repeat the test to attempt to overcome these artifacts. Although, in many instances, their effects may be small, recognition of these artifacts, especially in the context of the patient's aggregate clinical findings, remains critical for both automated and manual methods of visual field interpretation.

Acknowledgments

Disclosure: **J. Phu**, None; **M. Kalloniatis**, None

References

1. Jampel HD, Singh K, Lin SC, et al. Assessment of visual function in glaucoma: a report by the American Academy of Ophthalmology. *Ophthalmology*. 2011;118(5):986–1002.
2. Phu J, Khuu SK, Yapp M, et al. The value of visual field testing in the era of advanced imaging: clinical and psychophysical perspectives. *Clin Exp Optom*. 2017;100(4):313–332.
3. Chauhan BC, Garway-Heath DF, Goni FJ, et al. Practical recommendations for measuring rates of visual field change in glaucoma. *Br J Ophthalmol*. 2008;92(4):569–573.
4. Stewart WC, Hunt HH. Threshold variation in automated perimetry. *Surv Ophthalmol*. 1993;37(5):353–361.
5. Heijl A, Patella VM, Chong LX, et al. A New SITA Perimetric Threshold Testing Algorithm: Construction and a Multicenter Clinical Study. *Am J Ophthalmol*. 2019;198:154–165.
6. Phu J, Khuu SK, Agar A, Kalloniatis M. Clinical Evaluation of Swedish Interactive Thresholding Algorithm-Faster Compared With Swedish Interactive Thresholding Algorithm-Standard in Normal Subjects, Glaucoma Suspects, and Patients With Glaucoma. *Am J Ophthalmol*. 2019;208:251–264.

7. Thulasidas M, Patyal S. Comparison of 24-2 Faster, Fast, and Standard Programs of Swedish Interactive Threshold Algorithm of Humphrey Field Analyzer for Perimetry in Patients With Manifest and Suspect Glaucoma. *J Glaucoma*. 2020;29(11):1070–1076.

8. Phu J, Kalloniatis M. Viability of Performing Multiple 24-2 Visual Field Examinations at the Same Clinical Visit: The Frontloading Fields Study (FFS). *Am J Ophthalmol*. 2021;230:48–59.

9. Phu J, Kalloniatis M. Patient and technician perspectives following the introduction of frontloaded visual field testing in glaucoma assessment. *Clin Exp Optom*. 2021;17:1–7, doi:10.1080/08164622.2021.1965461. Online ahead of print.

10. Phu J, Kalloniatis M. A Strategy for Seeding Point Error Assessment for Retesting (SPEAR) in Perimetry Applied to Normal Subjects, Glaucoma Suspects, and Patients With Glaucoma. *Am J Ophthalmol*. 2021;221:115–130.

11. Heijl A, Patella VM, Flanagan JG, et al. False Positive Responses in Standard Automated Perimetry. *Am J Ophthalmol*. 2021;233:180–188.

12. Wang H, Kalloniatis M. Clinical outcomes of the Centre for Eye Health: an intra-professional optometry-led collaborative eye care clinic in Australia. *Clin Exp Optom*. 2021;104:795–804.

13. Jamous KF, Kalloniatis M, Hennessy MP, et al. Clinical model assisting with the collaborative care of glaucoma patients and suspects. *Clin Exp Ophthalmol*. 2015;43(4):308–319.

14. Prum BE, Jr., Rosenberg LF, Gedde SJ, et al. Primary Open-Angle Glaucoma Preferred Practice Pattern((R)) Guidelines. *Ophthalmology*. 2016;123(1):P41–P111.

15. Foster PJ, Buhrmann R, Quigley HA, Johnson GJ. The definition and classification of glaucoma in prevalence surveys. *Br J Ophthalmol*. 2002;86(2):238–242.

16. Olsson J, Bengtsson B, Heijl A, Rootzen H. An improved method to estimate frequency of false positive answers in computerized perimetry. *Acta Ophthalmol Scand*. 1997;75(2):181–183.

17. Keltner JL, Johnson CA, Quigg JM, et al. Confirmation of visual field abnormalities in the Ocular Hypertension Treatment Study. Ocular Hypertension Treatment Study Group. *Arch Ophthalmol*. 2000;118(9):1187–1194.

18. Gardiner SK, Mansberger SL. Effect of Restricting Perimetry Testing Algorithms to Reliable Sensitivities on Test-Retest Variability. *Invest Ophthalmol Vis Sci*. 2016;57(13):5631–5636.

19. Gardiner SK, Swanson WH, Goren D, et al. Assessment of the reliability of standard automated perimetry in regions of glaucomatous damage. *Ophthalmology*. 2014;121(7):1359–1369.

20. Camp AS, Long CP, Patella VM, et al. Standard reliability and gaze tracking metrics in glaucoma and glaucoma suspects. *Am J Ophthalmol*. 2021;234:91–98.

21. Heijl A, Patella VM, Bengtsson B. *5. Statpac Analysis of Single Fields. Excellent Perimetry - The Field Analyzer Primer*. Dublin, CA: Carl Zeiss Meditec, 2021.

22. Phu J, Kalloniatis M. The Frontloading Fields Study (FFS): Detecting Changes in Mean Deviation in Glaucoma Using Multiple Visual Field Tests Per Clinical Visit. *Transl Vis Sci Technol*. 2021;10(13):21.

23. Mills RP, Budenz DL, Lee PP, et al. Categorizing the stage of glaucoma from pre-diagnosis to end-stage disease. *Am J Ophthalmol*. 2006;141(1):24–30.

24. Choi AYJ, Nivison-Smith L, Phu J, et al. Contrast sensitivity isocontours of the central visual field. *Sci Rep*. 2019;9(1):11603.

25. Phu J, Khuu SK, Bui BV, Kalloniatis M. Application of Pattern Recognition Analysis to Optimize Hemifield Asymmetry Patterns for Early Detection of Glaucoma. *Transl Vis Sci Technol*. 2018;7(5):3.

26. Phu J, Khuu SK, Nivison-Smith L, et al. Pattern Recognition Analysis Reveals Unique Contrast Sensitivity Isocontours Using Static Perimetry Thresholds Across the Visual Field. *Invest Ophthalmol Vis Sci*. 2017;58(11):4863–4876.

27. Phu J, Kalloniatis M. Comparison of 10-2 and 24-2C Test Grids for Identifying Central Visual Field Defects in Glaucoma and Suspect Patients. *Ophthalmology*. 2021;128(10):1405–1416.

28. Hood DC, Kardon RH. A framework for comparing structural and functional measures of glaucomatous damage. *Prog Retin Eye Res*. 2007;26(6):688–710.

29. Phu J, Kalloniatis M. Ability of 24-2C and 24-2 Grids to Identify Central Visual Field Defects and Structure-Function Concordance in Glaucoma and Suspects. *Am J Ophthalmol*. 2020;219:317–331.

30. Gardiner SK, Demirel S, Johnson CA. Is there evidence for continued learning over multiple years in perimetry? *Optom Vis Sci*. 2008;85(11):1043–1048.

31. Wild JM, Dengler-Harles M, Searle AE, et al. The influence of the learning effect on automated perimetry in patients with suspected glaucoma. *Acta Ophthalmol (Copenh)*. 1989;67(5):537–545.

32. Wild JM, Searle AE, Dengler-Harles M, O'Neill EC. Long-term follow-up of baseline learning and fatigue effects in the automated perimetry of

glaucoma and ocular hypertensive patients. *Acta Ophthalmol (Copenh)*. 1991;69(2):210–216.

33. Yohannan J, Wang J, Brown J, et al. Evidence-based Criteria for Assessment of Visual Field Reliability. *Ophthalmology*. 2017;124(11):1612–1620.

34. Phu J, Kalloniatis M, Khuu SK. Reducing Spatial Uncertainty Through Attentional Cueing Improves Contrast Sensitivity in Regions of the Visual Field With Glaucomatous Defects. *Transl Vis Sci Technol*. 2018;7(2):8.

35. Phu J, Kalloniatis M, Khuu SK. The Effect of Attentional Cueing and Spatial Uncertainty in Visual Field Testing. *PLoS One*. 2016;11(3):e0150922.

36. Blumenthal EZ, Sapir-Pichhadze R. Misleading statistical calculations in far-advanced glaucomatous visual field loss. *Ophthalmology*. 2003;110(1):196–200.