tvst

Article

Improving Visual Field Forecasting by Correcting for the Effects of Poor Visual Field Reliability

Gabriel A. Villasana[1], Chris Bradley[2], Tobias Elze[3], Jonathan S. Myers[4], Louis Pasquale[5], C Gustavo De Moraes[6], Sarah Wellik[7], Michael V. Boland[3], Pradeep Ramulu[2], Greg Hager[1], Mathias Unberath[1], and Jithin Yohannan[1,2]

[1] Malone Center for Engineering in Healthcare, Johns Hopkins University, Baltimore, MD, USA
[2] Wilmer Eye Institute, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[3] Massachusetts Eye and Ear Infirmary, Harvard Medical School, Boston, MA, USA
[4] Wills Eye Hospital, Glaucoma Research Center, Philadelphia, PA, USA
[5] Icahn School of Medicine at Mount Sinai, New York, NY, USA
[6] Columbia University Irving Medical Center, New York, NY, USA
[7] Bascom Palmer Eye Institute, University of Miami Miller School of Medicine, Miami, FL, USA

Correspondence: Jithin Yohannan, Johns Hopkins Hospital, 600 N Wolfe St, Baltimore, MD 21287, USA.
e-mail: jithin@jhmi.edu

Received: July 22, 2021
Accepted: March 30, 2022
Published: May 26, 2022

Keywords: mean deviation (MD); visual field (VF); reliability

Citation: Villasana GA, Bradley C, Elze T, Myers JS, Pasquale L, De Moraes CG, Wellik S, Boland MV, Ramulu P, Hager G, Unberath M, Yohannan J. Improving visual field forecasting by correcting for the effects of poor visual field reliability. Transl Vis Sci Technol. 2022;11(5):27, https://doi.org/10.1167/tvst.11.5.27

Purpose: The purpose of this study was to accurately forecast future reliable visual field (VF) mean deviation (MD) values by correcting for poor reliability.

Methods: Four linear regression techniques (standard, unfiltered, corrected, and weighted) were fit to VF data from 5939 eyes with a final reliable VF. For each eye, all VFs, except the final one, were used to fit the models. Then, the difference between the final VF MD value and each model's estimate for the final VF MD value was used to calculate model error. We aggregated the error for each model across all eyes to compare model performance. The results were further broken down into eye-level reliability subgroups to track performance as reliability levels fluctuate.

Results: The standard method, used in the Humphrey Field Analyzer (HFA), was the worst performing model with an average residual that was 0.69 dB higher than the average from the unfiltered method, and 0.79 dB higher than that of the weighted and corrected methods. The weighted method was the best performing model, beating the standard model by as much as 1.75 dB in the 40% to 50% eye-level reliability subgroup. However, its average 95% prediction interval was relatively large at 7.67 dB.

Conclusions: Including all VFs in the trend estimation has more predictive power for future reliable VFs than excluding unreliable VFs. Correcting for VF reliability further improves model accuracy.

Translational Relevance: The VF correction methods described in this paper may allow clinicians to catch VF worsening at an earlier stage.

Introduction

Automated visual field (VF) tests play a key role in the diagnosis of glaucoma and assessment of disease worsening.[1–4] Whereas VF tests help clinicians monitor longitudinal changes in glaucoma disease trajectory, they are associated with known variability, making interpretation of changes difficult.[5] There is growing evidence that mean deviation (MD) values depend on reliability indices, such as false positive (FP) percentages, false negative (FN) percentages, and test duration. Worse levels of these reliability metrics make the determination of change over time more challenging.[6] By correcting for the effects of these reliability indices on algorithms used to determine change, it may be possible to more accurately identify disease worsening and improve patient care accordingly.

Our prior work has shown that increases in FPs, FNs, and test duration have effects on the MD value depending on the stage of glaucoma and the values of the reliability indices (Table 1).[7] In general, as

Table 1. Correction Chart for MD Values

	Effect of 1% Increase of False Positives on Mean Deviation, dB		Effect of 1% increase of false negatives on Mean Deviation, dB		Effect of 1 Minute Increment of Test Duration on Mean Deviation, dB
	0% $\leq$ *FP*% $\leq$ 20%	20% $\leq$ *FP*%	0% $\leq$ *FN*% $\leq$ 20%	20% $\leq$ *FN*%	
Mild/suspect	0.042	0.157	−0.007	−0.127	−0.400
Moderate	0.073	0.206	−0.014	−0.053	−0.350
Advanced	0.066	0.353	0.029	−0.051	−0.380

unreliability increases, the measured MD value deviates further from the true MD value. False positives have the largest effect on this MD error, followed by FNs and test duration. The Guided Progression Analysis (GPA) used in the Humphrey Field Analyzer (HFA) displays a linear regression for MD values over time that does not "correct" for poor reliability.[8] Rather, the GPA model – henceforth called the "standard model" – simply excludes "unreliable" VF MD values, defined as VFs with more than 20% fixation losses or 15% FPs, from the MD values over time regression in order to better predict future MD values from past reliable VFs.[9]

Here, we attempt to improve the accuracy of forecasting future MD values by applying three different techniques to linear regression: (1) including all available tests (i.e. not excluding data from "unreliable" VFs), (2) correcting the unreliable VFs MD values using results from our previous study,[7] and (3) weighting MD values by their reliability. We compare the performance of these three models – henceforth referred to as "unfiltered," "corrected," and "weighted," respectively – to the performance of the "standard" model (used in GPA) by predicting MD values of future reliable VFs. Because we wish to forecast MD values which are as close to a true MD value as possible and in order to compare our results to the standard model and other models used for MD forecasting,[10–14] we restrict our analysis to forecasting future reliable VFs.

Methods

Institutional review board approval was obtained at the Johns Hopkins University School of Medicine and at the centers contributing data to the Glaucoma Research Network. The study adhered to the tenets of the Declaration of Helsinki.

Study Participants

We included 3614 participants from the Glaucoma Research Network dataset. Data were collected over a period of 21 years starting from 1996. In order to improve the accuracy of estimation of disease progression with linear regression, we adapted previous work and chose to study eyes with at least five VFs.[15] Specifically, we included eyes which had five or more VFs obtained with the Humphrey Field Analyzer (HFA II; Carl Zeiss Meditec Inc., Dublin, CA) using the Swedish Interactive Threshold Algorithm (SITA) Standard test protocol and the 24-2 pattern. Patients could have either one or both eyes included in the analyses.

VF and Clinical Data Collection

For each eye, disease severity at baseline was calculated by taking the average MD value for the first two VFs. Baseline MD $>$ −6 dB was categorized as mild/suspect disease, baseline − 6 dB $\geq$ MD $\geq$ −12 dB was categorized as moderate disease, and the remaining eyes with MD $<$ −12 dB were considered advanced glaucoma using established guidelines.[16]

Each VF contains reliability indices, such as the percentage of FPs, the percentage of FNs, and the test duration. With these reliability indices, we can compute a measure of how unreliable the MD is based on a previous study which we modeled the difference between predicted MD values and observed MD values as a function of these reliability indices.[7] Table 1 provides the average effect that each reliability index has on ΔMD = *measured* MD − *true* MD. To predict "true MD" from measured MD values, we used Table 1 to add up the effects of FP, FN, and test duration. Fixation loss percentages were also available, but they were not used to estimate the level of unreliability as they have been shown to not significantly affect the MD values.[7] For example, an eye with moderate glaucoma that had 10% FPs, 0% FNs, and a duration that was 30 seconds longer than the average for moderate eyes

would have an expected $\Delta MD = 10 \times 0.073 + 0.5 \times -0.35 = 0.55$ dB. Assuming the measured MD value was -8.25 dB, we would expect that the true value is *true* MD $=$ *measured* MD $- \Delta MD = -8.25 - 0.55 = -8.80$ dB.

In a clinical setting, a 1 dB error in MD can be considered an acceptable level of error.[7] Because we are primarily interested in predicting future reliable VFs, we restricted the error threshold that defines reliability even further. Thus, we only included eyes in the analysis where the final VF was a "gold standard VF" defined as having an error less than 0.25 dB (i.e. $|\Delta MD| \leq 0.25$ dB). For each eye, we calculated the percentage of VFs which were labeled as unreliable ($|\Delta MD| > 0.25$ dB). Then, we divided eyes into subgroups based on the percentage of visits with unreliable VFs. The subgroups were: 0% unreliable, 0% to 10%, 10% to 20%, 20% to 30%, 30% to 40%, 40% to 50%, and eyes with more than 50% unreliable VFs.

Modeling MD Values Over Time

The same set of eyes was used to test all models, that is, each eye had at least five VFs with the last VF being reliable ($|\Delta MD| \leq 0.25$ dB). For each eye, four linear regression models were fit using all VF visits except the final visit. For all four models, the date of the VF was the independent variable. For the standard model, the measured MD value was the dependent variable; however, only VFs which met the HFA reliability criteria (cutoff of less than 20% FLs and 15% FPs) were included in the regression.[9] The unfiltered model also used measured MD values as the dependent variables but did not exclude any points for the regression. The corrected model used "corrected MD" (i.e. *measured* MD $-$ ΔMD) as the dependent variable; individual MD values were first corrected and then fit with a regression line. The weighted model used measured MD values as the dependent variables but weighed MD values by their reliability. The precise form of the weight for the ith MD value is $weight\ of\ MD_i = \frac{\gamma}{\gamma + |\Delta MD_i|^{\rho}}$ where γ and ρ were optimized to fit the data using a 2D grid search over the entire dataset. That is, we searched over a large range of non-negative values and chose the γ and ρ, which minimized the average difference between the measured MD values for the final VF and the estimated MD values from the weighted model, across all eyes.

Once we determined the optimal model fit using these four approaches, for each eye, we used the date of the most recent VF to predict an MD value for each model at that date. We then calculated the magnitude of the residual (i.e. the absolute value of the difference between predicted MD and measured MD values). We will refer to the magnitudes of these residuals as the "standard residual magnitude," "unfiltered residual magnitude," "corrected residual magnitude," and "weighted residual magnitude." This entire process is depicted in Figure 1.

Following these residual calculations, the distribution of the standard residual magnitudes was compared to the distribution of the unfiltered, corrected, and weighted residual magnitudes. Note, these distributions were not normally shaped, so a parametric statistical test which assumes normality would not be valid for comparison. In order to compare the distributions with a statistical test, we performed a Wilcoxon signed-rank test across each pair of distributions generated when comparing two models to test the hypothesis that the means of the magnitudes of the residuals were significantly different.[17] This procedure was repeated for each reliability subgroup as well as for the overall dataset. Last, we performed the same regressions and residual calculations/comparisons for subsets of the dataset with varying upper bounds of percentage unreliability. All analysis was done using Python version 3.7.

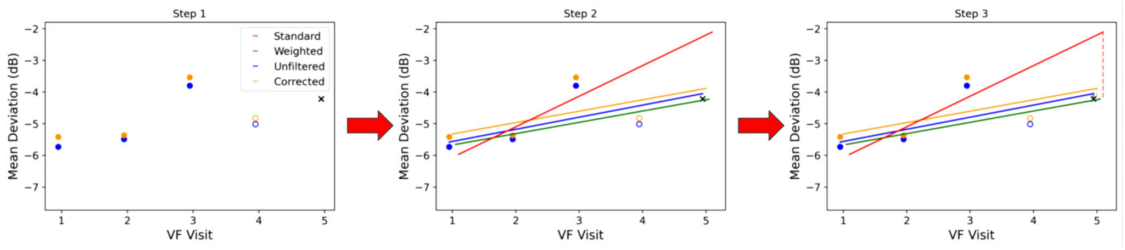

Figure 1. The first plot shows the corrected and measured mean deviation (MD) values for each visual field (VF) of a random eye. The final visual field is marked with an "x" to distinguish it from the others. The fourth MD value is hollow to indicate that the standard model would consider this VF unreliable, thereby excluding it in its regression. In the second plot, the unfiltered, corrected and weighted regression lines are fit using the first four MD values. The standard regression line is fit using only the first three MD values. In the final plot, the residuals for the most recent VF are calculated as the true MD minus the point-wise estimates from the regression fits.

Results

Demographics and Ocular Characteristics

A total of 41,120 VFs from 5939 eyes across 3614 patients were included in this study (Table 2). The mean time interval between VFs was 355 days. The mean age was 62.25 (SD = 12.94). The largest proportion of eyes (35.78%) included in the study had 0 unreliable VFs over time (see Table 2). There were 15.44% of the eyes that had more than 50% unreliable VFs over time, making this the next largest eye-level reliability subgroup. Roughly half of all eyes had between 0% and under 50% unreliable VFs.

Modeling Results

The major finding, as presented in Table 3 and Figure 2, is that the standard model performed worse than all other models, with the average of the standard residual magnitudes being higher than that of any other model (0.69 dB higher than the unfiltered model and 0.79 dB higher than the weighted and corrected models). The best performing model, measured by having the smallest average of the residual magnitudes across all the data, was the weighted model. We searched over a wide range of non-negative values to find the optimal γ and ρ parameters in the weighted regression and found the optimal parameters to be $\gamma = 0.23$ and $\rho = 1.32$. The weighted model narrowly beat the corrected model (0.006 dB, $P < 10^{-16}$), whereas both were roughly 0.13 dB better ($P < 10^{-16}$ in both cases) than the unfiltered model. However, a shortcoming of the weighted model is the need to optimize the γ and ρ parameters to derive weights of the weighted regression.

The Wilcoxon signed-rank test on the distributions of the residuals found all the means for three of our approaches to be significantly better with $\alpha = 0.05$ compared to the standard model. The same applies to each difference marked with an asterisk in Table 3. We found the mean of the weighted residual magnitudes to

Table 2. Demographics

	Mild/Suspect ($n = 4383$)	Moderate ($n = 908$)	Advanced ($n = 648$)	Overall ($n = 5939$)
Mean age, y (SD)	61.09 (12.70)	65.64 (12.81)	65.30 (13.41)	62.25 (12.94)
Mean MD, dB (SD)	−1.74 (2.16)	−8.56 (3.07)	−17.37 (5.15)	−4.49 (5.83)
Mean number of VFs (SD)	6.91 (2.37)	7.03 (2.39)	6.89 (2.36)	6.92 (2.37)
Mean unfiltered MD slope, dB/y (SD)	−0.07 (1.55)	−0.16 (1.46)	−0.07 (1.52)	−0.08 (1.54)
Mean corrected MD slope, dB/y (SD)	−0.07 (1.53)	−0.16 (1.42)	−0.07 (1.46)	−0.08 (1.50)
Mean weighted MD slope, dB/y (SD)	−0.07 (1.46)	−0.20 (1.43)	−0.17 (1.27)	−0.10 (1.43)
Mean standard MD slope, dB/y (SD)	0.01 (3.18)	−0.12 (2.44)	0.05 (5.19)	−0.01 (3.36)
Mean FP percentage (SD)	3.05 (4.94)	2.72 (3.83)	2.32 (4.43)	2.92 (4.74)
Mean FN percentage (SD)	2.96 (4.73)	7.31 (8.14)	11.02 (18.93)	4.51 (8.56)
Mean duration, s (SD)	338 (57)	433 (68)	447 (65)	364 (74)
Age brackets				
5<50	695 (15.86%)	87 (9.58%)	75 (11.57%)	857 (14.43%)
50–59	1032 (23.55%)	140 (15.42%)	102 (15.74%)	1274 (21.45%)
60–69	1405 (32.06%)	291 (32.05%)	185 (28.55%)	1881 (31.67%)
70–79	1019 (23.25%)	265 (29.19%)	201 (31.02%)	1485 (25.00%)
80–89	221 (5.04%)	117 (12.89%)	81 (12.50%)	419 (7.06%)
≥90	11 (0.25%)	8 (0.88%)	4 (0.62%)	23 (0.39%)
Percent unreliable				
0%	1824 (41.62%)	228 (25.11%)	73 (11.27%)	2125 (35.78%)
(0%, 10%])	83 (1.89%)	14 (1.54%)	2 (0.31%)	99 (1.67%)
(10%, 20%])	686 (15.65%)	141 (15.53%)	68 (10.49%)	895 (15.07%)
(20%, 30%])	584 (13.32%)	131 (14.43%)	81 (12.50%)	796 (13.40%)
(30%, 40%])	320 (7.30%)	83 (9.14%)	92 (14.20%)	495 (8.33%)
(40%, 50%])	376 (8.58%)	126 (13.88%)	110 (16.98%)	612 (10.30%)
> 50%	510 (11.64%)	185 (20.37%)	222 (34.26%)	917 (15.44%)

Table 3. Difference in Residuals for Each Eye-Level Reliability Subgroup

	Percentage of VFs Unreliable							
	$x = 0$	$0 < x \leq 10$	$10 < x \leq 20$	$20 < x \leq 30$	$30 < x \leq 40$	$40 < x \leq 50$	$50 < x$	Overall
Difference in residual (standard - unfiltered), dB	0.34*	0.20*	0.35*	0.65*	0.54*	1.57*	1.43*	0.69*
Difference in residual (standard - corrected), dB	0.35*	0.23*	0.41*	0.74*	0.68*	1.71*	1.73*	0.79*
Difference in residual (standard - weighted), dB	0.34*	0.25*	0.44*	0.81*	0.70*	1.75*	1.65*	0.79*
N	2125	99	895	796	495	612	917	5939

Significance using Wilcoxon signed-rank test: * $\equiv$ $p < 0.05$.

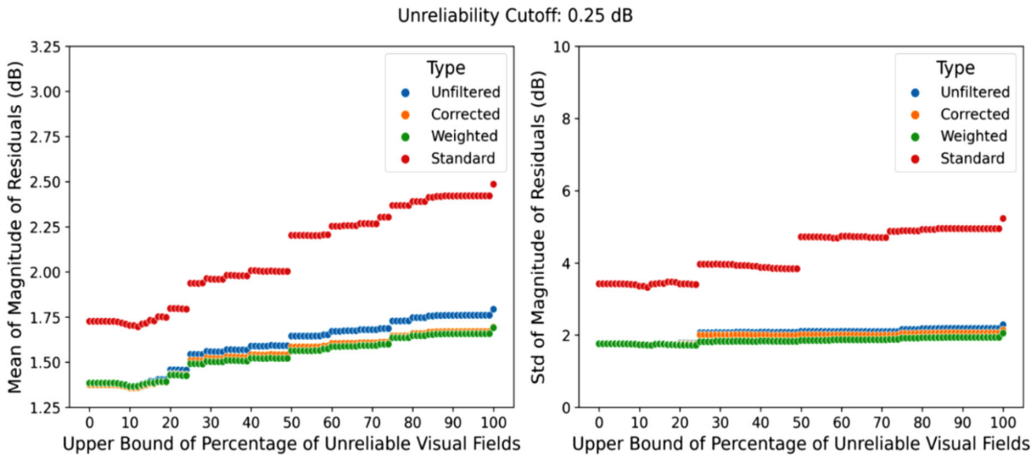

Figure 2. Mean magnitude of residuals (*left*) and their standard deviations (*right*) are shown for the standard, unfiltered, corrected and weighted models as a function of the percentage of maximum eye-level unreliability (x-axis). Each point P on the x-axis includes all eyes were at most P% of the VFs were unreliable.

be smaller than the means of the unfiltered, corrected, and standard residual magnitudes.

Inspecting the eye-level reliability subgroups in Table 3 reveals that some subgroups performed best using the corrected model. Those were eyes with no unreliable VFs and eyes with greater than 50% unreliable VFs. Although, the improvement gained by using the corrected model over the weighted model is minimal. For all subgroups, the mean of the magnitudes of the unfiltered, corrected, and weighted residuals was significantly smaller than that of the standard residuals. The largest differences in the mean of the standard residual magnitudes versus the mean of the weighted residual magnitudes occurs in the 40% to 50% eye-level reliability subgroup, in which the average of the weighted residual magnitudes is 1.75 dB smaller than the average of the standard residual magnitudes. The biggest difference in performance between the standard and corrected models (1.73 dB) occurs in the greater than 50% unreliable VF subgroup.

Figure 2 shows the mean magnitudes of standard, unfiltered, corrected, and weighted residuals as a function of the percentage of maximum eye-level unreliability – each point **P** on the x-axis includes all eyes were at most **P**% of their VFs were marked as unreliable. Note, as the percentage of maximum unreliability increased, the differences between each regression method widened. For eyes with no unreliability, the corrected regression performed slightly better than the weighted and unfiltered regressions. Yet, as more unreliability is introduced, the weighted regression begins to outperform the corrected and unfiltered regressions. The standard deviation graph shows that as more unreliable VFs are included in the analysis, the distribution of the weighted regression experiences the smallest variance. This signifies that as unreliability increases, the residuals are growing in magnitude and experiencing more variance across all eyes. Even so, the weighted regression has stronger predictive power than both the corrected and unfiltered regression methods. The weighted model also exhibited greater precision in its estimates. The average size of the 95% prediction interval for the weighted model was 7.67 compared to 8.70 for the corrected model and 8.80 for unfiltered linear regression.

Sensitivity Analysis

Our analysis required the final VF for each eye to be marked as reliable with $|\Delta MD| \leq 0.25$ dB. Recall,

this final VF is not used in the model fit and is reserved for measuring predictive power of the model. Other error thresholds were tested, including 0.5 dB, 0.75 dB, and 1 dB. Compared to 5939 eyes with a 0.25 dB threshold, there were 7235 eyes with a 0.5 dB threshold, 7531 eyes with a 0.75 dB threshold, and 7652 eyes with a 1 dB threshold. The γ and ρ parameters for the weights on the weighted regression were optimized for each set of data, respectively. In all analyses, the three approaches beat the standard model and the corrected and weighted models outperformed the unfiltered model. Within each of these threshold

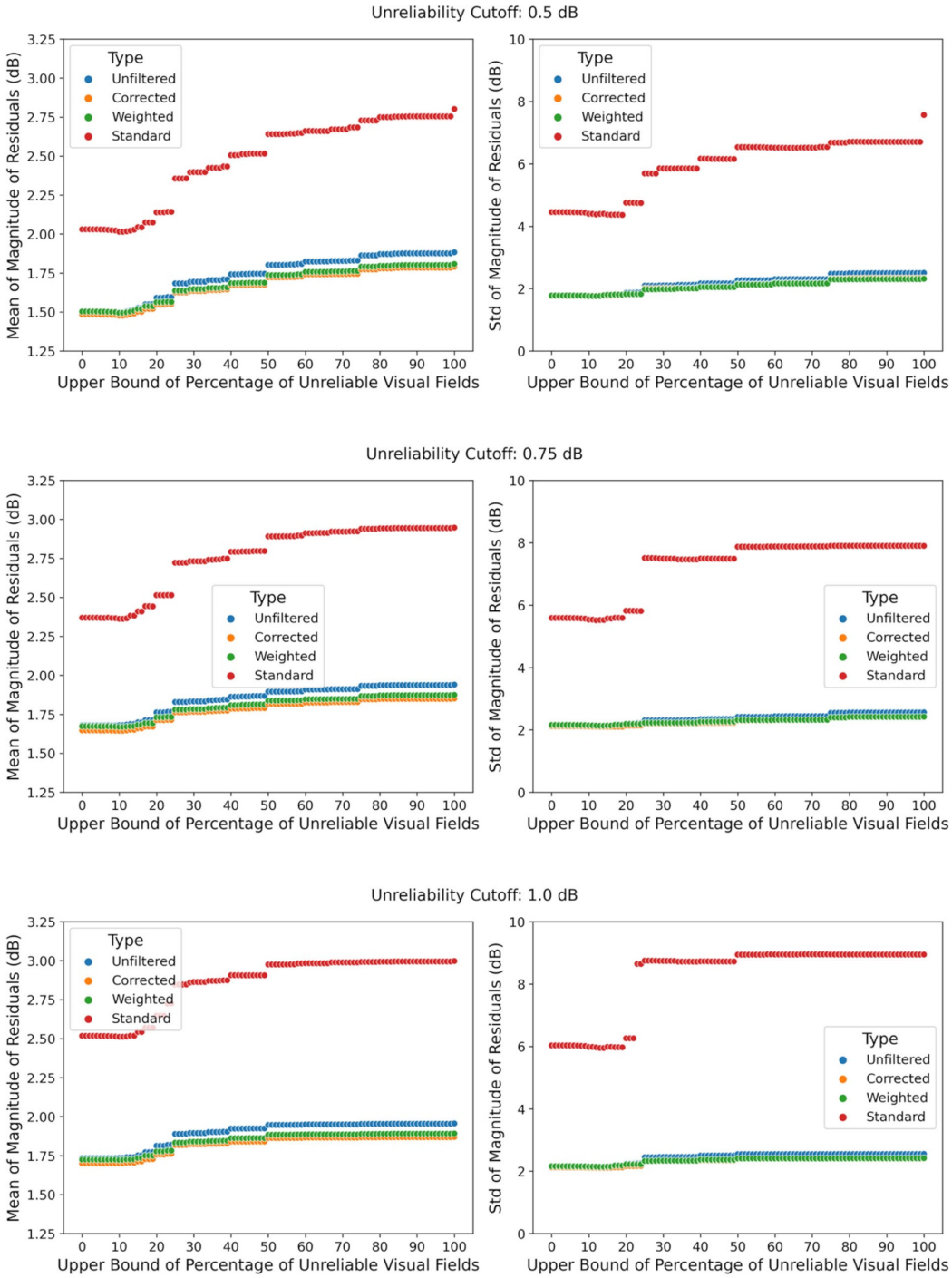

Figure 3. Mean magnitude of residuals (*left column*) and their standard deviations (*right column*) are shown for the standard, unfiltered, corrected, and weighted models as a function of the percentage of maximum eye-level unreliability (x-axis) and different error cutoffs for the last VF: 0.5 dB (*top row*), 0.75 dB (*middle row*), and 1.0 dB (*bottom row*).

analyses, the corrected model had the lowest mean magnitude of the residuals, although only marginally better than the weighted model.

Using the various error cutoffs for the last VF described above (0.25 dB, 0.5 dB, etc.), we explored the effect of eye-level reliability on the residual of the model prediction (Fig. 3). As unreliability increased, the difference between the mean magnitude of the standard residuals and those of the other methods increased. The largest difference in the means occurred in the 1 dB cutoff group, where the corrected model performed 1.13 dB better than the standard model. Moreover, the mean residuals for all methods grew as unreliability increased. This is as expected because more unreliable VFs were included in the analysis, so the model was not as good at estimating the future gold-standard MD values. For example, in the 0.25 dB group, the mean of the magnitude of corrected residuals was 1.70 dB, followed by 1.79 dB, 1.85 dB, and 1.87 dB in the 0.5 dB, 0.75 dB, and 1 dB cutoff groups, respectively.

Last, we wanted to test the generalizability of the optimal weighted regression parameters for other thresholds. First, we wanted to see if the optimal γ and ρ parameters used for the 0.25 dB error threshold achieved similar levels of error for other thresholds. After running the weighted regression with the 0.25 dB parameters on the other thresholds, we determined the results were comparable to the original threshold. The weighted model consistently performed within 0.02 dB of the corrected model and consistently beat the standard model by 1 dB or more. Thus, our analysis suggests that the parameters obtained using the 0.25 dB threshold ($\gamma = 0.23$ and $\rho = 1.32$) are applicable to other thresholds without a meaningful change in model error. Second, we wanted to see if in the 0.25 dB error threshold, there was high variability as the γ and ρ parameters fluctuated. We ran the same analysis using 121 combinations of the γ and ρ parameters within a 0.1 dB range from the optimal. We found there was a negligible difference in performance on the order of 0.0001 dB for mean magnitude of residuals. Third, we wanted to see how sensitive the predictions of the weighted model were to the training set. We compared the mean magnitude of residuals when using 50% of the data as the training set and 50% as the test set — model parameters were optimized based on a randomly chosen 50% of the data — to the mean magnitude of residuals without splitting the data. The standard deviation of the difference in residuals between the train/test split weighted model and the weighted model without the train/split was 0.013 dB over 100 simulations, which is larger than the 0.006 dB difference between the non-split weighted and corrected models. Thus, the statistically significant difference between the weighted and corrected models may not necessarily hold with different datasets. However, it is likely that the difference between the weighted and standard models remains statistically significant.

Discussion

The results of this study suggest that all three approaches – unfiltered, corrected, and weighted models – outperform the standard model for predicting future VF changes. Importantly, using all the VF data (including data that is labeled by the HFA as unreliable) results in more accurate predictions of future MD values as demonstrated by the unfiltered model improving upon the standard model by 0.69 dB. Correcting MD values for poor reliability using either an arithmetic approach (corrected model) or weighted regression approach (weighted model) results in further improvements (on the order of 0.1dB) in predicting future VF change.

To the best of our knowledge, no other study has tried to correct for poor reliability when building predictive models for VF change. Although, other groups have tried to forecast VF change with novel algorithms as well. Notably, JC Wen et al. (2019) experimented with various machine learning algorithms to predict future VFs using early VF data.[18] Their experiment design was different in that their models used the entire VF as input with the goal of predicting a future VF. After testing hundreds of models, their best performing model achieved a mean MD difference of +0.41 dB. Note, this value appears to reflect the average of the raw residuals in the forecasting model. Our reported results look at magnitudes of the residuals which are likely superior for calculating model error, as magnitude of error could be large, but if it is perfectly symmetrical then the average raw residual would be 0 dB which is uninformative. If we were to compare our mean raw residuals, our corrected and weighted models achieve a mean MD difference of +0.01 dB and −0.02 dB, respectively. Garcia et al. (2019) used Kalman filtering to predict MD values 5 years into the future and were able to predict within 2.5 dB for the majority of eyes.[12] Although the prediction time interval is much wider than in our study (most of the MD values we predict in our analysis were 1 year into the future), our model performance was better in terms of MD error.

The reliability subgroup analysis depicted in Figure 2 demonstrates that, as more unreliable VGs are introduced into the regression to predict future MD, the performance gap between the standard

model and our three models widens. That is, the more unreliable a patient is at taking VF tests, the better our models become at predicting future VF change compared to the standard approach (GPA) used in the HFA. Granted, the predictions are still not as good when more unreliability is introduced. Specifically, the weighted and corrected models achieve their greatest performance improvement over the standard model when roughly half of an eye's VFs are unreliable. Moreover, as the reliability error threshold gold standard increased (as is done in the sensitivity analysis), the performance gap between the best performing (corrected and weighted model) and worse performing models (standard model) widened further. This once again suggests that our models tend to make a larger impact on correctly predicting future VF change when there is a larger amount of unreliability. The sensitivity analysis suggests that the corrected and weighted models should be used for predicting future reliable VFs over the standard model, as across levels of unreliability, their performance surpasses that of the standard model.

Our study has several limitations. First, our analysis was limited to eyes with at least one reliable VF. Therefore, the results of this study cannot be generalized to eyes where all the VFs are unreliable. Second, as stated previously, the weighted model's parameters were optimized for the dataset used in the analysis; yet we demonstrated in the sensitivity analysis that the chosen parameters generalized to different error thresholds and even maintained comparable performance when the parameters themselves fluctuated. However, the statistically significant difference between the weighted and corrected models may not necessarily generalize. Third, the model performance is subject to the inclusion criteria. Whereas the proposed three approaches outperformed the standard model across all tested error thresholds (0.25 dB, 0.5 dB, 0.75 dB, and 1 dB), we have not compared the model performance for larger thresholds; however, as the threshold increases, the VF is considered less reliable, and we anticipate the gap in performance between the weighted and corrected models would widen. Fourth, defining "gold standard" VFs based on a threshold error (e.g. 0.25 dB) compared to a linear regression model artificially selects eyes where linear regression is more accurate. An alternative is to select eyes based on reliability indices alone, for example, where the final VF had <5% FNs and FPs. It turns out these two sets are highly similar, with 97.4% of all eyes with <5% FNs and FPs being in the 0.25 dB set. The 0.25 dB set was also the larger superset: 16.2% of eyes in the 0.25 dB set were not in the <5% FN and FP set, whereas only 2.6% in the <5% FN and FP set were not in the 0.25 dB set. Fifth, by default, the GPA performs a regression of visual field index (VFI) over time, although many clinicians will use the MD value over time trend instead,[9] as there is practically no difference in trend estimation (VFI versus MD over time) for most eyes.[19] Finally, we note that the 95% prediction interval (PI) for all models were relatively large, which means that these models may not be very useful for predicting future measurements.

In conclusion, we found the standard modeling approach built into the HFA perimetry to be the worst-performing for this analysis, as it excludes eyes with unreliable data. By excluding these data, the regression is unable to fit as well to the eye's overall VF MD value trend, resulting in a less accurate prediction of future VF change. By the same token, we found including all data, even unreliable points, allows one to more accurately predict future VF results, and in turn better assess change. Finally, we verified that correcting for unreliability and weighting by reliability further increases the accuracy of future VF prediction, which is likely to better facilitate the accuracy of progression judgments.

Acknowledgments

Supported in part by a Seed Grant of the Malone Center for Engineering in Healthcare at Johns Hopkins University; NIH 1K23EY032204-01.

Disclosure: **G.A. Villasana,** None; **C.Bradley,** None; **T. Elze,** None; **J.S. Myers,** None; **L. Pasquale,** None; **C.G. De Moraes,** None; **S. Wellik,** None; **M.V. Boland,** Carl Zeiss Meditec (C); **P. Ramulu,** None; **G. Hager,** None; **M. Unberath,** None; **J. Yohannan,** None

References

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*. 2002;120(6):701–713; discussion 829-830.
2. 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(9):1557–1565.
3. 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(4):653–662.
4. Heijl A, Leske MC, Bengtsson B, et al. Reduction of intraocular pressure and glaucoma progression: results from the Early Manifest Glaucoma Trial. *Arch Ophthalmol*. 2002;120(10):1268–1279.
5. Broadway DC. Visual field testing for glaucoma – a practical guide. *Community Eye Health*. 2012;25(79-80):66–70.
6. Recent developments in visual field testing for glaucoma : Current Opinion in Ophthalmology. https://journals.lww.com/co-ophthalmology/Fulltext/2018/03000/Recent_developments_in_visual_field_testing_for.7.aspx. Accessed April 29, 2021.
7. Yohannan J, Wang J, Brown J, et al. Evidence-based Criteria for Assessment of Visual Field Reliability. *Ophthalmology*. 2017;124(11):1612–1620.
8. Rao HL, Yadav RK, Begum VU, et al. Role of Visual Field Reliability Indices in Ruling Out Glaucoma. *JAMA Ophthalmol*. 2015;133(1):40.
9. Heijl A, Patella VM, Bengtsoon B. *The Field Analyzer Primer: Effective Perimetry*, Fifth Edition. Dublin, CA: Carl Zeiss Meditec USA, Inc.; 2012.
10. Nouri-Mahdavi K, Hoffman D, Coleman AL, et al. Predictive factors for glaucomatous visual field progression in the Advanced Glaucoma Intervention Study. *Ophthalmology*. 2004;111(9):1627–1635.
11. Park K, Kim J, Lee J. Visual Field Prediction using Recurrent Neural Network. *Sci Rep*. 2019;9(1):8385.
12. Garcia G-GP, Lavieri MS, Andrews C, et al. Accuracy of Kalman Filtering in Forecasting Visual Field and Intraocular Pressure Trajectory in Patients With Ocular Hypertension. *JAMA Ophthalmol*. 2019;137(12):1416–1423.
13. Nakazawa M, Terasaki H, Yamashita T, Uemura A, Sakamoto T. Changes in visual field defects during 10-year follow-up for indocyanine green-assisted macular hole surgery. *Jpn J Ophthalmol*. 2016;60(5):383–387.
14. Christopher M, Bowd C, Belghith A, et al. Deep Learning Approaches Predict Glaucomatous Visual Field Damage from OCT Optic Nerve Head En Face Images and Retinal Nerve Fiber Layer Thickness Maps. *Ophthalmology*. 2020;127(3):346–356.
15. Chauhan BC, Garway-Heath DF, Goñi FJ, et al. Practical recommendations for measuring rates of visual field change in glaucoma. *Br J Ophthalmol*. 2008;92(4):569–573.
16. Hodapp E, Parrish RK, Anderson DR. *Clinical Decisions in Glaucoma*. New York, NY: Elsevier Mosby; 1993.
17. Wilcoxon F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*. 1945;1(6):80–83.
18. Wen JC, Lee CS, Keane PA, et al. Forecasting future Humphrey Visual Fields using deep learning. *PLoS One*. 2019;14(4):e0214875.
19. Bengtsson B, Heijl A. A visual field index for calculation of glaucoma rate of progression. *Am J Ophthalmol*. 2008;145(2):343–353.