More Accurate Modeling of Visual Field Progression in Glaucoma: ANSWERS

Haogang Zhu,1,2 David P. Crabb,1 Tuan Ho,2 and David F. Garway-Heath2

1School of Health Sciences, City University London, London, United Kingdom
2National Institute for Health Research Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital National Health Service Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom

Correspondence: Haogang Zhu, School of Health Sciences, City University London, Northampton Square, London, EC1V 0HB UK, haogangzhu@gmail.com.
Submitted: March 25, 2015
Accepted: July 23, 2015
Citation: Zhu H, Crabb DP, Ho T, Garway-Heath DF. More accurate modeling of visual field progression in glaucoma: ANSWERS. Invest Ophthalmol Vis Sci. 2015;56:6077–6083. DOI:10.1167/iovs.15-16957

PURPOSE. To validate a method for visual field (VF) progression analysis, called ANSWERS (Analysis with Non-Stationary Weibull Error Regression and Spatial Enhancement), which takes into account increasing measurement variability as glaucoma progresses and spatial correlation among test locations.

METHODS. ANSWERS outputs both a global index of progression and a pointwise estimate of rate of change at each VF location. ANSWERS was compared with linear regression of mean deviation (MD) and permutation of pointwise linear regression (PoPLR). Visual field series of up to 2 years from the United Kingdom Glaucoma Treatment Study were used. This consists of 9104 Swedish Interactive Thresholding Algorithm Standard 24-2 VFs. ANSWERS and PoPLR rate of change were used to predict the VF at the next visit using subseries that were within 7, 13, 18, or 22 months from the baseline. The comparison was carried out on the statistical sensitivity, specificity, and accuracy of predicting future VF change.

RESULTS. Across all subseries, statistical sensitivity of ANSWERS in detecting VF deterioration was significantly better than the linear regression of MD and PoPLR, especially in short time series. Prediction accuracy of ANSWERS was better than PoPLR at all series lengths, and the improvement was particularly marked in shorter series. Seventy-five percent of VF series were better predicted by ANSWERS compared with PoPLR. The average prediction error of ANSWERS was 15% lower than that of PoPLR.

CONCLUSIONS. ANSWERS is more sensitive to detect VF progression and predicts future VF loss better than linear regression of MD and PoPLR, especially over short observation periods.

Keywords: visual field, progression, change detection, nonstationary, spatial correlation

Management of glaucoma relies on visual field (VF) measurement using standard automated perimetry (SAP), which assesses differential light sensitivity (DLS) across a subject’s field of view.1 Accurate and precise assessment of VF change over time is essential for appropriate clinical management of glaucoma so that patients whose condition is worsening receive prompt treatment intervention while those with a stable condition are not overtreated. Currently, however, VF measurement is highly imprecise and has complex statistical properties, which make monitoring changes in VF challenging.

Clinical evaluation of glaucomatous change in VF series can be supported by analytical algorithms. Probability of change and rate (velocity) of change can be derived from these algorithms. These two parameters can be estimated with methods known as trend analyses. Pointwise linear regression (PLR), the most widely used trend analysis, fits an ordinary linear regression model for each location in the VF and assesses the significance and slope of the fit.2 Summary measures, such as mean deviation (MD) from the average DLS of healthy eyes, are also often used in trend analysis; but, since glaucoma tends not to affect all locations to the same extent, global indices often have inadequate statistical sensitivity to detect worsening compared with methods assessing deterioration at individual locations.3 Permutation analyses of pointwise linear regression (PoPLR),4 a recent advance in PLR trend analysis, involves a random permutation of the order of VFs in a series. It has been reported to provide a better estimate of overall statistical significance of change compared with PLR.5 The significance of change in PoPLR is estimated in the context of permuted series of VFs, which assumes no change in reordered series. As there is a need for permutation in VF series, this technique cannot estimate reliably the significance of change in series with fewer than six VFs because of the limited number of possible permutations in such series. Moreover, despite a different method to estimate the overall statistical significance by PoPLR, the underlying regression model is still that of ordinary linear regression, and therefore the estimate of rate of change and the statistical significance of change at individual locations are identical to those of PLR.

Two important properties in VF measurement are not accounted for in the current methods for detecting change in VF series: nonstationary variability (increasing variability as DLS declines) and spatial correlation among test locations. Visual field measurements are subject to considerable variability, which increases as DLS deteriorates with disease progression, and eventually decreases in blind regions.6–8 For instance, the repeat measurement range (90% confidence interval [CI]) is 7 dB (26–33 dB) when DLS is healthy at 32 dB, while this range...
increases to 18 dB (5–27 dB) when DLS deteriorates to 20 dB. This changing variability over time is referred to as nonstationary measurement variability. Furthermore, the traversing of the VF test grid by retinal nerve fibers results in correlation between spatially related locations. The most widely used SAP VF measurements, such as those taken by Humphrey Visual Field Analyzer (HFA; Carl Zeiss Meditec, Dublin, CA, USA), are made in a regular grid across a patient’s field of view. Aside from the neighborhood of test locations, the spatial correlation is also governed by the anatomical arrangement of the retinal nerve fibers. Betz-Stablein et al. incorporated such spatial correlation in six regions of VF corresponding to the six sectors of the optic disc and demonstrated improved performance in detecting VF progression. Therefore, without taking into account these statistical properties, the detection of change in VF with current methods is potentially delayed or requires more clinic visits than necessary.

A new trend analysis method, Analysis with Non-Stationary Weibull Error Regression and Spatial Enhancement (ANSWERS), was proposed and validated on a large dataset acquired from electronic health records. In contrast to commonly used ordinary linear regression models, which assume fixed and normally distributed errors, ANSWERS incorporates the nonstationary variability at different levels of DLS modeled as mixtures of Weibull distributions. Spatial correlation of measurements was also included in the model using a Bayesian framework. Despite its optimized statistical attributes, ANSWERS still acts as a linear regression model that outputs both the probability of no deterioration and rate of change at individual locations in the VF and can be interpreted in the same way as PLR. It also produces a global deterioration index summarizing the overall probability of change in the series. The details about derivation and implementation of ANSWERS can be found elsewhere.

This study compared ANSWERS with PoPLR and the linear regression of MD on a dataset from a clinical trial. The assumption was that a more accurate method for modeling VF series should (1) be more sensitive in detecting change under the same specificity and (2) predict future VFs more accurately.

**METHODS**

The methods under comparison included ANSWERS, PoPLR, and linear regression of MD. Additionally, in order to investigate the effect of incorporating spatial correlation in ANSWERS, its effect can be switched off and the method is referred to as ANSWER.

**Datasets**

All VFs were measured with the HFA (Carl Zeiss Meditec) using the 24-2 test pattern and the SITA (Swedish Interactive Thresholding Algorithm) Standard testing algorithm. The test measures retinal DLS at 52 test locations excluding two points in the blind spot region.

Two datasets collected at different centers were used in this study. The first dataset contains VF series from the United Kingdom Glaucoma Treatment Study (UKGTS), a randomized, double-masked placebo-controlled clinical trial testing the hypothesis that treatment with a topical prostaglandin analogue, compared with placebo, reduces the frequency of VF deterioration events. Patients were followed up for 2 years or until reaching the endpoint criteria. During the 2-year period, patients were tested at 2, 4, 7, 10, 13, 16, 18, 20, 22, and 24 months from baseline, and two repeated VF tests were taken at baseline and at 2, 16, 18, and 24 months from baseline. Details of the dataset have been described elsewhere.

Visual field tests with false-positive reliability responses over 15% were discarded. Only series that were obtained over at least 4 months (three visits) were included in the analysis. Note that the length of series is purely for evaluation purposes and is not necessitated by ANSWERS. The resulting dataset consisted of 9104 VF tests from 659 series of 437 patients. The median (interquartile range [IQR]) time of follow-up was 22 (15–24) months, and the median (IQR) number of VFs in the series was 11 (6–12).

The second dataset was from a study examining the test-retest variability of VF test conducted at Moorfields Eye Hospital, London, United Kingdom, in a cohort of glaucoma patients. As changes in retinal function are slow in glaucoma, it is possible to estimate test measurement variability by taking repeat measurements in a short period of time under the assumption that no measurable deterioration can occur over the observation period. Fifty-two eyes of 27 patients were tested 10 times over a short period (maximum 10 weeks). The variance among VFs in these repeat measures indicates the inherent measurement variability. Furthermore, the VF series for each eye, and the same series with arbitrary reordering, represent a stable series with no underlying change. The use of randomly reordered series for the estimate of measurement variability is an established method in various studies.

Patients’ data were anonymized prior to investigation and did not contain personal or sensitive information. The data were held in a secure database at City University London. As such, patients’ written consent for their data to be used in the study was not required. The study adhered to the tenets of the Declaration of Helsinki and was approved by the research governance committee of City University London, United Kingdom. The anonymized dataset can be accessed upon request.

**False-Positive Rate and Statistical Sensitivity for Change Detection**

The false-positive rate and statistical sensitivity were compared for the four trend analysis methods. A false positive is a type I error when change is detected in a series with no true progression. The false-positive rate can be estimated in the series of repeated measurements acquired in a short period of time. Moreover, randomly reordering these repeated measurements produces additional pseudo-series where there is also no true deterioration.

The series of 10 VFs from each eye in the test-retest dataset were randomly reordered 300 times. With 52 eyes in the test-retest data, 124,800 (52 × 300 × 8) pseudo-series with eight different series lengths of between 3 and 10 VFs were generated. It was assumed that five VFs per year were taken in these pseudo-series (the median test frequency in the UKGTS dataset). The false-positive rate was then estimated as the proportion of series identified as progressing. In a clinical situation, false positives may lead to overtreatment and unnecessary cost, so methods with high false-positive rates are considered not to be clinically useful.

Comparison of the different methods should be made at equivalent false-positive rates, which is dependent on the chosen change criterion and the length of the series. For PoPLR, the deterioration criterion was an overall statistical significance of change smaller than a given threshold. For ANSWERS and ANSWER, the criterion was a deterioration index higher than a given threshold. For linear regression of MD, the criteria were a negative slope and slope value lower than a set threshold. For each method, a set of thresholds was chosen to achieve specified false-positive rates, and the statistical sensitivity of each method was then compared at equivalent false-positive rates.
Statistical sensitivity is a measure of identifying true change. Ideally, the sensitivity should be evaluated as the proportion of detected progression in VF series with true underlying deterioration. However, due to the lack of a gold standard and ground truth classification for glaucomatous deterioration, the underlying progression status of each VF series was unknown. Therefore, the methods were compared using the positive rate, which is the proportion of series flagged as progressing in the UKGTS dataset. Given an unknown proportion (p%) of truly progressing series in the dataset, the positive rate was linked to statistical sensitivity as positive rate \( = (p\% \times \text{sensitivity}) + [(1 - p\%) \times \text{false-positive rate}] \). Note that if the false-positive rate is controlled to be equivalent for all the methods, a higher positive rate implies better sensitivity of a method. Therefore, the positive rates of all the methods were compared as a surrogate comparison for statistical sensitivity. Moreover, when the false-positive rate is low, the positive rate is dominated by the sensitivity. Therefore when comparing two methods at lower false-positive rate, the ratio of positive rate between the methods is closer to the ratio of sensitivity. The comparison was made with series of 7, 13, 18, and 22 months from baseline.

**Prediction of Future Visual Field**

One important clinical question regarding modeling of VF progression is the projection of future VF loss, which is closely related to the rate of change in VF series. It was assumed that better modeling of progression would lead to a more accurate estimate for the rate of change and hence a better prediction of future VF. The comparison was carried out for ANSWERS, ANSWER, and PoPLR using the raw DLS measurements at all locations. Note that the rates of change and predictions from PoPLR are exactly those from ordinary linear regression at individual VF locations.

Subseries of the UKGTS data were used to predict the VF at the next visit, with the shortest subseries including VFs at the first three visits (4 months) from baseline. The subseries increased in length to include more visits in a chronologically ascending order. For each subseries of the same length, the three methods were used to estimate the rates of change at individual locations, which were then used to predict the VF in the subsequent visit. The prediction performance was evaluated as mean normalized squared error (MNSE) between the predicted and measured VFs across 52 locations:

\[
\text{MNSE} = \frac{1}{52} \sum_{i=1}^{52} \frac{A_i^2}{\sigma_i^2}
\]

The MNSE is the average squared prediction error \( A_i^2 \) in percentage with regard to the measurement variance \( \sigma_i^2 \).

![Figure 1](image1.png)  
**Figure 1.** ANSWERS change criterion at various false-positive rates and lengths of series (number of fields in the series). The ANSWERS threshold was estimated with false-positive rates between 2% and 10% and length of series between 4 and 10. Each curve represents the ANSWERS threshold for the false-positive rate indicated at the end of the curve. The thresholds with series longer than 10 (the part on the right side of the vertical dashed line) are extrapolated.

![Figure 2](image2.png)  
**Figure 2.** Positive rates of ANSWERS, ANSWER, PoPLR, and linear regression of MD in VF subseries at 7, 13, 18, and 22 months from baseline. The positive rates are estimated at false-positive rates between 0% and 15%.
The trend analysis methods were also compared at the 5% false-positive rate. The ratios of positive rates between pairs of methods are shown in Table 2, where a ratio > 1 indicates a better positive rate. For instance, with subseries of 7 months, the ratio of ANSWERS over PoPLR was 1.71, indicating that the positive rate of ANSWERS is 1.71 times that of PoPLR.

In all subseries, the positive rates of ANSWERS and ANSWER were higher than those of PoPLR and linear regression of MD. Improvement was even greater in short subseries. The spatial enhancement included in ANSWERS also increased the positive rate compared with ANSWER, especially in short subseries. However, this improvement became marginal as the length of the subseries increased.

Estimate of Rate of Change
In all subseries, the average rate of change (median [IQR]) across all VF locations estimated by ANSWERS, ANSWER, and PoPLR was 0.12 (−0.44 to 0.67), 0.12 (−0.41 to 0.65), and 0.16 (−0.73 to 0.98) dB/year, respectively. The (unsigned) magnitude (median [IQR]) of average rate of change is 0.55 (0.26–1.07), 0.53 (0.25–0.99), and 0.87 (0.40–1.70) dB/year for ANSWERS, ANSWER, and PoPLR.

Both ANSWERS and ANSWER made significantly ($P < 0.01$, Wilcoxon signed rank test) smaller estimates of the magnitude of the rate of change compared with PoPLR. ANSWERS provided a statistically significant ($P < 0.01$, Wilcoxon signed rank test) greater rate magnitude compared with ANSWER. The comparison of rate of change between ANSWERS, ANSWER, and PoPLR in subseries of 7, 13, 18, and 22 months is presented in Figures 3 and 4, in which the relative relationship between the magnitude of the rate of change from the three methods (PoPLR > ANSWERS > ANSWER, $P < 0.01$ Wilcoxon signed rank test) was consistent in all subseries, except that for 22 months, where the rate magnitude did not differ between ANSWERS and ANSWER ($P = 0.20$, Wilcoxon signed rank test). The results can be seen in Figure 3, where the points scatter around a line with a slope of less than 1, and in Figure 4, where the points scatter around a line with a slope of more than 1, except for those at 22 months. For 13 and 18 months, although the difference between ANSWERS and ANSWER is statistically different, the amount of the difference is minimal so the points scatter closely to the diagonal line in Figure 4.

Prediction of Future Visual Field
In all subseries of VFs, the MNSE (median [IQR]) for ANSWERS, ANSWER, and PoPLR was 54% (33%–113%), 60% (36%–122%), and 76% (45%–146%), respectively. ANSWERS provided better prediction (lower MNSE) than PoPLR and ANSWER in 75% and 71% of VFs, respectively. The MNSE from ANSWERS was significantly smaller ($P = 0.01$, Wilcoxon signed rank test) than those from PoPLR and ANSWER (median [95% CI] difference: 15% [10%–19%] and 2% [1%–4%], respectively). The comparison between the three methods for prediction of VFs at 10, 16, 20, and 24 months using subseries of 7, 13, 18 and 22 months is summarized in Table 3. ANSWERS outperformed PoPLR in all subseries.

Table 2. Ratio of the Positive Rates (at 5% False-Positive Rate) for ANSWERS and ANSWER Over Those of ANSWER, PoPLR, and Linear Regression of MD

<table>
<thead>
<tr>
<th></th>
<th>7 mo</th>
<th>13 mo</th>
<th>18 mo</th>
<th>24 mo</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANSWERS</td>
<td>1.47</td>
<td>1.18</td>
<td>1.06</td>
<td>1.06</td>
</tr>
<tr>
<td>ANSWER</td>
<td>1.71</td>
<td>1.47</td>
<td>1.41</td>
<td>1.75</td>
</tr>
<tr>
<td>PoPLR</td>
<td>2.03</td>
<td>1.70</td>
<td>1.75</td>
<td>2.06</td>
</tr>
</tbody>
</table>

Comparison was carried out with VF subseries at 7, 13, 18, and 24 months.
FIGURE 3. Scatterplot of average rate of change from ANSWERS against that from PoPLR in VF subseries at 7, 13, 18, and 22 months from baseline. Semitransparent points are used to relieve overlapping in the plot. The straight dashed line represents a reference line with a slope of 1.

FIGURE 4. Scatterplot of average rate of change from ANSWERS against that from ANSWER in VF subseries at 7, 13, 18, and 22 months from baseline. Semitransparent points are used to relieve overlapping in the plot. The straight dashed line represents a reference line with a slope of 1.
**DISCUSSION**

ANSWERS is a more sensitive method to detect VF progression, and is more accurate in predicting future VFs, compared to the other trend-based methods evaluated. At equivalent false-positive rates, it detected a greater number of eyes with change compared with PoPLR and linear regression of MD. In addition, the results indicate that future VFs can be better predicted by ANSWERS than the other methods. The Weibull mixture retest distribution, compared with a normally distributed error in ordinary regression models, captures the mixture retest distribution, compared with a normally distributed error in ordinary regression models. Notably, the prediction of future VF acted as a separate validation of the statistical methods, independent of the test–retest data.

**TABLE 3.** Median (Interquartile Range) Mean Normalized Squared Error (MNSE) of ANSWERS, ANSWER, and PoPLR for Prediction of VFs at 10, 16, 20, and 24 Months From Baseline

<table>
<thead>
<tr>
<th></th>
<th>10 mo</th>
<th>16 mo</th>
<th>20 mo</th>
<th>24 mo</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. VFs for prediction</td>
<td>6</td>
<td>8</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>ANSWERS, %</td>
<td>57 (38 to 114)</td>
<td>53 (33 to 104)</td>
<td>49 (30 to 104)</td>
<td>43 (31 to 121)</td>
</tr>
<tr>
<td>ANSWER, %</td>
<td>67 (45 to 124)</td>
<td>58 (36 to 111)</td>
<td>49 (30 to 112)</td>
<td>46 (32 to 129)</td>
</tr>
<tr>
<td>PoPLR, %</td>
<td>101 (65 to 192)</td>
<td>70 (44 to 131)</td>
<td>63 (53 to 110)</td>
<td>49 (32 to 109)</td>
</tr>
<tr>
<td>PoPLR – ANSWERS</td>
<td>38 (26 to 51)*</td>
<td>25 (12 to 34)*</td>
<td>16 (13 to 19)*</td>
<td>10 (* to 15)*</td>
</tr>
<tr>
<td>ANSWER – ANSWERS</td>
<td>15 (12 to 14)*</td>
<td>6 (5 to 8)*</td>
<td>1 (0.01 to 2.02)*</td>
<td>1 (* to 0.6 to 2.7)*</td>
</tr>
</tbody>
</table>

The last two rows summarize the median (95% CI) MNSE difference between ANSWER and ANSWERS, and between PoPLR and ANSWERS.

* A statistically significant (P < 1% in Wilcoxon signed rank test) difference.

Subseries. The improvement was greater in shorter subseries. The spatial enhancement in ANSWERS made it a better predictor than ANSWER for all VF predictions except for those at the 24th month.

Figure 5a shows the improvement in VF prediction by ANSWERS compared with PoPLR, where PoPLR MNSE minus ANSWERS MNSE was plotted against the amount of change (average difference between the VF being predicted and baseline VF). Seventy-five percent of VFs (above dashed line) were better predicted by ANSWERS than by PoPLR. Table 4 shows the median (95% CI) improvement of ANSWERS, compared with PoPLR and ANSWER, when predicted VFs were between −5 and 5 dB different from baseline. ANSWERS provided better prediction irrespective of the amount of change from baseline. Therefore, compared with PoPLR, although ANSWERS produced smaller magnitude of slopes (slower rate of change), it did not make larger prediction errors in faster-progressing eyes by generally flattening the slope. Moreover, the spatial correlation used in ANSWERS made it a better predictor for future VF compared with ANSWER, regardless of the difference from baseline (Fig. 5b).

**FIGURE 5.** Improvement of ANSWERS over (a) PoPLR and (b) ANSWER, stratified by the amount of change from baseline determined as the average difference between the VF being predicted and baseline VF. Semitransparent points are used to relieve overlapping. Positive values (above dashed lines) on the y-axis indicate a better performance by ANSWERS.
All the trend analysis methods compared in this study assumed a linear change in the VF subseries. This is because there are insufficient data to identify nonlinear change, should it exist, owing to the relatively short VF series acquired in clinical practice. A recent study indicated that change in VF series may follow a nonlinear trend such as an exponential function. It is, however, simple to configure ANSWERS to model nonlinear change in long VF series. Moreover, PoPLR was used to determine criteria for progression in PLR; however, other criteria defined on the combinations of slope and statistical significance are possible.

It is important to note that a perfect prediction of future VF cannot be achieved currently, owing to the variability in measurements; the VF being predicted itself includes measurement error. The performance of statistical methods is thus limited by data acquisition techniques.

In conclusion, ANSWERS provides an analytical tool in a “landscape of uncertainty” in modeling VF progression. This new technique has the potential to help improve clinical management decisions. In addition, it can be used to help define better and more relevant endpoints for clinical trials, which could help increase the efficiency of trials and decrease their duration and cost.

Acknowledgments

Supported by the National Institute for Health Research, National Health Service, United Kingdom, which gave rise to a Research Fellow Award and Health Technology Assessment Grant supporting this independent research. The views expressed in this publication are those of the authors and not necessarily those of the National Health Service, the National Institute for Health Research, or the Department of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Disclosure: H. Zhu, P.; D. P. Crabb, P.; T. Ho, None; D. F. Garway-Heath, P

References