City Research Online

City St George's, University of London

Citation: Montesano, G., McKendrick, A. M., Turpin, A., Brusini, P., Oddone, F., Fogagnolo, P., Perdicchi, A., Johnson, C. A., Lanzetta, P., Rossetti, L. M., et al (2021). Do Additional Testing Locations Improve the Detection of Macular Perimetric Defects in Glaucoma?. Ophthalmology, 128(12), pp. 1722-1735. doi: 10.1016/j.ophtha.2021.06.012

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/26573/

Link to published version: https://doi.org/10.1016/j.ophtha.2021.06.012

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

Do additional testing locations improve the detection of macular perimetric defects in glaucoma?

Giovanni Montesano[1, 2]; Allison M. McKendrick[3]; Andrew Turpin[4]; Paolo Brusini[5]; Francesco Oddone[6]; Paolo Fogagnolo[7]; Andrea Perdicchi[8]; Chris A. Johnson[9]; Paolo Lanzetta[10]; Luca M. Rossetti[7]; David F. Garway-Heath[2]; David P. Crabb[1]

1. City, University of London, Optometry and Visual Sciences, London, United Kingdom
2. NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
3. University of Melbourne, Department of Optometry and Vision Sciences, Melbourne, Australia
4. University of Melbourne, School of Computing and Information System, Melbourne, Australia
5. Department of Ophthalmology, "Città di Udine" Health Center, Udine, Italy
6. G. B. Bietti Eye Foundation-IRCCS, Rome, Italy
7. University of Milan – ASST Santi Paolo e Carlo, Milan, Italy
8. Ophthalmology Unit, St. Andrea Hospital, NESMOS Department, University of Rome "Sapienza," Rome, Italy
9. Department of Ophthalmology and Visual Sciences, University of Iowa Hospitals and Clinics, Iowa City, Iowa
10. Department of Medical and Biological Sciences, Ophthalmology Unit, University of Udine, Udine, Italy

Corresponding author: David P. Crabb
Email: david.crabb.1@city.ac.uk
Phone number: +44 (0)20 7040 0191
Fax number: None
Address: City, University of London
Northampton Square, London
EC1V 0HB, United Kingdom

Meeting presentation: this work was presented during the ARVO Online Meeting 2021

Financial support(s): The contribution of IRCCS G. B. Bietti Foundation to this work was supported by the Italian Ministry of Health and by Fondazione Roma. The original data collection was sponsored by CenterVue, Padua, Italy. The sponsor or funding organization had no role in the design or conduct of this research, data analysis, interpretation of the data, or preparation, review, or approval of the manuscript..

Key words: macula, glaucoma, perimetry, visual field

Conflict of interest: G.M, A.M.M., A.T., P.B., F.O., P.F., C.A.J., P.L., L.M.R., D.F.G-.H and D.P.C are or were Consultants for CenterVue, Padua, Italy.

Running head: combined perimetric grids to detect macular defects in glaucoma

Abstract

Purpose: to evaluate the ability of additional central testing locations to improve detection of macular visual field (VF) defects in glaucoma.

Design: prospective cross-sectional study.

Participants: 440 healthy people and 499 patients with Glaucomatous Optic Neuropathy (GON) were tested with a fundus tracked perimeter (CMP, CenterVue, Italy) using a 24-2 grid with 12 additional macular locations (24-2+).

Methods: GON was identified based on expert evaluation of optic nerve head photographs and optical coherence tomography scans, independently of the visual field (VF). We defined macular defects as locations with measurements outside the 5% and 2% normative limits on Total Deviation (TD) and Pattern Deviation (PD) maps within the VF central 10 degrees. Classification was based on the total number of affected macular locations (*overall detection*) or on the largest number of affected macular locations connected in a contiguous cluster (*cluster detection*). Criteria based on the number of locations and cluster size were used to obtain equivalent specificity between the 24-2 and the 24-2+, calculated using false detections in the healthy cohort. Partial Areas Under the detection Curve (pAUCs) were also compared at specificities ≥ 95%.

Main Outcome Measure: matched specificity comparison of the ability to detect glaucomatous macular defects between the 24-2 and 24-2+ grids.

Results: at matched specificity, *cluster detection* identified more macular defects with the 24-2+ compared to the 24-2. For example, the mean (95% confidence interval) increase in percentage of detection was 8 (5, 11)% and 10 (7, 13)% for TD-5% and PD-5% maps, respectively, and 5 (2, 7)% and 6 (4, 8)% for the TD-2% and PD-2% maps respectively. There was good agreement between the two grids. The improvement measured by pAUCs was also significant, but generally small. The percentage of eyes with macular defects ranged from 30 to 50%. Test time for the 24-2+ was longer (21% increase). Between 74% and 98% of defects missed by the 24-2 had at least one location with sensitivity < 20 dB,

Conclusions: VF examinations with additional macular locations can modestly improve the detection of macular defects in GON without loss of specificity when appropriate criteria are selected.

Glaucoma is an optic neuropathy with damage to Retinal Ganglion Cells (RGCs) and progressive loss of the visual field (VF) [1, 2]. The use of Standard Automated Perimetry (SAP) to measure the VF is central to management of glaucoma. SAP stimuli are organised in a fixed grid of regularly spaced locations, typically covering the central 30 degrees with locations 6 degrees apart. However, this coarse uniform sampling could be insufficient for the macular region, which contains more than 40% of all RGCs[3]. Testing patterns employed by some perimeters, such as the Octopus (Haag-Streit, Köniz, Switzerland), might be less affected by this problem. However, the widely used 24-2 test program of the Humphrey Field Analyser (HFA, Zeiss Meditec, Dublin, CA) assesses 54 points but only 12 (less than one quarter) are located within the central 10 degrees of the VF. Some reports suggest denser perimetric sampling of the macular region to be more effective in identifying early glaucomatous defects, using the HFA 10-2 test program for example[4-6]. These findings are of great importance both for the correct diagnosis of patients and for the identification of *de novo* sight threatening central defects in patients with confirmed glaucoma. Other very recently reported data suggests there is little evidence that a 10-2 test will reveal central VF loss not already detected by a 24-2[7]. This conflicting evidence is unhelpful for those managing glaucoma. Of course, performing both tests is often not practical in a clinical setting. One alternative could be the use of SAP grids that combine 24-2 with additional testing locations in the macular area, like the 24-2C program for the HFA [8-11].

Studies investigating improvement in the detection of macular defects with additional testing locations (or with a 10-2 grid) all have limitations. For example, increasing the number of tested locations for the central 10 degrees is bound to make finding abnormal locations more likely but will increase false positive detection reducing the specificity of the assessment. West et al. sensibly used matched-specificity criteria to compare the central values of the 24-2 grid and the 10-2 grid[7], yet other investigators failed to do this and reported the detection of macular defects using fixed rules (usually three contiguous affected locations) independently of the grid[4, 11]. This has led to conflicting results. Another source of disagreement lies in the selection bias that might have affected previous reports, since the presence or absence of VF defects in the 24-2 test was used to define the study groups (glaucoma or healthy controls).

We have previously presented the results of the main outcome of a clinical trial[12] comparing the diagnostic precision of the Compass fundus perimeter (CMP, CenterVue, Padova, Italy) with the HFA. The study was done in a large cohort of healthy people and patients with Glaucomatous Optic Neuropathy (GON), defined independently of VF loss; an advantageous design since it minimises selection bias towards the type of VF test. A secondary outcome of the study was to evaluate whether additional macular testing locations improved the detection of macular defects in patients with GON. To this aim, the testing grid used for the CMP was designed to include all the usual locations of a 24-2 grid with 12 additional locations within the central 10 degrees. A prospectively planned study in a large number of people such as this represents an incredible opportunity to address the controversy surrounding the benefit of more accurate testing of the macular region. We adopt matched-specificity criteria

to quantify the improvement in the detection of macular defects offered by the additional testing locations over the conventional 24-2.

Methods

Data collection

This was a cross-sectional case-control study conducted in accordance with the Declaration of Helsinki after written informed consent was acquired from each participant. This study received approval from the ethics committee (International Ethics Committee of Milan, Zone A, 22/07/2015, ref: Prot. n° 0019459) and was registered as a clinical trial (ISRCTN13800424). Participants were recruited (14/09/2015 - 31/07/2017) at eight different study sites[12]. The primary aim of the study was to compare the relative diagnostic performance of CMP and HFA perimeters in their ability to distinguish a large number of people with healthy vision from those with GON. All participants in the study were tested with both the HFA and the CMP, in randomised order. The main results have already been presented elsewhere[12].

A secondary aim of the trial was to quantify the impact on detection of macular defects in the glaucoma cohort brought by adding testing locations to the macular region of the VF. This will be the focus of this work. To this aim, the CMP was equipped with a custom grid containing all the 52 locations of a 24-2 grid (excluding the two blind spot locations) and 12 additional macular locations. This grid, called 24-2+ ("New Grid" in our previous report[12], **Figure 1**), was used for all participants. The additional 12 testing locations were excluded from the previous analysis[12].

Selection criteria have been previously detailed[12] and are available from the published protocol[12] (http://www.isrctn.com/ISRCTN13800424). Participants (consecutive adults) eligible for inclusion had both eyes examined but only one eye per subject was used in the final analysis, chosen randomly if both eyes were eligible. Each subject underwent complete ophthalmological evaluation involving biometry to measure axial length (AL), Spectral Domain Optical Coherence Tomography of the ONH and circumpapillary Retinal Nerve Fibre Layer (cp-RNFL), perimetric demonstration (only for subjects naïve to perimetry); one examination with HFA 24-2 grid to both eyes (data not used for this analysis, except to define inclusion criteria as explained below) and one examination with CMP 24-2+ to both eyes; colour fundus photo with CMP. The reference standard to diagnose GON was clinical evaluation by an expert based on cp-RNFL SD-OCT and/or optic nerve head photography acquired during the protocol examination. Experts were required to evaluate OCT scans for RNFL thinning and ONH colour pictures for disc cupping, rim narrowing, focal notching or peripapillary haemorrhages. The rationale for this reference standard was to avoid any classification based on VF testing that could bias the results towards either perimeter. For the current work, this also reduced the bias towards the 24-2 grid, since participants with GON were included regardless of their VF.

Details of the visual field examination

The CMP uses continuous infrared imaging of the retina designed to track and compensate for

eye movements during the test[12-14]. Threshold acquisition used a Bayesian testing strategy, an adaptation of the Zippy Estimation by Sequential Testing (ZEST) [15, 16]. No near correction was needed because the CMP is equipped with auto-focusing. All locations in the 24-2+ were tested independently in randomised order. Therefore, unlike the HFA SITA (Swedish Interactive Thresholding Algorithm) test strategy, the CMP test does not use spatial correlations between neighbouring locations or specific spatial patterns[15]; this was a noteworthy convenience for our analyses because it allows the evaluation of the isolated contribution of additional macular locations without this confounding effect, unavoidable with SITA strategies. Differently from the 24-2C adopted by the HFA, the 24-2+ was designed as a general-purpose pattern and not specifically to detect macular defects from glaucoma[11]. VF examinations were considered unreliable if the frequency of false positive errors was > 18% or the Blind Spot response frequency was > 25%. Same criteria were applied to the HFA examination[12]. If either the HFA or the CMP VF was deemed unreliable, the eye was excluded from the analysis and was therefore not present in the final dataset.

Statistical analysis

Only locations within the central 10 degrees from fixation were considered (**Figure 1**). The 24-2+ contained all central 24-2 locations (N = 12) plus 12 additional macular locations (total N = 24).

Calculation of probability maps

Data from the healthy cohort collected in this study represent the only available normative database for the CMP and were therefore used to define normative values. Total Deviation (TD) and Pattern Deviation (PD) maps were calculated as previously described[12]. In short, a linear regression for each test location was used to model the normal sensitivity decay with age using the healthy cohort. The TD is simply the difference between the observed sensitivity and the expected age corrected value at each location. PD is then calculated by subtracting the General Height (GH) of the field from the TD map. Following the definition of the Imaging and Perimetry Society (IPS)[17], the GH was the value corresponding to the 7th highest location TD map. This was calculated separately for the 24-2+ and the 24-2 grid, considering all locations (64 and 52 respectively). Normative limits were calculated using quantile regression of TD and PD values to account for age-related changes in variability as previously described[12]. For this analysis we calculated which sensitivity values in each VF were within or outside the 5% or 2% normative limits for each map. For healthy subjects, the normative limits were calculated using a leave-one-out procedure, so that each healthy subject was excluded from the normative database when their maps were calculated. The whole normative dataset was used to calculate the maps for patients with GON. Mean Deviation (MD) reported in this study is the mean of total deviation values. MD was only used as a descriptive measure and was calculated for the central 24-2 locations, the central 24-2+ locations and the whole 24-2 grid.

Calculation of the matched specificity criteria

Macular defects were identified using the probability maps. Hence, we obtained a defect identification for the TD and the PD maps, each one with two possible probability thresholds (5% or 2% normative limits). The macular defect for each probability map was identified when a certain number of locations fell outside their normative limit (*abnormal locations*). We

used two defect definition strategies:

1) **Overall detection**: the central locations from either the 24-2 grid or the 24-2+ were treated as a whole. Identification of the defect depended only on the total number of locations outside the normative limits in each grid. This analysis is similar to what has been proposed by West et al.[7].
2) **Cluster detection**: the central locations outside the normative limits were pooled together only if belonging to a cluster of contiguous points. The number used for detection is then the largest cluster size identified in the central VF. Defining contiguous locations is trivial in a regular grid such as the 24-2 but poses a challenge for grids with irregular spacing, such as the 24-2+. We defined a neighbourhood system for the 24-2+ based on a nearest-neighbour triangulation (**Figure 1**). The locations were not allowed to connect across the horizontal midline. Hence, the maximum size of one cluster was six locations for the 24-2 and 12 locations for the 24-2+ (i.e. half the total number of central locations). The details of the computation are reported in the **Appendix**. This method resembles a typical definition of glaucomatous VF defects, used for example by De Moraes et al.[4] and Phu et al.[11] in a similar analysis, but allows for a flexible selection of the cut-off for the cluster size to match specificity between the two grids (see below).

The objective of this study was to compare the detection rate of macular defects between 24-2+ and 24-2 at the same specificity. To achieve this, the number of locations used for detection needed to be different for the two grids. Specificity is defined based on the False Positive Rate (FPR) as S = (1 – FPR). We considered all locations outside the normative limits detected in the healthy cohort as false positives. From this assumption, we could calculate the specificity of increasing threshold criteria on the number of abnormal locations for each probability map (**Figure 2**). For each map, we selected the two threshold criteria (one for each grid) that yielded the same specificity. Given the discrete nature of the criterion, however, an exact match in specificity is unlikely. For our analysis, we selected the smallest pair of threshold criteria that provided a specificity ≥ 95% and a difference in specificity between the two grids ≤ 0.5%, assuming that any smaller difference would be clinically irrelevant. This calculation was performed for both grids with all probability maps and for both detection strategies (*overall detection* and *cluster detection*, **Figure 2**). This was meant to provide a practical evaluation of the performance with criteria that could actually be applied. We also performed a comparison of the partial Area Under the detection Curve (pAUC) limited to specificities ≥ 95%, to provide an analysis that was not affected by these practical limitations. Note that these pAUCs are not meant to evaluate the diagnostic precision but only the detection of macular defects.

Analysis of the detection rate

The matched-specificity criteria were then applied to the GON cohort to detect macular defects. The two grids were compared for all maps in terms of detection rate. Confidence Intervals (CIs) and p-values for the difference in detection rate were calculated via bootstrap (N = 5000 samples). Note that the bootstrap procedure was performed by resampling only eyes with GON, so that specificity was held constant at each draw. The agreement between the

two grids was also calculated and represented using Venn diagrams. The pAUCs were calculated by interpolating between different threshold criteria and normalising their value over 0.05, the maximum AUC achievable with the selected specificity range (95% - 100%). The pAUCs were compared using the same bootstrap procedure explained above. We also explored the spatial distribution of locations outside the normative limits by calculating their frequency for each location of the 24-2+ within the GON cohort. We compared the four quadrants (**Figure 1**) in terms of probability of finding an abnormal location, using a logistic regression. For this latter comparison, the p-values were corrected for multiple testing using the Bonferroni-Holm method and considered significant when < 0.05. Finally, we reported the distribution of the defect depth missed by either grid to assess their clinical relevance. For this descriptive analysis, we quantified the distribution of TD values for all the locations identified as abnormal by each probability map that met the criteria to detect a macular defect in one grid but not the other (i.e. the cases in which the 24-2 and the 24-2+ were in disagreement). We also defined as "deep" defects all locations with TD ≤ -20 dB and quantified them for each grid.

All statistical analyses were performed in R (R Foundation for Statistical Computing, Vienna, Austria).

Comparing test times

The duration of the test with the 24-2+ was recorded by the device. We obtained an accurate estimate of the time taken to test the additional macular locations by analysing the recorded history of each test extracted from the device, which reports the number of stimulus presentations for each location. With this, we calculated the average time for each presentation which was then used to estimate the time needed to test the 24-2 locations only. (Note this is different from our first report[12], where the time estimate for the 24-2 was simply derived as a proportion of the number of tested locations, not of the presentations themselves.)

Results

Description of the sample

Of the 1249 people screened for eligibility, 177 did not satisfy the inclusion criteria and 59 did not complete the examination protocol. Finally, 70 subjects were excluded because they had at least one unreliable VF test and four healthy subjects were erroneously tested only with the 24-2 grid. Therefore, 440 healthy subjects and 499 patients with GON were included in the final analysis. Descriptive statistics are reported in **Table 1**. As previously reported[12], despite a significant gap in the average age between the healthy and GON cohort, the age range for the healthy subjects was large (18 - 84 years) and allowed for reliable estimates of the normative limits for all GON patients. The range of VF damage in the GON cohort was wide (range of MD for the whole 24-2 grid: -27.85, +2.89 dB). MD for the central VF was very slightly higher when measured with the 24-2+ compared to the 24-2 (**Table 1**). This difference was statistically significant for people with GON ($p < 0.001$, paired Mann-Whitney test) but not for the healthy cohort ($p = 0.99$).

Specificity analysis

The 440 subjects in the healthy cohort were used to calculate matched-specificity criteria for the detection of macular defects. **Figure 2** shows how increasing the number of abnormal locations required to identify a defect progressively increases the specificity. **Table 2** reports the selected criteria and their respective specificity values for each map and detection strategy, according to the overall and *cluster detection*.

Detection of macular defects

The 499 eyes with GON were used to test the performance of the two grids in detecting macular defects with matched-specificity criteria. The improvement with the 24-2+ was variable for the *overall detection*, with a significant improvement detected only for the 2% probability maps (**Table 3**). The relative improvement was larger and significant for all maps when contiguous *cluster detection* was employed. In this case, the largest improvement was for the TD 5% map. The pAUC analysis partially replicated the results of the matched specificity analysis, but the differences were in general much smaller (**Table 4**). All differences were still significant when *cluster detection* was employed except for the TD 2% map ($p = 0.068$). The *overall detection* was no longer statistically significant with the TD 2% map but reached significance with the TD 5%. **Figure 3** shows the detection rate curves for all methods and all maps, marking the operating points used for the matched specificity analysis. The curves show that most of the difference between the two grids was obtained at very high specificities.

Agreement between the two grids was good but not perfect, as shown by the Venn diagrams (numerical values are reported in **Figure 4**). Interestingly, some macular defects were identified by the 24-2 and not by the 24-2+. This was either a consequence of healthy locations in the 24-2+ breaking up contiguous clusters or because the number of affected locations was sufficient to meet the detection criteria with the 24-2 but not the 24-2+.

Two illustrative examples of the results obtained with the cluster detection with the selected same-specificity criteria are reported in **Figure 5**. The distribution of the depth of the defects missed by either grid for the cases of disagreement is reported in **Figure 6**, including the number of locations with a deep defect. Many of the locations missed by the 24-2 had a deep defect. It should be noted, however, that these locations came from a relatively small number of eyes (largest N = 54, see **Figure 4**). The majority of the eyes that had a defect missed by the 24-2 but detected by the 24-2+ also had at least one location with a deep defect. This percentage ranged between 74% and 98% when all 24-2+ locations were considered, depending on the detection method and map, and between 47% and 90% using only locations in common between the two grids. More details are available as **supplementary material**.

Spatial distribution of abnormal locations

The frequency of locations outside the normative limits in the GON cohort was higher in the superior-temporal quadrant (**Figure 7**) for all probability maps. All pair-wise differences between quadrants were significant except for the comparison between Quadrant 2 and 4 with the TD at 5% probability ($p = 0.329$). All other p-values were < 0.001 except for the comparison between Quadrant 2 and 4 with the TD at 2% probability ($p = 0.002$). **Figure 7** reports the predicted frequency and 95% CIs for each quadrant from the logistic regression.

Discussion

We evaluated the effect of additional perimetric testing locations on detecting macular defects using prospectively collected data in a large cohort of patients with GON. Detection was based on the number of abnormal locations in TD or PD probability maps, detected with matched-specificity criteria derived from a large group of healthy eyes. We compared two different detection strategies, where the abnormal locations were either pooled independently of their spatial pattern (*overall detection*) or only if belonging to clusters of contiguous locations (*cluster detection*). Improvement in the detection rate provided by the additional macular locations was variable, ranging from an average of less than 1% to almost 10% depending on the detection method used. All analyses were conducted at a specificity level that would be clinically useful (≥ 95%). When taken together, the size of these experimental effects suggest perimetric grids with additional macular locations can modestly improve sensitivity of detecting macular defects in glaucomatous optic neuropathy without loss of specificity, provided appropriate criteria are applied to define a defect. Note, this is different from comparing the diagnostic precision of the two grids. In fact, the 24-2 might be able to correctly identify glaucoma cases based on peripheral damage while failing to detect central defects. The pAUC analysis partially confirmed this view, showing a consistent improvement with the *cluster detection*, less so with the *overall detection*. It should be noted that the improvement in interpolated pAUC with the 24-2+ grid was modest for both methods, indicating very similar overall performance (**Figure 3** and **Table 4**). However, it should be stressed that this is not reflective of how the two grids would be used in practice, i.e. with discrete cut-offs on the number of affected locations to identify a defect. This is why it was important to define a clinically acceptable difference in matched specificity for our practical evaluation. For example, many comparisons yielded a difference in detection rate between 8% and 10% (**Table 3**), which we believe is meaningful accepting a clinically insignificant change in specificity ≤ 0.5%.

Our findings are in partial agreement with those from previous reports. For example, De Moraes et al.[4] showed that more central defects were detected by the 10-2 when compared to 24-2 grid in patients with ocular hypertension (OHT), glaucoma or suspected glaucoma. The largest number of missed central defects in their dataset was recorded within the OHT and glaucoma suspect cohort. In these groups, they reported an improved detection of approximately 12% with the 10-2, which is generally higher than we found in our dataset. The results by De Moraes et al.[4] could, however, be biased by their definition for OHT and suspects requiring a normal 24-2 VF as an inclusion criterion. We had no such bias in our study design because the inclusion criteria did not involve VF loss. Moreover, the analysis employed by De Moraes et al.[4] applied the same detection criteria to both the 24-2 and the 10-2 grid failing to account for the loss of specificity introduced by the 10-2, which tests a higher number of locations in a smaller area of the VF. The reported percentage of false positives with the 10-2 grid was 4.6% in their healthy dataset, but they could not assess specificity for the 24-2 grid because all healthy subjects needed to have a normal 24-2 test result. Besides, in contrast to our analysis, De Moraes et al.[4] compared the ability of identifying any defect in the whole 24-2 grid versus the 10-2 grid; the percentage of missed central defects was then inferred from the

mismatch in the results from the two grids. This would not identify any occurrence when the 24-2 would identify glaucoma due to the presence of a peripheral defect but fail to highlight macular damage (**Figure 5, A**). Instead, we focused our analysis only on the central region, with the specific intent of evaluating the improvement in the detection of macular defects. We think this makes for a fairer comparison. Of course, other reasons could explain the differences between our results and those presented by De Moraes et al.[4]. For example, our additional macular locations did not sample the central field at the same density as a 10-2 grid, possibly underestimating the number of central defects. Our analysis, however, directly addresses the clinical question pertaining to the usefulness of combined VF grids, which are more likely to be an acceptable compromise for everyday clinical practice.

More recently, West et al.[7] performed a matched-specificity analysis with a method identical to our *overall detection* to compare the 24-2 and the 10-2 grids. Their results are in close agreement with ours, concluding for minimal improvement in the detection of macular defects with the 10-2 grid when no spatial patterns are considered. Their analysis, however, was performed on a much smaller sample (97 eyes with glaucoma and 65 controls) and suffered from another bias, since glaucoma patients were required to have an early damage on the 24-2 VF test for inclusion. In contrast, our recruitment scheme was originally designed to compare the diagnostic ability of two perimeters and purposely avoided any inclusion criterion based on VF tests, be it with a 24-2 or 10-2 grid[12].

Another novel aspect of our analysis was the use of an objective rule to detect contiguous abnormal locations that was independent of the specific arrangement of the testing grid (*cluster detection*). Differently from static criteria used previously[4, 11], our method allowed the selection of matched-specificity thresholds in the same way as with the *overall detection*. This is important because it is unusual for clinicians to consider abnormal locations in isolation. Simply counting the number of abnormal values would likely deviate from clinical practice. However, empirical combination rules are often applied to different grids without accounting for changes in specificity[4, 11], which was instead central in our analysis. Interestingly, our *cluster detection* provided the largest improvement with the 24-2+, which reached significance in all probability maps; this suggests that additional testing locations are beneficial in defining the spatial patterns of glaucoma damage, even when matched-specificity criteria are applied. This may explain the differences observed between the results by De Moraes et al. [4] and West et al.[7], since spatial continuity of abnormal locations were not considered by the latter. Wu et al.[18] also failed to show any significant improvement in the detection of macular defects between 24-2 and 10-2 using only global metrics such as pattern standard deviation.

Phu et al.[11] recently published a report comparing the 24-2 and the new 24-2C grid provided by the HFA. The 24-2C has 10 additional testing locations derived from the 10-2 grid designed to optimise the detection of macular glaucoma defects. Their main results showed no significant differences between the two grids in detecting macular defects. Their study did not select glaucoma patients based on VF criteria, hence removing the selection bias, but had some technical limitations in the testing procedure. First, the HFA only allows the 24-2C test with the SITA-Faster strategy. Moreover, all SITA strategies employ spatial correlations

between neighbouring locations[15] to improve the speed of the test. This prevents an independent evaluation of the isolated central locations of the 24-2 grid and of the contribution of the additional locations in the 24-2C. Each location was instead tested independently in the ZEST strategy implemented in the CMP. Although this might not be optimal for practical perimetry, it provides a convenient and ideal condition for our experiment. In fact, limitations imposed by the spatial correlations employed by SITA strategies are common to all the aforementioned studies[4, 7, 11] and would also greatly affect any analysis aimed at isolating the central locations from a larger grid such as the 24-2[7, 11]. This therefore represents another strength of our study that allowed us to make additions to the literature that would not be possible with conventional SITA methods.

Other important differences with the work by Phu et al.[11] pertain to the definition of VF defects. They also used a cluster based approach, but specificity was not calculated in a healthy cohort and was not matched between the two grids. Moreover, they did not define the neighbourhood used in the 24-2C to identify contiguous locations. Importantly, their main results relied on the definition of "additive" detection, where the contribution of additional locations was only evaluated in terms of improvement over the 24-2. This would fail to highlight all the instances where the additional normal locations would break up contiguous clusters found in the 24-2 (**Figure 5, B**) and hence reducing the detection rate. They reported only two such cases in their dataset. However, they also separately reported, in the same paper, the detection of defects based only on the central locations of the two grids with different cluster sizes, where this effect more clearly shows, i.e. some macular defects were only detected with the 24-2 and not with the 24-2C. This analysis is similar to ours and the results largely agree, showing that the vast majority of the defects were detected by both grids, with some improvement in detection with the additional macular locations. In their case, these differences were not significant, but their sample size (N = 64) was much smaller than ours. It is also important to notice that in our specific case, the missed identification of macular defects with additional central locations could also be due to the different criteria (i.e. more abnormal locations required) used for the 24-2+ to maintain the same specificity as the 24-2. It is also worth noting that the many of the affected locations not identified as macular defects by the 24-2 showed a deep defect (**Figure 6**). This might be of clinical relevance despite the small number of eyes in which there was disagreement between the two grids. However, in practice, it is unusual to consider probability thresholds in isolation and the depth of the defect is often taken into account. In fact, many of these deep defects were also detected by locations in common with the 24-2 grid, indicating that accounting for the magnitude of loss might improve the detection with the 24-2 despite its lower spatial resolution. Future work should focus on the development and validation of criteria that combine different probability levels from TD and PD maps. These exist for the 24-2[19] but have not yet been formalised for other grids.

Our study has other important strengths. Due to our recruiting strategy, no stratification was planned for VF damage. Despite this, the range of overall VF loss was large. However, most of the eyes with GON had early to moderate central damage (**Table 1**), possibly as a consequence of the lower bounds imposed on visual acuity for recruitment. This constitutes a

very convenient setting to evaluate the performance of the two grids in detecting macular defects which in some cases could be subtle. Moreover, all locations were tested within the same session and in no preferred order within the test. In all previous studies subjects were tested with the 24-2 or the alternative grid (either the 10-2 or the 24-2C) in separate sessions[4, 7, 11]. Moreover, we employed a fundus tracked perimeter that compensates for eye movements. This can be particularly valuable when closely spaced locations are tested near fixation, removing the effect of eye movements on the spatial resolution of the grid[20, 21]. Another important aspect is that the GH was chosen as the 7th highest value in the TD and could therefore be different for the two grids. We believe this is more reflective of how the calculation would be performed in clinical practice, where all the available data-points would be used. Moreover, this is also in line with previous studies involving the 10-2 test, for which the GH would be calculated in a similar way. Also note that, differently from the 24-2, this value for the GH is not exactly the 85th percentile for the 24-2+[17].

Our study has some limitations. Our main analysis only focussed on the detection improvement provided by the central locations. This could be a limitation for the *cluster detection* method, since it would force the clusters to be fully contained within the central 10 degrees. For example, this approach may miss any macular defect identified by isolated central locations that connect to larger clusters outside the central 10 degrees. This choice was specifically meant to increase the specificity for macular defects which would likely lie on nerve fibre bundles fully contained within the central 10 degrees[22]. In response to this potential limitation the **supplementary material** reports a secondary analysis allowing clusters to extend to the whole grid and detecting a macular defect if any of these clusters invaded the central 10 degrees. As expected, more macular defects were identified with both grids and the results leaned more towards an equivalence between the 24-2 and 24-2+. However, we reiterate that such a liberal approach to cluster connectivity might not be specific for anatomically plausible macular defects and should be interpreted with caution. It should also be noted that the 24-2+ grid was not specifically designed to reflect the topology of macular defects from glaucoma, unlike the 24-2C. However, while patterns tailored to glaucomatous defects might be desirable for this specific application, they make the test less generalisable, possibly compromising the detection of other sight threatening diseases such as age-related macular degeneration. Finally, our healthy cohort represented the only available normative database for CMP. We accounted for this by using a leave-one-out approach in the calculation of the normative limits for the healthy cohort. One limitation of our definition of the healthy cohort is that only one grader evaluated the structural data to exclude the presence of GON. However, given our additional constraints on the intraocular pressure and the relatively low prevalence of GON in the general population, misclassification of eyes with GON as healthy is unlikely. The opposite misclassification (healthy as GON) would only dilute the percentage of detected macular defects but would not compromise the same specificity comparison between the two grids. Future evaluations might also benefit from more detailed characterisation of structural damage with dense macular OCT maps, not available for this cohort. The lack of structural confirmation also prevented us from assessing the actual prevalence of macular damage in eyes with GON. Hence, we could not report the

sensitivity of the methods but only the relative difference in detection rate (proportional to sensitivity).

A definitive judgment on the usefulness of increased spatial sampling of the central VF in glaucoma patients is hard to come by. With our results, we showed a consistently significant increase in the detection rate of macular defects when spatially connected clusters are considered. It would be therefore inappropriate to dismiss any improvement from denser macular grids as a simple consequence of a loss in specificity[7]. Indeed, a previous report by Grillo et al.[5] showed that macular defects identified by the 10-2 grid and confirmed by independent structural measurements, such as dense macular OCT scans, were missed by the 24-2 grid in 52% of the cases. Previous analysis on structural tests showed that many of the missed bundle defects were within the central 4 degrees, not tested by the 24-2[22]. In our dataset, the four testing points within 4 degrees from fixation in the 24-2+ identified at least one damaged location in 56%, 39%, 35% and 28% of the subjects with GON for the TD-5%, TD-2%, PD-5% and PD-2% maps respectively. This speaks to another important point of discussion: similar detection power of central damage does not always correspond to equivalent characterization of the defect itself. This could have important effects on the definition of the spatial spread of glaucoma damage and on the accuracy of structure-function relationship.

Improvement in detection seen in our results came at the cost of an estimated 21% average increase in test time (**Table 1**). Whether this is clinically acceptable is outside the scope of the questions we have asked in this study. Integration of structural metrics to guide the acquisition of additional data from the macular region where damage is expected could be a solution to this issue and could be the subject of future work. Such a strategy has proven successful in reducing the test time and improving the overall efficiency of the test, [13, 23-25] but might also bias the estimates of VF sensitivity. This would not only reduce testing time but also improve specificity. Another approach is to incorporate prior population based knowledge about what locations are more likely to detect a macular defect. This is the strategy chosen for the 24-2C grid implemented in the newest HFA perimeters[8-11], which includes only few locations from the 10-2 grid. Our data indicate a spatial distribution of macular defects in the GON cohort that largely reflects the expectation from previous knowledge[22]. However, many abnormal locations were also distributed across the whole central VF (**Figure 7**) and their detection would certainly be compromised by such an approach. Additional locations might also be added dynamically based on the spatial features of VF defects, providing customisable testing grids[26, 27].

In conclusion, VF tests with additional macular locations can modestly improve detection of macular defects in glaucoma patients with minimal to no loss in specificity. However, the overall difference between the two grids is small. A more tailored approach, possibly based on structural evaluations, could help select people who are more likely to benefit from more precise macular testing.

References

1. Garway-Heath DF, Hitchings RA. Quantitative evaluation of the optic nerve head in early glaucoma. Br J Ophthalmol 1998;82(4):352-61.
2. Weinreb RN, Aung T, Medeiros FA. The pathophysiology and treatment of glaucoma: a review. JAMA 2014;311(18):1901-11.
3. Curcio CA, Allen KA. Topography of ganglion cells in human retina. J Comp Neurol 1990;300(1):5-25.
4. De Moraes CG, Hood DC, Thenappan A, et al. 24-2 Visual Fields Miss Central Defects Shown on 10-2 Tests in Glaucoma Suspects, Ocular Hypertensives, and Early Glaucoma. Ophthalmology 2017;124(10):1449-56.
5. Grillo LM, Wang DL, Ramachandran R, et al. The 24-2 Visual Field Test Misses Central Macular Damage Confirmed by the 10-2 Visual Field Test and Optical Coherence Tomography. Transl Vis Sci Technol 2016;5(2):15.
6. Traynis I, De Moraes CG, Raza AS, et al. Prevalence and nature of early glaucomatous defects in the central 10 degrees of the visual field. JAMA Ophthalmol 2014;132(3):291-7.
7. West ME, Sharpe GP, Hutchison DM, et al. Utility of 10-2 Visual Field Testing in Glaucoma Patients with Early 24-2 Visual Field Loss. Ophthalmology 2020.
8. Lee GC, Monhart M, Callan T, et al. Performance of a modified 24-2 test pattern using SITA Faster. Investigative Ophthalmology & Visual Science 2018;59(9):6032-.
9. Lee GC, Yu S, Callan T, et al. Diagnostic Efficacy of 24-2 and 24-2C SITA Faster Global Summary Indices. Investigative Ophthalmology & Visual Science 2019;60(9):2455-.
10. Yu S, Lee GC, Callan T, et al. Comparison of SITA Faster 24-2C test times to legacy SITA tests. Investigative Ophthalmology & Visual Science 2019;60(9):2454-.
11. Phu J, Kalloniatis M. Ability of 24-2C and 24-2 grids in identifying central visual field defects and structure-function concordance in glaucoma and suspects. Am J Ophthalmol 2020.
12. Montesano G, Bryan SR, Crabb DP, et al. A Comparison between the Compass Fundus Perimeter and the Humphrey Field Analyzer. Ophthalmology 2019;126(2):242-51.
13. Montesano G, Rossetti LM, Allegrini D, et al. Improving Visual Field Examination of the Macula Using Structural Information. Transl Vis Sci Technol 2018;7(6):36.
14. Rossetti L, Digiuni M, Rosso A, et al. Compass: clinical evaluation of a new instrument for the diagnosis of glaucoma. PLoS One 2015;10(3):e0122157.
15. Turpin A, McKendrick AM, Johnson CA, Vingrys AJ. Properties of perimetric threshold estimates from full threshold, ZEST, and SITA-like strategies, as determined by computer simulation. Invest Ophthalmol Vis Sci 2003;44(11):4787-95.
16. King-Smith PE, Grigsby SS, Vingrys AJ, et al. Efficient and unbiased modifications of the QUEST threshold method: theory, simulations, experimental evaluation and practical implementation. Vision Res 1994;34(7):885-912.
17. Sample PA, Dannheim F, Artes PH, et al. Imaging and Perimetry Society standards and guidelines. Optom Vis Sci 2011;88(1):4-7.
18. Wu Z, Medeiros FA, Weinreb RN, Zangwill LM. Performance of the 10-2 and 24-2 Visual Field Tests for Detecting Central Visual Field Abnormalities in Glaucoma. Am J Ophthalmol 2018;196:10-7.
19. Hodapp E, Parrish RK, Anderson DR. Clinical decisions in glaucoma: Mosby Incorporated, 1993.
20. Matsuura M, Murata H, Fujino Y, et al. Evaluating the Usefulness of MP-3 Microperimetry in Glaucoma Patients. Am J Ophthalmol 2018;187:1-9.
21. Wyatt HJ, Dul MW, Swanson WH. Variability of visual field measurements is correlated with the gradient of visual sensitivity. Vision Res 2007;47(7):925-36.
22. Hood DC, Raza AS, de Moraes CG, et al. Glaucomatous damage of the macula. Prog Retin Eye Res 2013;32:1-21.

23. Ballae Ganeshrao S, Turpin A, Denniss J, McKendrick AM. Enhancing Structure-Function Correlations in Glaucoma with Customized Spatial Mapping. Ophthalmology 2015;122(8):1695-705.
24. Ballae Ganeshrao S, Turpin A, McKendrick AM. Sampling the Visual Field Based on Individual Retinal Nerve Fiber Layer Thickness Profile. Invest Ophthalmol Vis Sci 2018;59(2):1066-74.
25. Ehrlich AC, Raza AS, Ritch R, Hood DC. Modifying the Conventional Visual Field Test Pattern to Improve the Detection of Early Glaucomatous Defects in the Central 10 degrees. Transl Vis Sci Technol 2014;3(6):6.
26. Turpin A, Morgan WH, McKendrick AM. Improving Spatial Resolution and Test Times of Visual Field Testing Using ARREST. Translational Vision Science & Technology 2018;7(5):35-.
27. Muthusamy V, Turpin A, Walland MJ, et al. Increasing the Spatial Resolution of Visual Field Tests Without Increasing Test Duration: An Evaluation of ARREST. Translational Vision Science & Technology 2020;9(13):24-.

Figure legends

Figure 1. Schematic depiction of the 24-2 (red circles) and the 24-2+ (black dots), excluding the blind spot locations. The additional macular locations were symmetric for all four quadrants. The coordinates for Quadrant 2 (X; Y) were [(1.5; 1.5), (1.5; 6), (6; 6)] degrees. The area within 10 degrees from fixation was used in this analysis and is highlighted. On the right, neighbourhood relationships for the central region of the two grids (see Methods).

Figure 2. Changes in specificity according to different threshold criteria. The criterion represents the minimum number of locations outside the normative limits (at either 5% or 2% probability) used to define a macular defect for the TD and PD maps. In the overall selection (left) the central field is considered as a whole (maximum N is 12 for the 24-2 and 24 for the 24-2+). In the cluster selection (right) locations are pooled only if spatially contiguous and the size of the largest cluster is then used (maximum N is 6 for the 24-2 and 12 for the 24-2+). Black circles highlight the matched-specificity criteria selected for this analysis. These calculations were performed with data from the healthy cohort (440 eyes), using a leave-one-out approach. FPR = False Positive Rate.

Figure 3. Curves representing the detection rate at different specificity levels for the different maps and criteria applied to the 24-2 and 24-2+ grids. The operating points for the matched specificity comparisons are marked by a target.

Figure 4. Venn diagrams representing the agreement between 24-2 and 24-2+ in detecting macular defects with different detection strategies. The area of the squares is proportional to the percentage of eyes for which a macular defect was detected. The overlapping area represents the percentage of defects detected by both grids. Percentages (number of eyes) are reported separately for each grid in colour-coded labels and in black for simultaneous detections.

Figure 5. Panel (A) shows an example of a macular defect identified by the 24-2+ grid (outlined in red) but missed by the 24-2 according to the *cluster detection*. Notice how the whole 24-2 grid would still detect a peripheral defect (superior-temporal) and correctly identify this as a glaucoma case without detecting the central damage. Panel (B) shows an opposite example, where the 24-2+ grid misses a defect identified by the 24-2 because the cluster is interrupted by a healthy location. Both examples use the criteria for the TD-5% map. The criterion refers to the minimum size of the largest cluster considered to identify a defect. The central locations considered for the analysis are enclosed by the blue dashed outline. These maps were built using only the $P < 5\%$ and $P < 2\%$ symbols, reflecting the two maps used for this analysis.

Figure 6. Histograms of the distribution of the depth of the defect for all the abnormal locations missed by one grid but detected by the other, i.e. the cases in which the 24-2 and the 24-2+ disagreed. The distribution of the deep defects (TD ≤ -20 dB) is outlined in black. The text-boxes report the total number of missed deep defects/the total number of missed affected locations. TD = Total Deviation; PD = Pattern Deviation.

Figure 7. Spatial distribution of locations outside the 5% and 2% normative limits in the cohort of patients with GON. The size of the circles is proportional to the frequency of values outside the normative limits for each tested location. Estimated probability of finding abnormal values [95% CIs] are also reported for each quadrant, as estimated by logistic regression. GON = Glaucomatous Optic Neuropathy; Q = Quadrant.