A comparison between the Compass fundus perimeter and the Humphrey Field Analyzer

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

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

Corresponding author: David P. Crabb

e-mail: David.Crabb.1@city.ac.uk

Meeting presentation: part of the content of this manuscript has been accepted for oral presentation at the ARVO Meeting 2018.

Financial Support: CenterVue, Padua, Italy sponsored the study. The sponsor or funding organization had no role neither in the design or conduct of this research nor in data analysis, interpretation of the data, preparation, review or approval of the manuscript. The contribution of IRCCS G.B.Bietti Foundation to this work was supported by the Italian Ministry of Health and by Fondazione Roma

Conflict of Interest: Paolo Fogagnolo, Francesco Oddone, Allison M. McKendrick, Andrew Turpin, Paolo Lanzetta, Chris A. Johnson, David F. Garway-Heath, David P. Crabb, Paolo Brusini, MD and Luca M. Rossetti are Consultants for CenterVue, Padua, Italy.

This article contains additional online-only material. The following should appear online-only: Supplementary Figure 1 and 2.

Running head: Comparison of Compass and Humphrey Field Analyzer

Abstract

Purpose: To evaluate relative diagnostic precision and test retest variability of two devices, the Compass (CMP, CenterVue, Italy) fundus perimeter and the Humphrey Field Analyzer (HFA, Zeiss, Dublin), in detecting glaucomatous optic neuropathy (GON).

Design: Multicentre cross-sectional case–control study.

Subjects: We sequentially enrolled 499 glaucoma patients and 444 normal subjects to analyse relative precision. A separate group of 44 glaucoma patients and 54 normal subjects was analysed to assess test – retest variability.

Methods: One eye of the recruited subjects was tested with the index tests: HFA (SITA Standard strategy) and CMP (ZEST strategy) with a 24-2 grid. The reference test for GON was specialist evaluation of fundus photographs or OCT, independent of the visual field. For both devices, linear regression was used to calculate the sensitivity decrease with age in the normal group to compute pointwise Total Deviation (TD) values and Mean Deviation (MD). We derived 5% and 1% pointwise normative limits. MD and the total number of TD values below 5% (TD 5%) or 1% (TD 1%) limits per field were used as classifiers.

Main Outcome Measures: We used partial Receiver Operating Characteristic (ROC) curves and partial Area Under the Curve (pAUC) to compare the diagnostic precision of the devices. Pointwise Mean Absolute Deviation (MAD) and Bland Altman plots for the mean sensitivity (MS) were computed to assess test- retest variability.

Results: Retinal sensitivity was generally lower with CMP, with an average mean difference of 1.85 ± 0.06 dB (Mean ± Standard Error, $p < 0.001$) in healthy subjects and 1.46 ± 0.05 dB (Mean ± Standard Error, $p < 0.001$) in patients with glaucoma. Both devices showed similar discriminative power. The MD metric had marginally better discrimination with CMP (pAUC difference ± Standard Error, 0.019 ± 0.009, $p = 0.035$). The 95% limits of agreement for the MS were reduced by 13% in CMP compared to HFA in glaucoma subjects, and by 49% in normal subjects. MAD was very similar, with no significant differences.

Conclusions: Relative diagnostic precision of the two devices is equivalent. Test-retest variability of mean sensitivity for CMP was better than for HFA.

Standard Automated Perimetry (SAP) is used to assess the visual field (VF) and is a key examination for detection, diagnosis and follow up in glaucoma. SAP typically uses stimuli of varying intensities to assess the differential light sensitivity at static locations across the VF. The examination demands strong cooperation[1] from test subjects; they are required to maintain central fixation and respond timely and accurately to the presented stimuli. Fixation instability might be an unavoidable feature of a person's vision, especially with advanced age and macular damage[2]. One proposed solution has been to incorporate live fundus tracking in the macular perimetric exam to compensate for eye movements in unstable fixation [3].

Recently, a novel instrument, the COMPASS fundus perimeter (CMP, CenterVue, Padua, Italy), has successfully employed a live fundus tracking technology for wide field (30 degrees) VF assessment [4, 5] yielding results comparable with the Humphrey Field Analyzer (HFA) in a preliminary study [4]. The CMP captures images of the fundus during the perimetric examination using a scanning laser ophthalmoscope. This design feature is intended to afford compensation for eye movements when the stimuli are presented at predetermined test locations. Moreover, the instrument provides colour images of the fundus and optic nerve that can be mapped to the final perimetric results potentially providing clinically useful information about structure and function in one assessment.

Diagnostic accuracy studies are used to certify new examinations before they are brought into clinical practice. The CMP has not yet been scrutinised in this way and this is the main purpose of our investigation. Studies investigating relative diagnostic accuracy are at risk of bias due to shortcomings in design and conduct. For this reason, we designed our study to follow appropriate guidelines on this specific aim [6, 7].

Our cross-sectional and multicentre study was designed to evaluate and compare two index tests, namely the CMP and the HFA. One objective was to evaluate and compare test – retest variability of the two index tests in healthy subjects and patients with glaucomatous optic neuropathy (GON). We hypothesised that the CMP could obtain a 20% reduction in test-retest variability on the measurement of the Mean Sensitivity (MS) of the VF. Another objective was to build a normative database for the CMP and analyse its relative discriminative ability, compared to HFA, in detecting subjects with GON. We specifically hypothesised that the two index tests will have equivalent relative diagnostic precision as assessed by partial area under the receiver operating characteristic (ROC) curve at >75% specificity, across a spectrum of disease severity. In both analyses, the reference assessment for GON was specialist evaluation based on the inspection of fundus photograph or Spectral Domain – Optical Coherence Tomography (SD-OCT) evaluation of the Retinal Nerve Fibre Layer (RNFL), independent of the VF. A further objective was to evaluate examination times for the CMP and HFA.

Methods

Data collection for the normative database and discrimination analysis

People were recruited at eight study sites. These were: ASST - Santi Paolo e Carlo, Milan, Italy; Azienda Ospedaliero Universitaria Santa Maria della Misericordia di Udine, Udine, Italy; NIHR Clinical Research Facility at Moorfields Eye Hospital, London, UK; Department of Ophthalmology and Visual Sciences University of Iowa, 200 Hawkins Drive, Iowa City, IA; Department of Optometry & Vision Sciences, The University of Melbourne, Parkville, Australia; IRCCS Fondazione "G.B. Bietti", Clinica Oculistica Università degli Studi di Roma "La Sapienza", Rome, Italy; and Azienda Ospedaliera Sant'Andrea, Rome, Italy). Recruitment started on 14/09/2015 and concluded on 31/07/2017. Data collection was planned before the index test and reference standard were performed. The study was designed to achieve a target number of 1000 glaucoma subjects and 600 healthy subjects for the normative database and discrimination analysis. However, these targets were not reached by the termination date of the study.

Participants eligible for inclusion were consecutive adults (18-90 years) with:

- Best corrected visual acuity > 0.8 (if ≤ 50 years old) or >0.6 (if >50 years old) in the study eye;
- Refraction -10D / +6D; astigmatism ±2D;
- Absence of systemic pathologies that could affect the VF;
- No use of drugs interfering with the correct execution of the perimetric test;

Additional specific inclusion criteria for healthy subjects were:

- Normal optic nerve head in both eyes (no evidence of excavation, rim narrowing or notching, disc haemorrhages, RNFL thinning);
- Intraocular Pressure (IOP) less than 21 mmHg in both eyes;

- No ocular pathologies, trauma, surgeries (apart from uncomplicated cataract surgery) in both eyes;

Additional specific inclusion criteria for glaucoma subjects were:

- GON defined as glaucomatous changes to the optic nerve head (ONH) or retinal nerve fibre layer (RNFL) as determined by a specialist from fundus photograph or SD-OCT, independently of the VF.
- Patients had to be receiving anti-glaucoma therapy;
- No ocular pathologies, trauma, surgeries (apart from uncomplicated cataract surgery), other than glaucoma, in both eyes;

Eligible patients were identified based on a clinical diagnosis of GON from the clinical registry of the glaucoma clinics in the recruiting centres. An expert clinician confirmed the diagnosis of GON using the imaging data (RNFL SD-OCT or optic nerve photograph) acquired during the protocol examination (see below). Subjects were recruited consecutively. Since the VF metrics were not included in the identification of patients with GON, no stratification was planned according to disease severity.

Eligible healthy participants were identified among staff in the clinics, volunteer registries, patients' spouses or partners and patients attending the clinic for reasons other than glaucoma (for example, for preoperative assessment for cataract in the fellow eye).

If deemed eligible for the study, healthy subjects were recruited consecutively.

Both eyes were examined but only one eye per subject was used in the final analysis, chosen randomly if both eyes were eligible. All patients gave their written informed consent to participate in the study. Ethics Committee approval was obtained (International Ethics Committee of Milan, Zone A, 22/07/2015, ref: Prot. n° 0019459) and the study was registered as a clinical trial (ISRCTN13800424). This study adhered to the tenets of the Declaration of Helsinki.

Each subject had an ophthalmological evaluation following a standard operating procedure involving assessment of axial length (AL) measurement with the IOL Master (Zeiss) biometer, SD-OCT of the Optic Nerve Head (ONH) and RNFL, perimetric demonstration (only for subjects naïve to perimetry); one examination with HFA 24-2 grid SITA Standard to both eyes and one examination with CMP New Grid (see below), ZEST strategy to both eyes; colour fundus photo with CMP.

The reference standard to diagnose GON was clinical evaluation by an expert based on RNFL SD-OCT and/or optic nerve head photography. The rationale for this choice was to avoid any classification based on VF testing that could have affected the analysis of the relative discriminative power of the index tests. The two index tests were VF examinations with the HFA and the CMP. The order of CMP and HFA tests was randomized. The VF examination performed with the HFA used a 24-2 grid and the SITA – Standard algorithm. Near correction was used. Fixation was monitored with blind spot tests using the Heijl-Krakau method [8].

The VF examination performed with the CMP employed a testing grid termed 'New Grid' which differs from the HFA 24-2 grid (Supplementary Figure 1, available at www.aaojournal.org). The New Grid contains all the 52 locations tested with a 24-2, only one blind spot location (instead of 2 as in the 24-2) and 12 additional points in the macular region of the VF. The testing strategy was an adaptation of the Zippy Estimation by Sequential Testing (ZEST) [9, 10]. Since the CMP is equipped with autofocusing, no near correction was needed. Blind spot responses were monitored by projecting stimuli on the location of the ONH, identified manually by the operator on the baseline infrared fundus image captured at the beginning of the test. In all the analyses, only the 52 locations in common between the 24-2 and the New Grid were used.

For both devices, VF examinations were considered reliable if the false positive frequency (FP) was <=18% and the Blind Spot response frequency (BP) was <=25%. If either the HFA or the CMP VF was deemed unreliable, the eye was excluded from the analysis.

Statistical analysis

All analyses were based exclusively on the 52 locations in common between the 24-2 grid (HFA) and the New Grid (CMP).

Differences between the two devices in terms of Mean Sensitivity (MS) and its decrease with age in healthy subjects were analysed. Since the same eyes were tested with both devices, a mixed model was used to account for repeated measurements.

Linear regression was used to estimate expected decrease in sensitivity with age in healthy subjects (dB/years) at each VF location. Total deviation (TD) values for each VF in normal and glaucoma subjects were calculated as the deviation from the mean trend in the age model for each location. Mean Deviation (MD) was calculated as the mean of all 24-2 grid TD values in each VF. Mixed models were used to compare MS and MD values between the two devices in both the glaucoma and normal groups. MD values were only compared for the glaucoma group since subjects in the normal group were used to calculate the TD values and are bound to have a mean MD equal to zero with both devices.

Normative lower limits for each location were calculated for TD values using quantile regression [11, 12] to account for changes in normal variability with age. Since the variability of thresholds in healthy subjects is known to increase with age [12, 13], we only allowed for negative slopes in quantile regression, meaning that normative limits could not shrink with age. Only the lower 5% and 1% limits for TD values were used in this analysis.

For a fair comparison, TD values and their normative limits were calculated in the same fashion for HFA and CMP, using the dataset of healthy subjects acquired with each respective device in this study.

For each VF, we calculated the total number of TD values below the 5% and 1% limits, which we refer to as TD 5% and TD 1% respectively.

Discrimination ability of the two index tests was measured using MD, TD 5% and TD 1% as classifiers. These classifiers were used to build Receiver Operating Characteristics (ROC) curves. Instead of comparing the whole ROC curve, we analysed the Partial ROC curve (pROC) down to a minimum specificity of 0.75 to avoid comparing the two devices at too low specificity values that would fall far outside a clinically useful range. The 95% confidence intervals for Partial Areas Under the Curves (pAUCs) and p–values for differences were calculated via bootstrapping[14].

The normative data, used to calculate MD and TD metrics and their normative limits, was composed of the same set of healthy subjects used in the discrimination analysis to calculate pROC curves and their pAUCs. Therefore, they are only used here to compare the relative performance of the two devices and not to estimate or report their actual discriminative power.

To compare test times, CMP average time per location was calculated for each test and the result multiplied by the number of total points in a 24-2 grid (54 points). This made it comparable with the testing time read from the printout of the HFA.

Data collection for test - retest variability

A separate group of glaucoma and healthy subjects was recruited to assess test – retest variability with the two devices. The target number was 56 subjects with GON and 56 healthy subjects. The sample size calculation for this part of the study was based on previously reported data for test - retest in healthy subjects and glaucoma patients [15, 16]. All subjects underwent the same examinations reported for the previous section and the diagnosis of GON was again confirmed by expert evaluation of the RNFL on SD – OCT images or photographs of the optic nerve head. Subjects were sequentially recruited in the same way described for the previous part of the study. No stratification by disease (VF) severity was planned in the recruitment of glaucoma subjects. All subjects performed four VF tests: two with CMP with a 24-2 grid, ZEST strategy, and two with HFA with a 24-2 grid, SITA Standard strategy, in randomized order. All examinations were done within a time span of seven days.

Statistical analysis

Test – retest variability for the overall VF was assessed for MS using Bland – Altman plots and 95% limits of agreement. Any change in test-retest variability was evaluated by percentage reduction of the 95% interval of agreement of CMP over HFA. The 95% confidence intervals for the percentage variation were estimated using a paired bootstrap procedure with 50000 resamples. Mean Absolute Deviation (MAD) was used to assess pointwise test - retest variability. Differences in MAD, point-wise sensitivity and MS were tested using t-test statistics from linear mixed models with random effects to account for correlations between VF measurements from the same subject.

All analyses were done using R version 3.3.1 (R Foundation for Statistical Computing, Vienna, Austria).

Results

Normative database

For this part of the study, 1249 people were screened for eligibility and invited to participate between 14/09/2015 and 31/07/2017. Of these, 177 subjects 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 (48 with HFA, 20 with CMP and 2 with both devices).

Therefore, 444 healthy subjects and 499 glaucoma subjects (patients with GON) were included in the final analysis. Although no stratification by disease severity was planned, a wide spectrum of VF severity was obtained by the end of the recruitment. Glaucoma Staging System 2 (GSS2)[17] stage distribution for glaucoma participants is reported in Table 1 and depicted in Figure 1.

Subjects' age distributions are reported in Table 1. Mean age (± standard deviation [SD]) was 48 ± 16 and 68 ± 11 years for the normal and glaucoma group respectively.

Average MS was lower with CMP compared to HFA in healthy subjects (Mean ± SD, 27.6 ± 1.6 dB vs 29.4 ± 2.0 dB) and glaucoma subjects (20.5 ± 6.7 dB vs 21.9 ± 6.9 dB) and these differences were both statistically significant ($p < 0.001$). Comparison of the MD values in healthy subjects has not been performed since this group was used to calculate the normative average and therefore they were bound to have zero means for both devices. The MD values from the two devices showed good agreement (Figure 2). Indeed, the average MD (± SD) for glaucoma subjects was -6.55 ± 6.60 dB (Median: -4.37 dB, IQR: 8.92 dB) with CMP and -6.50 ± 6.63 dB (Median: -4.73 dB, IQR: 9.19 dB) with HFA and this difference was not statistically significant ($p = 0.54$).

Average number of presentations (± SD) per location in CMP was 3.02 ± 0.55 for healthy subjects and 3.70 ± 1.09 for glaucoma patients. Corrected test duration for CMP and test duration for HFA were similar in both the healthy and glaucoma subjects (see Table 2).

Point-wise sensitivity was generally lower for CMP compared to HFA (Figure 3). The average mean difference was 1.85 ± 0.06 dB (Mean ± Standard Error, $p < 0.001$) in healthy subjects and 1.46 ± 0.05 dB (Mean ± Standard Error, $p < 0.001$) in patients with glaucoma. Similarly to the MD, such a difference was reduced when total deviations were considered in glaucoma subjects (Figure 4), with 7 locations exceeding 1 dB difference.

The MS in the healthy group decreased with age in a similar fashion for both devices, with a small but statistically significant difference (-0.051 ± 0.005 dB/year for HFA and -0.027 ± 0.005 dB/year for CMP; Mean ± Standard Error; $p < 0.001$ for slope difference).

The rate of false positives was 1.6 ± 4.0 % for CMP and 1.6 ± 2.3 % for HFA (Mean ± SD).

Discrimination analysis

Relative discriminative power (relative diagnostic precision) was marginally greater for CMP when compared to HFA using the MD metric (pAUC difference ± Standard Error, 0.019 ± 0.009, $p = 0.035$, see Figure 5). There was no statistically significant difference in pAUC between CMP and HFA when using TD 5% ($p = 0.18$) or TD 1% ($p = 0.22$) as the classifier. Sensitivity values at selected specificities are reported in Table 3.

Test – retest variability

By the end of the study, 99 subjects were screened; one subject did not complete all the examinations and was excluded. In total 54 healthy subjects and 44 glaucoma patients, were recruited for the test – retest study. Bland – Altman plots are reported in Figure 6. The mean difference in MS between the first and the second test with the CMP was statistically different from zero in glaucoma subjects (Mean ± Standard Error, 0.44 ± 0.21 dB, $p = 0.041$). Bootstrap distributions of the percentage improvement for the glaucoma group are reported in Supplementary Figure 2 (available at www.aaojournal.org).

The 95% limits of agreement for MS are depicted in Figure 6. They were 49% (95% CIs: 17% to 67%) narrower for CMP (Limits of agreement: -1.31, 1.63 dB) compared to HFA (Limits of agreement: -2.84, 2.91 dB) in the healthy subjects. The 95% limits of agreement were 13% narrower for CMP (Limits of agreement: -2.26, 3.14 dB) compared to HFA (Limits of agreement: -3.11, 3.11 dB) in the glaucoma patients but the confidence intervals for these estimates were very large (95% CI: - 28% to 42%). In glaucoma subjects, the mean test - retest difference (± SD) was 0.44 ± 1.38 dB for CMP and 0 ± 1.59 dB for HFA. Bland – Altman plots for all sensitivities are reported in Figure 7. The 95% limits of agreement were generally narrower for CMP for sensitivities above or equal to 15 dB (Mean Difference: 1.80 dB, between 15 and 30 dB) and larger below 15 dB (Mean Difference: 5.46 dB).

Pointwise test – retest variability, calculated using the MAD was not significantly different between CMP and HFA for glaucoma patients (Mean ± SD, CMP: 1.03 ± 1.01 dB, HFA: 1.07 ± 1.16 dB; Mean Difference ± SE, 0.03 ± 0.2 dB, $p = 0.88$) and for healthy subjects (Mean ± SD, CMP: 0.59 ± 0.48 dB, HFA: 0.90 ± 1.15 dB; 0.08 ± 0.16 dB, $p = 0.62$).

Discussion

This study was designed to compare two index tests, CMP and HFA, in terms of test - retest variability and relative discriminative power. We recruited a large cohort of 943 subjects (499 patients with glaucoma and 444 healthy subjects) for the discrimination analysis and 98 subjects (44 glaucomatous and 54 healthy) to compare test-retest variability. The reference standard used for the diagnosis of GON was independent of VF assessment, based on specialist assessment of ONH colour photography and/or peripapillary RNFL thickness measured with SD-OCT.

The primary objective was to show a reduction of test – retest variability in the MS of at least 20%. Such a reduction was achieved in healthy subjects (49%), but not in glaucoma subjects, where the reduction was of 13%. Several factors might have contributed to this result, such as a more pronounced perimetric learning effect with CMP[18-21]. The mean difference in MS in CMP between the first and the second test was small but statistically significant and this may be indicative of a learning effect in the glaucoma test - retest cohort. This effect was not seen in the HFA data. Indeed, despite all glaucoma subjects in our sample having had previous experience with SAP, the new setup of a fundus perimeter might have created an unfamiliar testing condition for test takers. In fact, most of them were recruited from glaucoma clinics and were experienced with HFA. The different threshold acquisition strategies employed by the two devices may also explain this difference. SITA strategies incorporate spatial information between neighbouring test locations. Such an approach allows for faster threshold estimation, but it has been shown to bias the estimates introducing correlations between neighbouring points [22, 23]. On the other hand, the implementation of the ZEST strategy used in CMP tests each point independently. Moreover, test - retest variability is known to increase dramatically at lower sensitivities[24-27] and this effect may simply consume any improvements from adjusting for fixation stability afforded by the tracking in fundus perimetry. We speculate this is the reason we see much bigger improvement in test-retest variability in the healthy subjects compared to the patients in this study. This is supported by the results shown in the Bland-Altman plots for pointwise sensitivities, where it can be observed that the CMP offers no advantage in test-retest variability compared to HFA at values below 15 dB. Indeed, the 95% limits of agreement between 11 and 14 dB were larger for CMP than for HFA. The difference here might be explained by the spatial smoothing and the use of growth pattern to seed the priors [9, 22] in the SITA strategy, which might play a large role in reducing the test retest variability in this sensitivity range. However, the clinical utility of thresholds below 15 dB has been questioned. Indeed, recent evidence suggests that increasing perimetric contrast all the way to 0 dB may not be clinically useful and sensitivities obtained at severely damaged visual field locations (<15-19 dB) are unreliable and highly variable. It could be argued that improvements in tests-retest variability in the upper range of sensitivity values could be more clinically relevant for progression detection [24-29]. However, this is speculation because only analysis of long-term follow-up of glaucoma subjects with the CMP will allow the assessment of the real effect of such reduction in variability on earlier diagnosis of progression.

Additionally, Wyatt et al identified gaze instability as a possible source of variability at the edges of scotomata[30], and tracking might help reduce this effect. However, their analysis was performed with a 10-2 grid, which has a much finer spacing between locations (2 degrees). Hence, further investigation is needed to assess the effect of gaze instability in the estimation of edges on a typical testing grid, such as 24-2 or 30-2.

One limitation of our analysis is that the sample size of the glaucoma test – retest group was probably too small to reliably assess any differences, as shown by the large confidence intervals calculated via bootstrapping (Supplementary Figure 2, available at www.aaojournal.org). Post hoc power calculations based on bootstrap resampling estimated that 97 glaucoma subjects would have been needed to detect a 20% improvement at a significance level of 0.05 with 80% power. This is considerably above the initial estimates obtained from literature data [15, 16] used for designing of the study. Therefore, an additional investigation with longer test series on a larger sample might be needed to fully assess the effect of fundus tracking on test – retest variability.

Relative discriminative power for the two index tests (devices) was similar. When compared, pROC curves calculated using the number of abnormal points per field in the TD maps largely overlapped, with no evidence for any superiority of either index test (Figure 5). Statistically significant differences in pROC curves were observed when MD was used as a classifier but such differences are too small to be likely relevant in clinical situations. These results are compatible with the fact that, although the actual sensitivity estimates were lower for CMP compared to HFA, relative indices, such as the MD and TD values, showed only small differences in glaucoma subjects between the two devices, yielding similar diagnostic ability.

Our results are based on a large sample of individuals from different centres. The different age clusters, except for people older than 80 years of age, were well represented (Table 1). This was sufficient to reliably conduct an analysis on relative discriminative power. It is important to note that, for both devices, all indices used in the discrimination analysis (MD, TD 5% and TD 1%) and the normative limits for TD were recalculated in the same fashion from the raw sensitivities and are therefore comparable. However, since the normative limits have been derived from the same group of healthy subjects used in the discrimination analysis, the pAUCs are biased and can only be used to compare the relative discriminative ability of the two devices; they cannot be generalised to estimate the effective discriminative power of either the CMP or the HFA in clinical practice.

Examination times for the two devices were similar. Both devices took, on average, 5 to 6 minutes to complete. Testing times had to be corrected prior to comparison due to the greater number of tested locations with the New Grid used with CMP (65 locations) compared to the HFA 24-2 grid (54 locations). After corrections, no statistically significant differences could be detected between the two devices in healthy subjects. A statistical difference was observed in glaucoma subjects but it is clinically irrelevant (approximately an 11 second difference on average). Despite similarities in overall examination times, fewer presentations were needed to estimate thresholds in CMP when compared to HFA at the 52 matching locations. The number of presentations in healthy subjects was 157 ± 28, which is lower than that reported for SITA-Standard in the literature (276 for 52 locations) [13]. Unfortunately, interpretation of the examination times of the two devices is difficult for a variety of reasons. For example, CMP uses catch trials whereas HFA SITA algorithms use response times to estimate false positive error rates [31]. Moreover, the CMP does not project stimuli when the quality in the tracking signal is low, and this may increase overall examination time.

One limitation of our study is that the glaucoma subjects were not stratified according to disease severity, since VF data were not used in the diagnosis of GON. This could have resulted in an uneven representation of glaucoma stages. However, the range of visual field damage was sufficiently large to allow for a reliable evaluation across the whole spectrum of glaucoma damage (see Table 1 and Figure 1).

Our recruitment of healthy subjects was not population based and this is another potential limitation of our study. The main design bias potentially recruiting 'super-normals' in studies of diagnostic precision is to recruit the healthy control group using restriction criteria related to the outcome of interest [32], for example requiring the healthy controls to have normal visual fields. We explicitly avoided this bias. Nevertheless, volunteers to clinical studies may be healthier than an unselected population. This is very hard to avoid, because participants need to volunteer. However, when we analysed the MD values from the HFA printouts of the 444 healthy subjects, whose calculation is based on the independent internal normative database built in the device, we found that our sample did not show important deviations from the normative values. Indeed, the average MD was -1.12 ± 1.64 dB (Median: -0.91 dB, IQR: 1.97 dB).

Finally, the design of this study only allowed for a relative comparison of discriminative power. Evaluation of the actual diagnostic accuracy would need a further validation on an independent dataset, to assess how much these findings can be extracted on the general population. Furthermore, such an evaluation should be conducted on a set of subjects before the reference test (the clinical diagnosis of GON) is performed, as case-control scenarios are known to produce biased estimates in discrimination analyses. One option might be to test glaucoma suspects with the CMP before they are diagnosed as healthy or as having glaucoma.

References

1. Glen FC, Baker H, Crabb DP. A qualitative investigation into patients' views on visual field testing for glaucoma monitoring. BMJ Open 2014;4(1):e003996.

2. Bellmann C, Feely M, Crossland MD, et al. Fixation stability using central and pericentral fixation targets in patients with age-related macular degeneration. Ophthalmology 2004;111(12):2265-70.

3. Hanout M, Horan N, Do DV. Introduction to microperimetry and its use in analysis of geographic atrophy in age-related macular degeneration. Curr Opin Ophthalmol 2015;26(3):149-56.

4. 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.

5. Fogagnolo P, Modarelli A, Oddone F, et al. Comparison of Compass and Humphrey perimeters in detecting glaucomatous defects. Eur J Ophthalmol 2016;26(6):598-606.

6. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 2016;6(11):e012799.

7. Fidalgo BM, Crabb DP, Lawrenson JG. Methodology and reporting of diagnostic accuracy studies of automated perimetry in glaucoma: evaluation using a standardised approach. Ophthalmic Physiol Opt 2015;35(3):315-23.

8. Heijl A, Krakau CE. An automatic static perimeter, design and pilot study. Acta Ophthalmol (Copenh) 1975;53(3):293-310.

9. 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.

10. 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.

11. Bryan SR, Vermeer KA, Eilers PH, et al. Robust and censored modeling and prediction of progression in glaucomatous visual fields. Invest Ophthalmol Vis Sci 2013;54(10):6694-700.

12. Artes PH, Crabb DP. Estimating normative limits of Heidelberg Retina Tomograph optic disc rim area with quantile regression. Invest Ophthalmol Vis Sci 2010;51(1):355-61.

13. Heijl A, Lindgren G, Olsson J. Normal variability of static perimetric threshold values across the central visual field. Arch Ophthalmol 1987;105(11):1544-9.

14. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12(1):77.

15. Acton JH, Bartlett NS, Greenstein VC. Comparing the Nidek MP-1 and Humphrey field analyzer in normal subjects. Optom Vis Sci 2011;88(11):1288-97.

16. Heijl A, Lindgren A, Lindgren G. Test-retest variability in glaucomatous visual fields. Am J Ophthalmol 1989;108(2):130-5.

17. Brusini P, Filacorda S. Enhanced Glaucoma Staging System (GSS 2) for classifying functional damage in glaucoma. J Glaucoma 2006;15(1):40-6.

18. Heijl A, Lindgren G, Olsson J. The effect of perimetric experience in normal subjects. Arch Ophthalmol 1989;107(1):81-6.

19. Werner EB, Petrig B, Krupin T, Bishop KI. Variability of automated visual fields in clinically stable glaucoma patients. Invest Ophthalmol Vis Sci 1989;30(6):1083-9.

20. Kutzko KE, Brito CF, Wall M. Effect of instructions on conventional automated perimetry. Invest Ophthalmol Vis Sci 2000;41(7):2006-13.

21. Kulze JC, Stewart WC, Sutherland SE. Factors associated with a learning effect in glaucoma patients using automated perimetry. Acta Ophthalmol (Copenh) 1990;68(6):681-6.

22. Rubinstein NJ, McKendrick AM, Turpin A. Incorporating Spatial Models in Visual Field Test Procedures. Transl Vis Sci Technol 2016;5(2):7.

23. Bengtsson B, Heijl A, Olsson J. Evaluation of a new threshold visual field strategy, SITA, in normal subjects. Swedish Interactive Thresholding Algorithm. Acta Ophthalmol Scand 1998;76(2):165-9.

24. Artes PH, Iwase A, Ohno Y, et al. Properties of perimetric threshold estimates from Full Threshold, SITA Standard, and SITA Fast strategies. Invest Ophthalmol Vis Sci 2002;43(8):2654-9.

25. Russell RA, Crabb DP, Malik R, Garway-Heath DF. The relationship between variability and sensitivity in large-scale longitudinal visual field data. Invest Ophthalmol Vis Sci 2012;53(10):5985-90.

26. Gardiner SK, Swanson WH, Goren D, et al. Assessment of the reliability of standard automated perimetry in regions of glaucomatous damage. Ophthalmology 2014;121(7):1359-69.

27. Gardiner SK, Mansberger SL. Effect of Restricting Perimetry Testing Algorithms to Reliable Sensitivities on Test-Retest Variability. Invest Ophthalmol Vis Sci 2016;57(13):5631-6.

28. Pathak M, Demirel S, Gardiner SK. Reducing Variability of Perimetric Global Indices from Eyes with Progressive Glaucoma by Censoring Unreliable Sensitivity Data. Transl Vis Sci Technol 2017;6(4):11.

29. Gardiner SK, Swanson WH, Demirel S. The Effect of Limiting the Range of Perimetric Sensitivities on Pointwise Assessment of Visual Field Progression in Glaucoma. Invest Ophthalmol Vis Sci 2016;57(1):288-94.

30. 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.

31. Newkirk MR, Gardiner SK, Demirel S, Johnson CA. Assessment of false positives with the Humphrey Field Analyzer II perimeter with the SITA Algorithm. Invest Ophthalmol Vis Sci 2006;47(10):4632-7.

32. Garway-Heath DF, Hitchings RA. Sources of bias in studies of optic disc and retinal nerve fibre layer morphology. Br J Ophthalmol 1998;82(9):986.

Figure Legends

Figure 1. GSS2[17] plot showing the distribution of the 499 subjects with glaucomatous optic neuropathy in the different stages of the classification. The light grey lines indicate the boundaries for the different stages. Subjects are classified based on their MD and PSD values

directly taken from the HFA printout. The distribution is approximately uniform across the different stages.

Figure 2. The two panels show the agreement of MD (on the left) and MS (on the right) values between CMP (vertical axis) and HFA (horizontal axis). The black solid line indicates the ideal perfect agreement. The red dots represent the healthy subjects while the green dots indicate glaucoma subjects. Differently from MS, MD values did not show important differences between the two devices.

Figure 3. Average sensitivity (dB) for each of the 52 locations considered in this analysis for CMP (A) and HFA (B). The bottom panels report the average pairwise difference per location in the healthy subjects (C) and for glaucoma patients (D).

Figure 4. Average total deviation value (dB) for each of the 52 locations considered in this analysis for CMP (A) and HFA (B). Panel C reports the average pairwise difference (CMP – HFA) in Total Deviation per location in the glaucoma subjects (in bold all differences exceeding 1 dB).

Figure 5. Partial ROC curves built using the MD (in the leftmost panel) as a classifier. The middle and rightmost panels depict partial ROC curves built using the number of abnormal locations at two different cut-offs, 5% and 1%, on the probability maps for TD values. There was no significant difference in either the TD 5% or the TD 1%. MD = Mean Deviation; TD = Total Deviation.

Figure 6. Bland – Altman plots for MS. Red dots represent MS measurements from the HFA, blue dots from the CMP. The shaded grey area indicates the 95 % limits of agreement on the test-retest difference. The black solid line indicates the mean difference between test-retest MS measurements. A small offset in the mean difference can be detected in the glaucoma group with the CMP (bottom – left panel).

Figure 7. Bland – Altman plots for all sensitivities. Red dots represent MS measurements from the HFA, blue dots from the CMP. The shaded grey area indicates the 95 % limits of agreement on the test-retest difference. 95% Limits of agreement were narrower for sensitivities above or equal to 15 dB, larger between 11 dB and 14 dB and equivalent below 10 dB. The larger range in the differences was at 14 dB (-27 dB, 27 dB) for CMP and at 12 dB (-24 dB, 25 dB) for HFA.