Published in final edited form as:

Ophthalmology. 2017 November ; 124(11): 1612–1620. doi:10.1016/j.ophtha.2017.04.035.

Evidence-based Criteria for Assessment of Visual Field Reliability

Jithin Yohannan, MD, MPH, Jiangxia Wang, MS, MA, Jamie Brown, MD, Balwantray C. Chauhan, PhD, Michael V. Boland, MD, PhD, David S. Friedman, MD, MPH, PhD, and Pradeep Y. Ramulu, MD, MHS, PhD

Glaucoma Center of Excellence, Wilmer Eye Institute, Johns Hopkins University School of Medicine, Baltimore, MD, USA. Dana Center for Preventive Ophthalmology, Wilmer Eye Institute, Johns Hopkins University, Baltimore, MD, USA

Abstract

Objective—Assess the impact of false positives (FPs), false negatives (FNs), fixation losses (FLs) and test duration (TD) on visual field (VF) reliability at different stages of glaucoma severity.

Participants—10,262 VFs from 1,538 eyes of 909 subjects with suspect or manifest glaucoma and 5 VF examinations.

Design—Retrospective.

Methods—Predicted mean deviation (MD) was calculated with multilevel modeling of longitudinal data (>5 non-initial VFs). Differences between predicted and observed MD (MD) were calculated as a reliability measure. The impact of FPs, FN, FLs and TD on MD was assessed using multi-level modelling.

Main outcome measure— MD associated with a 10% increment in FPs, FN, and FLs, or a 1-minute change in test duration.

Results—FLs had little impact on MD (<0.2 dB per 10% abnormal catch trials) and no level of FL produced 1 dB of MD at any disease stage. FPs yielded greater than expected MD, with a 10% increment in abnormal catch trials associated with a MD=0.42, 0.73, and 0.66 dB in mild (MD>−6 dB), moderate (−6 MD<−12 dB), and severe (−12 MD −20 dB) disease, respectively, up to 20% abnormal catch trials, and a MD=1.57, 2.06, and 3.53 dB beyond 20% abnormal catch trials.

Corresponding Author: Pradeep Y. Ramulu, MD, MHS, PhD, Wilmer Eye Institute, 600 N. Wolfe Street, Baltimore, MD 21287, pramulu@jhmi.edu.

Meeting Presentations: Association for Research in Vision and Ophthalmology Annual Meeting 2013 Paper Presentation, American Glaucoma Society Annual Meeting Poster Presentation 2014, American Academy of Ophthalmology Poster Presentation 2016

Conflict of Interest: No conflict of interest exists for any author

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

FNs generally produced observed MDs below expected MDs. FNs were minimally impactful up to 20% abnormal catch trials (MD per 10% increment >-0.14 dB at all levels of severity). Beyond 20% abnormal catch trials, each 10% increment in abnormal FN catch trials was associated with a MD=-1.27, -0.53, and -0.51 dB in mild, moderate and severe disease. | MD| 1dB occurred with 22% FPs and 26% FNs in early, 14% FPs and 34% FNs in moderate, and 16% FPs and 51% FNs in severe disease. A 1-minute increment in TD produced a MD between -0.35 and -0.40 depending on disease severity.

Conclusion—FLs have little impact on test reliability in established glaucoma patients. FPs, and to a lesser extent FNs and test duration, impact reliability significantly. The impact of FP and FN varies with disease severity and over the range of abnormal catch trials. Based on our findings, we present evidence-based severity-specific standards for classifying VF reliability are presented for clinical or research applications.

Introduction

A number of seminal glaucoma studies including the Ocular Hypertension Treatment Study (OHTS)¹, Collaborative Initial Glaucoma Treatment Study², Early Manifest Glaucoma Trial³ and Advanced Glaucoma Intervention Study⁴ have used automated visual field (VF) testing to assess presence of glaucoma, gauge disease severity, and track progression. As such, automated visual field testing remains the primary tool that glaucoma practitioners use to assess glaucoma-related visual damage, monitor progression⁵⁻⁷ and determine the impact of glaucoma on patient functionality.⁸⁻¹¹

When performing VF testing, it is important to know whether the VF test was completed properly by the patient in order to determine how the results should be used to guide care. Classically, this question has been answered by using reliability measures based on the percentage of abnormal catch trials in metrics such as fixation losses (FL), false positives (FP) and false negatives (FN) to classify a VF as unreliable (untrustworthy and needing repetition) or reliable (usable for clinical decision making). The Humphrey Field Analyzer (HFA) software, for example, uses a cutoff of 33% FP or FN and a cutoff of 20% for FL to define a field as unreliable.¹² Initially, such cutoffs were validated by testing normal subjects, ocular hypertensives and early glaucoma patients to determine the percentage of subjects who met or exceeded the cutoff levels. This early work showed that less than 0.5% of patients and normal subjects exceeded the 33% cutoff for FP and FN but a large number (19–35%) exceeded the cutoff of 20% or more FL,^{13,14} and the suggestion was made to increase the cutoff for FL to 33%.¹⁴ Based on these early data, VFs with FP, FN or FL above 33% were not included in OHTS¹ and these same cutoff values gained acceptance as a means to judge whether a visual field exam was unreliable. However, there are several limitations to this approach. First, the initial development of such cutoffs was based solely on how many patients exceeded the cutoff rather than an assessment of whether the fields exceeding these cutoffs values were unlikely to represent the true degree of VF loss. Second, cutoffs create a binary categorization of VF results (reliable or unreliable) which precludes consideration of *how unreliable* (i.e. how disparate from the true, unknown level of VF damage at the time of the exam) a VF is likely to be based on test parameters (i.e. FP, FN, FL).

In order to more meaningfully assess the impact of FP, FN and FL on VF reliability, quantitative measures capturing the degree of VF reliability are needed to allow clinicians to make decisions based on the degree of error likely to be present in their patients' test results. Previously, Bengtsson studied quantitative reliability in patients who performed VF tests twice in one week and found that the standard reliability indices were not significantly associated with threshold reproducibility, though severity of field loss was.¹⁵ Junoy Montolio et. al¹⁴ employed a different method to measure quantitative reliability, first predicting what a particular VFs MD should be using modeling and then calculating the difference between this predicted MD and the actual MD of the VF test to define a measure of reliability known as MD. They then calculated the effect of FP, FN and FL on MD and found that FP had the largest impact on MD with a 10% increment in FP associated with a 1.5 dB higher MD. FN and FL, when compared to FP, had less dramatic effects on MD. Although this study moved us toward a quantitative assessment of visual field reliability, it had several limitations including: (1) lack of a complete investigation of the relative impact of different measures at different stages of disease severity, (2) equal weighting of each additional 10% FPs, FNs, or FLs (thus suggesting that going between 0 and 10% FPs and between 30 and 40% FPs has an equivalent impact on MD) and (3) a limited sample of 160 patients.

In this study, we build upon the work of Junoy Montolio¹⁴ and use a large VF database to determine the quantitative impact of FPs, FNs, FLs, and test duration on VF reliability as defined by MD, defined as the difference between observed MD values and those predicted by our regression models. In addition, we attempt to assess how disease severity and the range of abnormal catch trials where additional FPs, FNs and FLs occur, impact the relationship between these indices and VF reliability. Finally, we use our data to propose evidence-based criteria to quantitatively judge VF reliability in clinical practice and research settings.

Methods

The study protocol was approved by the Johns Hopkins institutional review board, and adhered to the tenets of the Declaration of Helsinki. A waiver of consent was obtained to review VF data as well as to obtain information via chart review

STUDY PARTICIPANTS

Patients age 18 and over evaluated at The Wilmer Eye Institute Glaucoma Center of Excellence between 2002 and 2012 were eligible to be included in the analysis if they had a glaucoma-related diagnosis (glaucoma suspect or any other form of glaucoma). Eyes for which 5 or more VFs were obtained with a HFA II (Carl Zeiss Medical Technologies Inc., Dublin, CA, USA) and the 24-2 SITA protocol were analyzed. Patients could have one or both eyes included in analyses. As the current study was designed to evaluate the impact of abnormal catch trials and other VF metrics on visual field reliability, no eyes or VFs were excluded because of poor reliability. Only VFs with an MD>-20 were included in the analysis.

VF data were retrieved for eyes meeting the inclusion criteria above. Mean deviation (MD) was extracted as the measure of disease severity. Visual field metrics potentially affecting measured MD were also extracted including test duration and the percentage of abnormal FL, FP, and FN catch trials. Finally, the date and time of each VF test were obtained. Time of day was categorized as early morning (7–10 am), late morning (10 am – noon), early afternoon (12 noon – 2 PM), or late afternoon (2–5 pm). The date of VF testing was used to determine the day of the week as well as the season of the year (spring, summer, fall, winter).

Patient age was determined directly from VF output. A chart review was performed to determine patient sex, race and the additional variables presented in Table 1.

MODELING OF RELIABILITY

Reliability of MD was taken as the difference between observed and predicted MD values ($MD_{\text{observed}} - MD_{\text{predicted}}$, referred to as ΔMD), and was derived in a three-step process as summarized in Figure 1. First, predicted MDs were calculated for eligible VFs tests in the database with linear mixed effects regression models. The dependent variable in this model was the MD for each eligible VF test in the database, while the independent variables included time and the features described in Table 1. Since the baseline disease condition categories generated from the first VF MD and the eye specific average VF MD were used as covariates, the first VFs were not excluded from the sample used in the regression model. A linear mixed effects regression model approach was employed to account for clustering between eyes within the same patient and VFs done on the same eye. The model employed random intercepts, random slopes and an unstructured variance-covariance matrix. Second,

ΔMD was calculated as a continuous, directional measure of reliability for each VF test included in the study by subtracting the predicted MD obtained from the mixed effects model above from the actual observed MD for that VF test ($MD_{\text{observed}} - MD_{\text{predicted}}$). Third, predictors of reliability (ΔMD) were identified with a multi-level linear mixed effects model employing random intercepts but not random slopes, as ΔMD was not expected to vary over time. In this final multivariate model, the dependent variable was ΔMD (representing reliability) and predictors that were used to explain ΔMD included FLs, FPs, FNs, test duration, time of day, day of week and season. Interaction terms between the severity of visual field loss and FLs, FPs, FNs, test durations were used to account for the fact that the effects of FLs, FPs, FNs, test durations on ΔMD vary by the severity of visual field loss. The estimated regression coefficients represent the effect of each factor on ΔMD assuming all other factors are held constant.

Derived regression coefficients were used to define the test duration or percentage of abnormal FL, FP, and FN catch trials required to produce various degrees of unreliability at different stages of disease severity. Acceptable levels of ΔMD were defined as: (1) an absolute level (i.e. >1 dB or <-1 dB), or (2) a level defined relative to the typical range of

MD values found at that stage of disease. To accomplish the latter, the distribution of within-eye standard deviations was first derived for ΔMD ($SD_{\Delta MD}$) at each disease stage (defined as mild, moderate, or severe based on an average MD of >-6 dB, -6 dB $\Delta MD < 12$ dB, or -12 dB $\Delta MD \geq -20$ dB, respectively, over the course of follow-up), and stage-

specific median SD_{MD} values were defined. Standards for acceptable MD related to the stage-specific SD_{MD} values were then derived, i.e., requiring MD to be less than the median SD_{MD} at that level of disease severity. All statistical analyses were performed in STATA (version 12.0, College Station, TX, USA).

Results

1,538 eyes of 909 subjects had VF data that met the inclusion criteria and were included in the study. The study population had a mean age (SD) of 64.9 (11.9) years at the time of the first VF. There were 497 (54.7%) male and 594 (65.4%) white subjects (Table 2). Subjects completed a total of 10,262 VF tests (mean (SD), 6.7 (1.8) tests per eye) over the study period. The majority of these VFs (72.7%) had mild field loss of <6 dB in the first test (Fig. 2). With regard to the conventional catch trials, 2,975 (29%), 261 (2.5%), and 262 (2.5%) of VF tests had > 20% FLs, FPs and FNs, respectively in the first VF test (Fig. 3). Similar percentages of visual fields demonstrated >20% FLs, FPs and FNs when only the last VF test that for each eye was considered.

In the full set of VFs, MD increased with a greater percentage of FPs, decreased with greater numbers of FNs, and was relatively unchanged across the observed range of FLs (Fig. 4). The impact of various VF parameters (FLs, FPs, FNs and test duration) on MD was further assessed quantitatively in multivariate regression models. Models were run to stratify by the severity of visual field loss. Given the observed appearance of a greater impact of additional FPs and FNs on MD at higher percentages of abnormal catch trials (Fig. 4), spline terms were incorporated to separately to define the effect of additional FLs, FPs, and FNs over the range of 0–20% abnormal catch trials, and beyond 20% abnormal catch trials (Table 3). We chose to use a spline term at 20% abnormal catch trials because the impact of FPs and FNs appeared to accelerate with higher percentages of abnormal catch trials in Figure 4, and because the number of FPs and false negatives become sparse beyond 20% (Figures 3b and 3c).

FLs had little to no impact on MD regardless of disease severity or the percentage of abnormal catch trials (less than 0.2 dB for 10% more FLs for all conditions, Table 3a). FPs had a significantly increased MD across all levels of disease severity. Moreover, the increase in MD attributable to additional FPs was greater beyond 20% abnormal catch trials (1.57 to 3.53 dB/10% depending on level of MD severity) as compared to additional FPs occurring over the range of 0–20% abnormal catch trials (0.42 to 0.73 dB/10% depending on level of MD severity) (Table 3b). Additionally, FPs occurring in eyes with severe visual field loss had a greater impact on MD than FPs occurring in eyes with mild visual loss (0.66 vs. 0.42 dB/10% over the range of 0–20% abnormal catch trials; 3.53 vs. 1.57 dB/10% beyond 20% abnormal catch trials) (Table 3b). FNs were generally associated with lower MD, though the absolute impact was not as large as the impact of FPs in most cases (Table 3). Additional FNs had a small impact on MD over the range of 0–20% abnormal catch trials (–0.07 to 0.29 dB/10% depending on level of glaucoma severity), but a more significant impact on MD beyond 20% abnormal catch trials (–1.27 to –0.51 dB/10% depending on the level of glaucoma severity) (Table 3c). Finally, longer test duration was associated with a significantly lower (more negative) MD values. Specifically,

each extra minute of test time was associated with a MD between a -0.40 to -0.35 dB, depending on the stage of disease (Table 3d).

The extent to which degree each parameter was required to produce ± 1 dB MD at various stages of visual field loss is shown in Table 4a. The impact of FLs on visual field reliability was so low such that no degree of fixation loss could produce a ± 1 dB MD. However, as few as a 14% or 16% FPs were enough to produce a +1 dB MD in eyes with moderate or severe visual loss, respectively. As compared to FPs, a higher rate of FNs was required to produce a -1 dB MD, and more FNs were required to produce the same -1 dB MD in eyes with more severe visual field loss as compared to eyes with mild visual field loss. While test duration was important, it has to be increased significantly (by 2.5 to 2.9 minutes) to produce a change in MD value of -1 dB.

An alternate approach to reliability is to allow a degree of MD defined by the normal variability of VF testing at that level of disease severity. To explore the normal degree of variability observed in VFs taken from eyes of a given level of severity, the within-eye standard deviation values for MD (SD_{MD}) were first examined (Table 5). Median SD_{MD} was greater in eyes with moderate and severe visual field loss (median of 1.53 and 1.38 dB respectively) than in eyes for mild field loss (median of 0.89 dB). In Table 4b, the degree to which each parameter was required to produce a MD value equivalent median SD_{MD} is shown. Again, no level of FLs produced a MD value equivalent to the corresponding SD_{MD} . At each level of visual field loss severity 21% FP was required to produce a MD value equal to median SD_{MD} . Only 26% FN was needed to produce MD equal to SD_{MD} in mild disease but this increased to 44 and 59% FNs in moderate and severe disease respectively. Test duration must increase by between 2.3 and 4.4 minutes to produce a MD equivalent to SD_{MD} . The appendix for Table 5 demonstrates how to calculate the percentage of FL, FP and FN for any desired value of MD. Test timing defined as time of day, day of the week and season had mild associations with MD (Table 6). While some values reached statistical significance, the absolute impact of these timing variables on visual field performance was minimal, and all predictors had an absolute impact of less than 0.20 dB on MD.

Discussion

In this study, we showed that in established VF test takers, FPs, FNs and test duration are all clinically significant predictors of VF reliability (defined as MD), with FPs producing the largest impact on MD and FLs producing almost no impact on MD. The impact of many traditional VF metrics (particularly FPs and FNs) on reliability varies by glaucoma disease severity, and the impact of additional FPs and FNs is greater if these additional FPs and FNs occur beyond the threshold of 20% abnormal catch trials, as opposed to over the range of 0–20% abnormal catch trials. For each VF metric, clinicians should consider stage of glaucoma severity, the degree of abnormality (i.e. percentage of abnormal catch trials), and the acceptable amount of unreliability when deciding whether to trust a VF result. Evidence-based standards for VF reliability derived from study data are proposed for use in future research studies.

FPs have the greatest impact on VF reliability amongst all VF metrics, and this impact is most prominent in eyes with moderate or severe disease, which likely have a greater potential for an artificially higher MD level. As false positives are produced when patients report a stimulus when there is none, they are expected to increase MD (positive MD) and this is consistent with our results. Indeed, just 14% FPs are expected to produce an observed MD 1 dB above predicted MD in moderate and severe glaucoma as compared to 21% FPs required to produce the same change in mild disease. Our work agrees with prior work by Junoy Montolio et. al¹⁶ and Lee et. al¹⁷ who used similar methods to find observed MDs above those predicted by regression models in VFs with high FPs. Junoy Montolio found that 10% more FPs was associated with an MD 1 dB greater than the expected true value, a value higher than that observed here. Differences in the findings from the two studies may reflect the use of a 30–2 pattern in the previous study versus the 24-2 pattern in the current study. Lee et. al demonstrated that 10% more FPs were associated with a 1.5 dB higher mean sensitivity (MS) using the Octopus 201 Perimeter with the G1 testing algorithm. The Octopus perimeter has a different test pattern and different strategies for computing catch trials than used in this study, suggesting that the association between VF reliability and degree of FPs may have other contributing factors. Finally, it should be noted that these prior studies considered the impact of all FPs without regard to disease severity, though work by Bengtsson¹⁵ et. al. also showed that disease severity is an important modifier of the impact of reliability indices (in agreement with the findings of our study). Additionally, unlike prior studies, the current work demonstrates that the impact of additional FPs varies across the observed range of abnormal catch trials, with additional FPs beyond 20% having a substantially greater impact than additional FPs below the 20% threshold. In fact, the MD associated with an additional 10% more abnormal FP catch trials occurring beyond the threshold of 20% abnormal catch trials nearly matches or exceeds the median within-eye standard deviation of MD for every visual field loss severity category (Table 4). Therefore, clinicians should interpret a VF result with caution when extreme values of FPs (>20%) are obtained.

FNs had a smaller impact on VF reliability than FPs, with a greater percentage of FNs generally associated with an observed MD more negative than that predicted by our models. These results are generally in agreement with prior work by Junoy Montolio¹⁶, who found a 0.1 dB increase in MD for 10% change in false negative in severe disease and a 0.5 dB decrease in mild disease using a 30–2 test, and Lee et. al¹⁷, who found a 10% change in FN resulting in a 1.2 dB decrease in mean sensitivity. Direct comparison between our study and this prior work is not possible, however, given that Lee et. al. did not stratify their results by disease severity, and neither study assessed the impact of additional FNs based on the range of abnormality where the FNs occurred. Indeed, additional FNs occurring beyond the 20% threshold generated a substantially greater degree of unreliability than additional FNs occurring up to the 20% threshold (−1.27 vs. −0.07 dB/10% FNs in mild disease). Therefore, one should be very cautious in interpreting a VF result when extreme values of FNs are obtained. We also identified that FNs had the greatest impact on reliability in mild disease, possibly reflecting that normal or near-normal fields had the greatest potential to overestimate true test results, or that FNs may in fact be expected in later disease where diseased locations may have variable responses¹⁸.

Increased test duration had a moderate impact on VF reliability, and to our knowledge this is the first study to quantify the impact of test duration on MD. Specifically, a one-minute increase in test duration resulted in an observed MD 0.35 to 0.40 dB lower than the modeled true value. Newkirk et. al¹⁹ demonstrated errors increased as test duration was artificially increased with SITA strategies, which could account for our findings. However, the authors did not assess the direct, real world impact that test duration had on VF reliability. In this analysis, the impact of test duration on reliability was not substantial, with 2–3 minutes of additional time required to produce an additional 1 dB of error and up to 4.4 minutes of test time needed to produce a 1 median within-eye standard deviation change in MD. Thus, even VFs with mild loss would need to be at least 8 minutes long to be associated with this degree of error due to test duration alone, and such long testing times are rarely encountered in clinical practice at this stage of disease. The reason that test duration is associated with a larger (more negative) MD is unclear as a longer test can be both a cause of an unreliable tests (i.e., test with high FP and FNs tend to take longer) or a consequence of an unreliable test (FP and FN are measured more times in a longer test). However, we did not seek to find a causal relationship between test duration and catch trials. Additionally, we performed testing for multicollinearity between test duration and the catch trials and found there was no effect significant enough to invalidate the models used in this study.

FLs had little impact on visual field reliability in this study, and this finding has been mirrored in other studies¹⁶. FLs can often occur when the blind spot is not in the correct position either because of mismapping of the blind spot, small head tilts, or other changes in patient position during the test. Our results do not support the use of FL as a clinically meaningful indicator for VF reliability in established patients, though it may be important in in new VF takers (who may be searching for stimuli. Gaze tracking data were not available for our subjects, and it therefore remains possible that alternate methods for evaluating fixation instability may better capture the impact of poor fixation on VF reliability. Finally, test timing, including time of day, day of week and season had minimal impact on VF performance, and these results are also in agreement with prior work.¹⁶

Determining VF reliability is an important aspect of caring for glaucoma patients, and our findings provide practical guidelines for quantifying the likely error in VF testing based on test duration and the percentage of abnormal catch trials. We propose the following evidence-based criteria as a guide to determine VF reliability:

1. FLs do not impact VF reliability meaningfully in non-naive field taker
2. Any level of FP decreases visual field reliability, but one should be especially cautious when FP occur in advanced disease or are found in >20% of catch trials.
3. FN have less of an impact than FP. FN do not significantly impact reliability in advanced disease unless they are very high (at least 35% but oftentimes more than this value); In early disease, FNs are more impactful where as few as 25% FN can have a significant impact.
4. Large increases in test duration (more than 2 to 3 minutes beyond what is typical) is an indicator of poor reliability and should be taken into consideration.

While we strongly recommend considering reliability along a continuum, there may also be times when explicit cutoffs may be useful for deeming a VF to be reliable or unreliable. The degree of likely error acceptable for such cutoff differs depending on the reason for why the VF is obtained and used. For example, if VFs are being used to assess the stage of disease, then 1 (or even 2) dB of error is unlikely to result in substantial misclassification. Alternately, if VFs are being used to judge worsening of disease, a small MD can have a significant impact. To this end, we have given the reader tools such as the standard deviation of MD (Table 5) as well as formulas for calculating custom cutoffs for VF metrics (Appendix) to tailor these data to a specific application. Here, we propose using 1 dB of error as a general cutoff when trying to classify disease severity (Table 4a), and the median SD of MD when determining if error is significant compared to the variability typical at that stage of disease (Table 4b).

VF testing is, of course, only one component of diagnosing glaucoma or judging glaucoma progression. Each VF test may be thought of as modifying the pre-test (pre -VF) probability that the patient has glaucoma, or that their glaucoma is progressing. The pre-VF probability is informed by all current and prior clinical data outside of the current VF test (optic nerve status, prior VF tests, IOP measurements, etc.). While a test with poor reliability could not be discounted completely, it would be less able to influence the pre-test probability, and thus have a lower ability to help the clinician gain certainty regarding the diagnosis of glaucoma, or whether meaningful glaucoma progression was occurring.²⁰

Our study was limited by the fact that predicted VF MD, which was used to calculate MD, is a prediction based on modeling as there is no way of knowing how accurately these linear models tell us the “true” level of VF loss at any point in time. Even though the prediction model incorporated the eye specific average VF MD as a proxy for the true condition of the eye, and an eye level random slope over time to improve prediction, the study results could still be affected by the fact that the use of Empirical Bayes algorithms give more weights to the average sample profile when deriving MD. Additionally, the choice to place a spline term at 20% of catch trials in our final model may influence the model results. We felt 20% this was the most conservative approach as moving the spline term to a lower value (i.e. 10%) would pair lower unreliability values with higher values and increase the impact of early unreliability on MD. Alternatively, moving the spline term higher (i.e. 30%) would result in very few data beyond the spline point. In sensitivity analysis (data not shown), adding a second spline term at 10% did not meaningfully impact the results of the final model. Despite the limitations, we feel that our modeling approach is based on rigorous statistical methodology and the most appropriate approach to apply to a large visual field data base. Bengtsson’s approach of repeated measures (having patients repeat VF within one week) to measure reliability is certainly a viable alternative methodology, though it is impractical to apply to this methodology (and therefore the findings) to a large sample of clinical patients.¹⁵ Furthermore, our study was limited to automated VF testing performed on the HFA using the SITA Standard strategy and 24-2 pattern. Similar studies that used either the HFA with the 30-2 pattern¹⁶ or the Octopus perimeter¹⁷ had results that were similar but not identical to ours. Thresholding algorithms may also alter results given that prior work showed significantly fewer FNs²¹ and FPs²² in tests incorporating SITA as compared to tests utilizing a full threshold algorithm. This suggests that although our results

may be generalizable to different machines and testing algorithms, further research must be done to determine the true quantitative impact of VF metrics in different test scenarios. Furthermore, although we only VF metrics calculated by the HFA, some studies have suggested that technician notes and experience may be an indicator as well as predictor of reliability.³

In summary, our findings and evidence-based guidelines are highly relevant to glaucoma clinicians and researchers who wish to determine whether, and to what extent, they can trust an automated VF test. Future research in this arena should focus on applying this method to different visual field patterns and strategies. In addition, efforts should be made incorporate the proposed quantitative measures (i.e. MD) into visual field machine output so that clinicians have easy access to this additional level of data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding: NIH R01 EY022976, Research to Prevent Blindness, Doris Duke Foundation, Canadian Institutes of Health Research (MOP-11359)

Bibliography

1. Kass MA, Heuer DK, Higginbotham EJ, et al. The Ocular Hypertension Treatment Study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma. *Arch Ophthalmol Chic Ill 1960*. 2002; 120:701, 713–830.
2. Musch DC, Lichter PR, Guire KE, Standardi CL. The Collaborative Initial Glaucoma Treatment Study: study design, methods, and baseline characteristics of enrolled patients. *Ophthalmology*. 1999; 106:653–662. [PubMed: 10201583]
3. Heijl A, Leske MC, Bengtsson B, et al. Reduction of intraocular pressure and glaucoma progression: results from the Early Manifest Glaucoma Trial. *Arch Ophthalmol Chic Ill 1960*. 2002; 120:1268–1279.
4. The AGIS Investigators. The Advanced Glaucoma Intervention Study (AGIS): 7. The relationship between control of intraocular pressure and visual field deterioration. The AGIS Investigators. *Am J Ophthalmol*. 2000; 130:429–440. [PubMed: 11024415]
5. Chauhan BC, Malik R, Shuba LM, et al. Rates of glaucomatous visual field change in a large clinical population. *Invest Ophthalmol Vis Sci*. 2014; 55:4135–4143. [PubMed: 24917147]
6. Artes PH, O'Leary N, Nicolela MT, et al. Visual field progression in glaucoma: what is the specificity of the Guided Progression Analysis? *Ophthalmology*. 2014; 121:2023–2027. [PubMed: 24878173]
7. Heijl A, Bengtsson B, Chauhan BC, et al. A comparison of visual field progression criteria of 3 major glaucoma trials in early manifest glaucoma trial patients. *Ophthalmology*. 2008; 115:1557–1565. [PubMed: 18378317]
8. Arora KS, Boland MV, Friedman DS, et al. The relationship between better-eye and integrated visual field mean deviation and visual disability. *Ophthalmology*. 2013; 120:2476–2484. [PubMed: 23993358]
9. Ramulu PY, Hochberg C, Maul EA, et al. Glaucomatous visual field loss associated with less travel from home. *Optom Vis Sci Off Publ Am Acad Optom*. 2014; 91:187–193.
10. Curriero FC, Pinchoff J, van Landingham SW, et al. Alteration of travel patterns with vision loss from glaucoma and macular degeneration. *JAMA Ophthalmol*. 2013; 131:1420–1426. [PubMed: 24030033]

11. van Landingham SW, Willis JR, Vitale S, Ramulu PY. Visual field loss and accelerometer-measured physical activity in the United States. *Ophthalmology*. 2012; 119:2486–2492. [PubMed: 22892152]
12. Heijl, A., Patella, VM., Bengtsson, B. *Effective Perimetry: The Field Analyzer Primer*. 4. Dublin, Calif: Carl Zeiss Meditec, Inc; 2012.
13. Nelson-Quigg JM, Twelker JD, Johnson CA. Response properties of normal observers and patients during automated perimetry. *Arch Ophthalmol Chic Ill* 1960. 1989; 107:1612–1615.
14. Bickler-Bluth M, Trick GL, Kolker AE, Cooper DG. Assessing the utility of reliability indices for automated visual fields. Testing ocular hypertensives. *Ophthalmology*. 1989; 96:616–619. [PubMed: 2748118]
15. Bengtsson B. Reliability of computerized perimetric threshold tests as assessed by reliability indices and threshold reproducibility in patients with suspect and manifest glaucoma. *Acta Ophthalmol Scand*. 2000; 78:519–522. [PubMed: 11037906]
16. Junoy Montolio FG, Wesselink C, Gordijn M, Jansonius NM. Factors that influence standard automated perimetry test results in glaucoma: test reliability, technician experience, time of day, and season. *Invest Ophthalmol Vis Sci*. 2012; 53:7010–7017. [PubMed: 22952121]
17. Lee M, Zulauf M, Caprioli J. The influence of patient reliability on visual field outcome. *Am J Ophthalmol*. 1994; 117:756–761. [PubMed: 8198159]
18. Gardiner SK, Swanson WH, Goren D, et al. Assessment of the reliability of standard automated perimetry in regions of glaucomatous damage. *Ophthalmology*. 2014; 121:1359–1369. [PubMed: 24629617]
19. Newkirk MR, Gardiner SK, Demirel S, Johnson CA. Assessment of false positives with the Humphrey Field Analyzer II perimeter with the SITA Algorithm. *Invest Ophthalmol Vis Sci*. 2006; 47:4632–4637. [PubMed: 17003461]
20. Fagan TJ. Letter: Nomogram for Bayes theorem. *N Engl J Med*. 1975; 293:257. [PubMed: 1143310]
21. Johnson CA, Sherman K, Doyle C, Wall M. A comparison of false-negative responses for full threshold and SITA standard perimetry in glaucoma patients and normal observers. *J Glaucoma*. 2014; 23:288–292. [PubMed: 23632399]
22. Wall M, Doyle CK, Brito CF, et al. A comparison of catch trial methods used in standard automated perimetry in glaucoma patients. *J Glaucoma*. 2008; 17:626–630. [PubMed: 19092457]
23. Johnson LN, Aminlari A, Sassani JW. Effect of intermittent versus continuous patient monitoring on reliability indices during automated perimetry. *Ophthalmology*. 1993; 100:76–84. [PubMed: 8433832]

Figure 1. Schematic demonstrating how ΔMD (Delta Mean Deviation) was calculated. Predicted MD at each time point was calculated using a regression model whose predictors included the average performance based visual field (VF) tests, time since visual field testing commenced, severity of visual field loss, intraocular pressure (IOP) and surgical history. MD was calculated by subtracting visual field MD measured at the given time point for the given patient from predicted MD at that time point.

Figure 2.
Histogram of mean deviation (MD) of the first field of each eye in the visual field database.

Figure 3.
Histogram of fixation losses false positives and false negatives in the visual field database

Figure 4. Effect of False positives, false negative and fixation losses on delta mean deviation (ΔMD).

Table 1

Variables used to model predicted Mean Deviation (MD)

Baseline severity of visual field (VF) loss judged by baseline MD in each eye
Interaction term between baseline MD and time since baseline VF in each [*]
Within-eye average MD from all prior VF (with maximum and minimum MD excluded)
Mean IOP at each visit in each eye
The interaction term between within eye maximum IOP and time since baseline VF in each eye [*]
Glaucoma diagnosis in each eye at baseline (No Glaucoma, POAG, POAS, PACG, PACS, PXG, PDG & Other Glaucoma)
The interaction term between baseline glaucoma diagnosis in each eye and time since baseline VF [*]
Number of glaucoma medications baseline and at the patient's final visit
Interaction term between number of glaucoma medications and time since baseline VF [*]
Presence of baseline incisional glaucoma surgery in each eye
Interaction term between presence baseline incisional glaucoma surgery in each eye and time since baseline VF [*]
Presence of interim incisional glaucoma surgery in each eye
Number of years since interim incisional glaucoma surgery in each eye
Baseline lens status in each eye
Presence of interim cataract surgery in each eye

^{*} Interaction terms including time were included for variables likely to be associated with more rapid rate of visual field progression.

Table 2

Demographic and Visual Field Characteristics of Patients Studied

Total Number of Subjects	909
Gender	
Male (%)	497 (54.7)
Female (%)	412 (45.3)
Race	
White (%)	594 (65.4)
Black (%)	232 (25.5)
Asian (%)	83 (9.1)
Age at first VF, mean (SD)	64.9 (11.8)
Total number of eyes	1538
Mean number of Fields per eye (SD)	6.7 (1.8)
Total Number of Fields	10,262
MD	
MD>-6 dB (%)	7,462 (72.7)
-6 dB > MD > -12 dB (%)	1,754 (17.1)
-12 dB > MD > -20 dB (%)	1,046 (10.2)
Fixation Loss	
0 to 10%	5,244 (51.1)
10 to 20%	2,027 (19.8)
>20%	2,975 (29.0)
False Positive	
0 to 10%	9,304 (90.7)
10 to 20%	696 (6.8)
>20%	262 (2.5)
False Negative	
0 to 10%	8,827 (86.0)
10 to 20%	1,174 (11.4)
>20%	261 (2.5)
Test duration > 8 minutes (%)	793 (7.7)

Table 3

Effects of visual field parameters on difference from predicted mean deviation based on severity of visual field loss.

Table 3a. Effect of Fixation Losses and Visual Field Loss on difference from predicted mean deviation.		
	Change in MD (Observed – Predicted) for 10% change in Fixation Losses	
Severity of Visual Field Loss	Fixation loss 0 to 20% (95% CI)	Fixation loss >20% (95% CI)
>–6 dB	0.07 dB (0.02 to 0.12) *	0.00 dB (–0.03 to 0.03)
–6 to –12 dB	0.08 dB (–0.03 to 0.18)	0.09 dB (0.04 to 0.15) *
–12 to –20 dB	0.18 dB (0.04 to 0.32) *	0.07 dB (–0.02 to 0.15)

Table 3b. Effect of False Positives and Visual Field Loss on difference from predicted mean deviation.		
	Change in MD (Observed – Predicted) for 10% change in False Positives	
Severity of Visual Field Loss	False Positives 0 to 20% (95% CI)	False Positives >20% (95% CI)
>–6 dB	0.42 dB (0.33 to 0.51) *	1.57 dB (1.37 to 1.77) *
–6 to –12 dB	0.73 dB (0.55 to 0.90) *	2.06 dB (1.85 to 2.26) *
–12 to –20 dB	0.66 dB (0.41 to 0.92) *	3.53 dB (3.08 to 3.97) *

Table 3c. Effect of False Negatives and Visual Field Loss on difference from predicted mean deviation.		
	Change in MD (Observed – Predicted) for 10% change in False Negatives	
Severity of Visual Field Loss	False Negatives 0 to 20% (95% CI)	False Negatives >20% (95% CI)
>–6 dB	–0.07 dB (–0.17 to 0.03)	–1.27 dB (–1.60 to –0.94) *
–6 to –12 dB	–0.14 dB (–0.27 to –0.01)	–0.53 dB (–0.77 to –0.29) *
–12 to –20 dB	0.29 dB (0.14 to 0.45) *	–0.51 dB (–0.76 to –0.25) *

Table 3d. Effect of Test Time and Visual Field Loss on Visual Field Reliability		
Severity of Visual Field Loss	Mean Test Duration in Minutes (SE)	Change in MD (Observed – Predicted) for 1 minute increase in Test Duration
>–6 dB	5.7 (0.03)	–0.40 dB (–0.44 to –0.36)
–6 to –12 dB	6.9 (0.05)	–0.35 dB (–0.39 to –0.32)
–12 to –20 dB	7.3 (0.06)	–0.38 dB (–0.42 to –0.34)

Results in tables 3a, 3b, 3c, 3d and 6 are from the linear mixed effects regression model including the predictors test duration, time of day, day of week and season, and the interaction terms between the severity of visual field loss and FLs, FPs, FNs and test duration.

* P-value<0.05

Table 4

Table 4a. Percentage of abnormal catch trials required to produce a 1 dB increase in absolute value of MD (Observed – Predicted) for each reliability parameter, stratified by degree of visual field loss					
Severity of Visual Field Loss	Fixation Loss	False Positives	False Negatives	Test Duration	
>–6 dB	No level	22%	26%	8.2'	
–6 to –12 dB	No level	14%	34%	9.8'	
–12 to –20 dB	No level	16%	51%	10.0'	

Table 4b. Percentage of abnormal catch trials required to produce the median within-eye standard deviation of MD (Observed – Predicted) for each reliability parameter, stratified by degree of visual field loss					
Severity of Visual Field Loss	Fixation Loss	False Positives	False Negatives	Test Duration	
>–6 dB	No level	21%	26%	8.0'	
–6 to –12 dB	No level	21%	44%	11.3'	
–12 to –20 dB	No level	21%	59%	11.0'	

Table 5

Within Eye Standard Deviation of Change in MD for Various Stages of Visual Field Loss

Severity of Visual Field Loss	Standard Deviation of MD			
	Mean	25 th Percentile	50 th Percentile (Median)	75 th Percentile
> -6 dB	1.14	0.60	0.89	1.36
-6 to -12 dB	1.94	0.96	1.53	2.13
-12 to -20 dB	1.82	0.96	1.38	2.37

Table 6

Effect of Diurnal, Weekly and Seasonal Variation on Visual Field Reliability

	Change in MD (95% CI)
Time of Day	
7 am to 10 am	REF
10 am to noon	−0.05 dB (−0.14 to 0.03)
Noon to 2 pm	−0.10 dB (−0.19 to −0.02)*
2 pm to 5pm	−0.07 dB (−0.16 to 0.02)
Day of Week	
Monday	REF
Tuesday	0.02 dB (−0.08 to 0.12)
Wednesday	−0.03 dB (−0.13 to 0.07)
Thursday	−0.05 dB (−0.16 to 0.06)
Friday	0.005 dB (−0.11 to 0.10)
Season	
Winter	REF
Spring	−0.09 dB (−0.17 to 0.00)*
Summer	−0.16 dB (−0.25 to −0.07)*
Fall	−0.16 dB (−0.25 to −0.08)*