City Research Online

City, University of London Institutional Repository

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/21784/

Link to published version: https://doi.org/10.1016/j.ophtha.2018.08.010

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

38

39   This article contains additional online-only material. The following should appear online-only:

40   Supplementary Figure 1 and 2.

41

42   Running head: Comparison of Compass and Humphrey Field Analyzer

43

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

69   normal subjects. MAD was very similar, with no significant differences.

70   **Conclusions**: Relative diagnostic precision of the two devices is equivalent. Test-retest

71   variability of mean sensitivity for CMP was better than for HFA.

72   Standard Automated Perimetry (SAP) is used to assess the visual field (VF) and is a key

73   examination for detection, diagnosis and follow up in glaucoma. SAP typically uses stimuli of

74   varying intensities to assess the differential light sensitivity at static locations across the VF.

75   The examination demands strong cooperation[1] from test subjects; they are required to

76   maintain central fixation and respond timely and accurately to the presented stimuli. Fixation

77   instability might be an unavoidable feature of a person's vision, especially with advanced age

78   and macular damage[2]. One proposed solution has been to incorporate live fundus tracking in

79   the macular perimetric exam to compensate for eye movements in unstable fixation [3].

80   Recently, a novel instrument, the COMPASS fundus perimeter (CMP, CenterVue, Padua, Italy),

81   has successfully employed a live fundus tracking technology for wide field (30 degrees) VF

82   assessment [4, 5] yielding results comparable with the Humphrey Field Analyzer (HFA) in a

83   preliminary study [4]. The CMP captures images of the fundus during the perimetric

84   examination using a scanning laser ophthalmoscope. This design feature is intended to afford

85   compensation for eye movements when the stimuli are presented at predetermined test

86   locations. Moreover, the instrument provides colour images of the fundus and optic nerve that

87   can be mapped to the final perimetric results potentially providing clinically useful

88   information about structure and function in one assessment.

89   Diagnostic accuracy studies are used to certify new examinations before they are brought into

90   clinical practice. The CMP has not yet been scrutinised in this way and this is the main

91   purpose of our investigation. Studies investigating relative diagnostic accuracy are at risk of

92   bias due to shortcomings in design and conduct. For this reason, we designed our study to

93   follow appropriate guidelines on this specific aim [6, 7].

94   Our cross-sectional and multicentre study was designed to evaluate and compare two index

95   tests, namely the CMP and the HFA. One objective was to evaluate and compare test – retest

96   variability of the two index tests in healthy subjects and patients with glaucomatous optic

97   neuropathy (GON). We hypothesised that the CMP could obtain a 20% reduction in test-retest

98   variability on the measurement of the Mean Sensitivity (MS) of the VF. Another objective was

99   to build a normative database for the CMP and analyse its relative discriminative ability,

100   compared to HFA, in detecting subjects with GON. We specifically hypothesised that the two

101   index tests will have equivalent relative diagnostic precision as assessed by partial area under

102   the receiver operating characteristic (ROC) curve at >75% specificity, across a spectrum of

103   disease severity. In both analyses, the reference assessment for GON was specialist evaluation

104   based on the inspection of fundus photograph or Spectral Domain – Optical Coherence

105   Tomography (SD-OCT) evaluation of the Retinal Nerve Fibre Layer (RNFL), independent of

106   the VF. A further objective was to evaluate examination times for the CMP and HFA.

107

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;

132   •   No ocular pathologies, trauma, surgeries (apart from uncomplicated cataract surgery)

133   in both eyes;

134 Additional specific inclusion criteria for glaucoma subjects were:

135   •   GON defined as glaucomatous changes to the optic nerve head (ONH) or retinal nerve

136   fibre layer (RNFL) as determined by a specialist from fundus photograph or SD-OCT,

137   independently of the VF.

138   •   Patients had to be receiving anti-glaucoma therapy;

139   •   No ocular pathologies, trauma, surgeries (apart from uncomplicated cataract surgery),

140   other than glaucoma, in both eyes;

141 Eligible patients were identified based on a clinical diagnosis of GON from the clinical registry

142 of the glaucoma clinics in the recruiting centres. An expert clinician confirmed the diagnosis of

143 GON using the imaging data (RNFL SD-OCT or optic nerve photograph) acquired during the

144 protocol examination (see below). Subjects were recruited consecutively. Since the VF metrics

145 were not included in the identification of patients with GON, no stratification was planned

146 according to disease severity.

147 Eligible healthy participants were identified among staff in the clinics, volunteer registries,

148 patients' spouses or partners and patients attending the clinic for reasons other than

149 glaucoma (for example, for preoperative assessment for cataract in the fellow eye).

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

151 Both eyes were examined but only one eye per subject was used in the final analysis, chosen

152 randomly if both eyes were eligible. All patients gave their written informed consent to

153 participate in the study. Ethics Committee approval was obtained (International Ethics

154 Committee of Milan, Zone A, 22/07/2015, ref: Prot. n° 0019459) and the study was registered

155 as a clinical trial (ISRCTN13800424). This study adhered to the tenets of the Declaration of

156 Helsinki.

157 Each subject had an ophthalmological evaluation following a standard operating procedure

158 involving assessment of axial length (AL) measurement with the IOL Master (Zeiss) biometer,

159 SD-OCT of the Optic Nerve Head (ONH) and RNFL, perimetric demonstration (only for

160 subjects naïve to perimetry); one examination with HFA 24-2 grid SITA Standard to both eyes

161 and one examination with CMP New Grid (see below), ZEST strategy to both eyes; colour

162 fundus photo with CMP.

163 The reference standard to diagnose GON was clinical evaluation by an expert based on RNFL

164 SD-OCT and/or optic nerve head photography. The rationale for this choice was to avoid any

165 classification based on VF testing that could have affected the analysis of the relative

166 discriminative power of the index tests. The two index tests were VF examinations with the

167 HFA and the CMP. The order of CMP and HFA tests was randomized. The VF examination

168 performed with the HFA used a 24-2 grid and the SITA – Standard algorithm. Near correction

169 was used. Fixation was monitored with blind spot tests using the Heijl-Krakau method [8].

170 The VF examination performed with the CMP employed a testing grid termed 'New Grid'

171 which differs from the HFA 24-2 grid (Supplementary Figure 1, available at

172 www.aaojournal.org). The New Grid contains all the 52 locations tested with a 24-2, only one

173 blind spot location (instead of 2 as in the 24-2) and 12 additional points in the macular region

174 of the VF. The testing strategy was an adaptation of the Zippy Estimation by Sequential

175 Testing (ZEST) [9, 10]. Since the CMP is equipped with autofocusing, no near correction was

176 needed. Blind spot responses were monitored by projecting stimuli on the location of the

177 ONH, identified manually by the operator on the baseline infrared fundus image captured at

178 the beginning of the test. In all the analyses, only the 52 locations in common between the 24-

179 2 and the New Grid were used.

180 For both devices, VF examinations were considered reliable if the false positive frequency

181 (FP) was <=18% and the Blind Spot response frequency (BP) was <=25%. If either the HFA or

182   the CMP VF was deemed unreliable, the eye was excluded from the analysis.

183

184   *Statistical analysis*

185   All analyses were based exclusively on the 52 locations in common between the 24-2 grid

186   (HFA) and the New Grid (CMP).

187   Differences between the two devices in terms of Mean Sensitivity (MS) and its decrease with

188   age in healthy subjects were analysed. Since the same eyes were tested with both devices, a

189   mixed model was used to account for repeated measurements.

190   Linear regression was used to estimate expected decrease in sensitivity with age in healthy

191   subjects (dB/years) at each VF location. Total deviation (TD) values for each VF in normal and

192   glaucoma subjects were calculated as the deviation from the mean trend in the age model for

193   each location. Mean Deviation (MD) was calculated as the mean of all 24-2 grid TD values in

194   each VF. Mixed models were used to compare MS and MD values between the two devices in

195   both the glaucoma and normal groups. MD values were only compared for the glaucoma

196   group since subjects in the normal group were used to calculate the TD values and are bound

197   to have a mean MD equal to zero with both devices.

198   Normative lower limits for each location were calculated for TD values using quantile

199   regression [11, 12] to account for changes in normal variability with age. Since the variability of

200   thresholds in healthy subjects is known to increase with age [12, 13], we only allowed for

201   negative slopes in quantile regression, meaning that normative limits could not shrink with

202   age. Only the lower 5% and 1% limits for TD values were used in this analysis.

203   For a fair comparison, TD values and their normative limits were calculated in the same

204   fashion for HFA and CMP, using the dataset of healthy subjects acquired with each respective

205   device in this study.

206   For each VF, we calculated the total number of TD values below the 5% and 1% limits, which

207   we refer to as TD 5% and TD 1% respectively.

208   Discrimination ability of the two index tests was measured using MD, TD 5% and TD 1% as

209   classifiers. These classifiers were used to build Receiver Operating Characteristics (ROC)

210   curves. Instead of comparing the whole ROC curve, we analysed the Partial ROC curve (pROC)

211   down to a minimum specificity of 0.75 to avoid comparing the two devices at too low

212   specificity values that would fall far outside a clinically useful range. The 95% confidence

213   intervals for Partial Areas Under the Curves (pAUCs) and p–values for differences were

214   calculated via bootstrapping[14].

215   The normative data, used to calculate MD and TD metrics and their normative limits, was

216   composed of the same set of healthy subjects used in the discrimination analysis to calculate

217   pROC curves and their pAUCs. Therefore, they are only used here to compare the relative

218   performance of the two devices and not to estimate or report their actual discriminative

219   power.

220   To compare test times, CMP average time per location was calculated for each test and the

221   result multiplied by the number of total points in a 24-2 grid (54 points). This made it

222   comparable with the testing time read from the printout of the HFA.

223

224   *Data collection for test - retest variability*

225   A separate group of glaucoma and healthy subjects was recruited to assess test – retest

226   variability with the two devices. The target number was 56 subjects with GON and 56 healthy

227   subjects. The sample size calculation for this part of the study was based on previously

228   reported data for test - retest in healthy subjects and glaucoma patients [15, 16]. All subjects

229   underwent the same examinations reported for the previous section and the diagnosis of GON

230   was again confirmed by expert evaluation of the RNFL on SD – OCT images or photographs of

231 the optic nerve head. Subjects were sequentially recruited in the same way described for the

232 previous part of the study. No stratification by disease (VF) severity was planned in the

233 recruitment of glaucoma subjects. All subjects performed four VF tests: two with CMP with a

234 24-2 grid, ZEST strategy, and two with HFA with a 24-2 grid, SITA Standard strategy, in

235 randomized order. All examinations were done within a time span of seven days.

236

237 *Statistical analysis*

238 Test – retest variability for the overall VF was assessed for MS using Bland – Altman plots and

239 95% limits of agreement. Any change in test-retest variability was evaluated by percentage

240 reduction of the 95% interval of agreement of CMP over HFA. The 95% confidence intervals

241 for the percentage variation were estimated using a paired bootstrap procedure with 50000

242 resamples. Mean Absolute Deviation (MAD) was used to assess pointwise test - retest

243 variability. Differences in MAD, point-wise sensitivity and MS were tested using t-test

244 statistics from linear mixed models with random effects to account for correlations between

245 VF measurements from the same subject.

246 All analyses were done using R version 3.3.1 (R Foundation for Statistical Computing, Vienna,

247 Austria).

248

249

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 ($p = 0.54$).

273 Average number of presentations (± SD) per location in CMP was 3.02 ± 0.55 for healthy

274 subjects and 3.70 ± 1.09 for glaucoma patients. Corrected test duration for CMP and test

275 duration for HFA were similar in both the healthy and glaucoma subjects (see Table 2).

276 Point-wise sensitivity was generally lower for CMP compared to HFA (Figure 3). The average

277 mean difference was 1.85 ± 0.06 dB (Mean ± Standard Error, $p < 0.001$) in healthy subjects

278 and 1.46 ± 0.05 dB (Mean ± Standard Error, $p < 0.001$) in patients with glaucoma. Similarly to

279 the MD, such a difference was reduced when total deviations were considered in glaucoma

280 subjects (Figure 4), with 7 locations exceeding 1 dB difference.

281 The MS in the healthy group decreased with age in a similar fashion for both devices, with a

282 small but statistically significant difference (-0.051 ± 0.005 dB/year for HFA and -0.027 ±

283 0.005 dB/year for CMP; Mean ± Standard Error; $p < 0.001$ for slope difference).

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

285

286 *Discrimination analysis*

287 Relative discriminative power (relative diagnostic precision) was marginally greater for CMP

288 when compared to HFA using the MD metric (pAUC difference ± Standard Error, 0.019 ±

289 0.009, $p = 0.035$, see Figure 5). There was no statistically significant difference in pAUC

290 between CMP and HFA when using TD 5% ($p = 0.18$) or TD 1% ($p=0.22$) as the classifier.

291 Sensitivity values at selected specificities are reported in Table 3.

292

293 *Test – retest variability*

294 By the end of the study, 99 subjects were screened; one subject did not complete all the

295 examinations and was excluded. In total 54 healthy subjects and 44 glaucoma patients, were

296 recruited for the test – retest study. Bland – Altman plots are reported in Figure 6. The mean

297 difference in MS between the first and the second test with the CMP was statistically different

298   from zero in glaucoma subjects (Mean ± Standard Error, 0.44 ± 0.21 dB, p = 0.041). Bootstrap

299   distributions of the percentage improvement for the glaucoma group are reported in

300   Supplementary Figure 2 (available at www.aaojournal.org).

301   The 95% limits of agreement for MS are depicted in Figure 6. They were 49% (95% CIs: 17%

302   to 67%) narrower for CMP (Limits of agreement: -1.31, 1.63 dB) compared to HFA (Limits of

303   agreement: -2.84, 2.91 dB) in the healthy subjects. The 95% limits of agreement were 13%

304   narrower for CMP (Limits of agreement: -2.26, 3.14 dB) compared to HFA (Limits of

305   agreement: -3.11, 3.11 dB) in the glaucoma patients but the confidence intervals for these

306   estimates were very large (95% CI: - 28% to 42%). In glaucoma subjects, the mean test -

307   retest difference (± SD) was 0.44 ± 1.38 dB for CMP and 0 ± 1.59 dB for HFA. Bland – Altman

308   plots for all sensitivities are reported in Figure 7. The 95% limits of agreement were generally

309   narrower for CMP for sensitivities above or equal to 15 dB (Mean Difference: 1.80 dB,

310   between 15 and 30 dB) and larger below 15 dB (Mean Difference: 5.46 dB).

311   Pointwise test – retest variability, calculated using the MAD was not significantly different

312   between CMP and HFA for glaucoma patients (Mean ± SD, CMP: 1.03 ± 1.01 dB, HFA: 1.07 ±

313   1.16 dB; Mean Difference ± SE, 0.03 ± 0.2 dB, p = 0.88) and for healthy subjects (Mean ± SD,

314   CMP: 0.59 ± 0.48 dB, HFA: 0.90 ± 1.15 dB; 0.08 ± 0.16 dB, p = 0.62).

315

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

341 perimetry. We speculate this is the reason we see much bigger improvement in test-retest

342 variability in the healthy subjects compared to the patients in this study. This is supported by

343 the results shown in the Bland-Altman plots for pointwise sensitivities, where it can be

344 observed that the CMP offers no advantage in test-retest variability compared to HFA at

345 values below 15 dB. Indeed, the 95% limits of agreement between 11 and 14 dB were larger

346 for CMP than for HFA. The difference here might be explained by the spatial smoothing and

347 the use of growth pattern to seed the priors [9, 22] in the SITA strategy, which might play a large

348 role in reducing the test retest variability in this sensitivity range. However, the clinical utility

349 of thresholds below 15 dB has been questioned. Indeed, recent evidence suggests that

350 increasing perimetric contrast all the way to 0 dB may not be clinically useful and sensitivities

351 obtained at severely damaged visual field locations (<15-19 dB) are unreliable and highly

352 variable. It could be argued that improvements in tests-retest variability in the upper range of

353 sensitivity values could be more clinically relevant for progression detection [24-29]. However,

354 this is speculation because only analysis of long-term follow-up of glaucoma subjects with the

355 CMP will allow the assessment of the real effect of such reduction in variability on earlier

356 diagnosis of progression.

357 Additionally, Wyatt et al identified gaze instability as a possible source of variability at the

358 edges of scotomata[30], and tracking might help reduce this effect. However, their analysis was

359 performed with a 10-2 grid, which has a much finer spacing between locations (2 degrees).

360 Hence, further investigation is needed to assess the effect of gaze instability in the estimation

361 of edges on a typical testing grid, such as 24-2 or 30-2.

362 One limitation of our analysis is that the sample size of the glaucoma test – retest group was

363 probably too small to reliably assess any differences, as shown by the large confidence

364 intervals calculated via bootstrapping (Supplementary Figure 2, available at

365 www.aaojournal.org). Post hoc power calculations based on bootstrap resampling estimated

366   that 97 glaucoma subjects would have been needed to detect a 20% improvement at a

367   significance level of 0.05 with 80% power. This is considerably above the initial estimates

368   obtained from literature data [15, 16] used for designing of the study. Therefore, an additional

369   investigation with longer test series on a larger sample might be needed to fully assess the

370   effect of fundus tracking on test – retest variability.

371   Relative discriminative power for the two index tests (devices) was similar. When compared,

372   pROC curves calculated using the number of abnormal points per field in the TD maps largely

373   overlapped, with no evidence for any superiority of either index test (Figure 5). Statistically

374   significant differences in pROC curves were observed when MD was used as a classifier but

375   such differences are too small to be likely relevant in clinical situations. These results are

376   compatible with the fact that, although the actual sensitivity estimates were lower for CMP

377   compared to HFA, relative indices, such as the MD and TD values, showed only small

378   differences in glaucoma subjects between the two devices, yielding similar diagnostic ability.

379   Our results are based on a large sample of individuals from different centres. The different age

380   clusters, except for people older than 80 years of age, were well represented (Table 1). This

381   was sufficient to reliably conduct an analysis on relative discriminative power. It is important

382   to note that, for both devices, all indices used in the discrimination analysis (MD, TD 5% and

383   TD 1%) and the normative limits for TD were recalculated in the same fashion from the raw

384   sensitivities and are therefore comparable. However, since the normative limits have been

385   derived from the same group of healthy subjects used in the discrimination analysis, the

386   pAUCs are biased and can only be used to compare the relative discriminative ability of the

387   two devices; they cannot be generalised to estimate the effective discriminative power of

388   either the CMP or the HFA in clinical practice.

389   Examination times for the two devices were similar. Both devices took, on average, 5 to 6

390   minutes to complete. Testing times had to be corrected prior to comparison due to the greater

391 number of tested locations with the New Grid used with CMP (65 locations) compared to the

392 HFA 24-2 grid (54 locations). After corrections, no statistically significant differences could be

393 detected between the two devices in healthy subjects. A statistical difference was observed in

394 glaucoma subjects but it is clinically irrelevant (approximately an 11 second difference on

395 average). Despite similarities in overall examination times, fewer presentations were needed

396 to estimate thresholds in CMP when compared to HFA at the 52 matching locations. The

397 number of presentations in healthy subjects was 157 ± 28, which is lower than that reported

398 for SITA-Standard in the literature (276 for 52 locations) [13]. Unfortunately, interpretation of

399 the examination times of the two devices is difficult for a variety of reasons. For example, CMP

400 uses catch trials whereas HFA SITA algorithms use response times to estimate false positive

401 error rates [31]. Moreover, the CMP does not project stimuli when the quality in the tracking

402 signal is low, and this may increase overall examination time.

403 One limitation of our study is that the glaucoma subjects were not stratified according to

404 disease severity, since VF data were not used in the diagnosis of GON. This could have

405 resulted in an uneven representation of glaucoma stages. However, the range of visual field

406 damage was sufficiently large to allow for a reliable evaluation across the whole spectrum of

407 glaucoma damage (see Table 1 and Figure 1).

408 Our recruitment of healthy subjects was not population based and this is another potential

409 limitation of our study. The main design bias potentially recruiting 'super-normals' in studies

410 of diagnostic precision is to recruit the healthy control group using restriction criteria related

411 to the outcome of interest [32], for example requiring the healthy controls to have normal visual

412 fields. We explicitly avoided this bias. Nevertheless, volunteers to clinical studies may be

413 healthier than an unselected population. This is very hard to avoid, because participants need

414 to volunteer. However, when we analysed the MD values from the HFA printouts of the 444

415 healthy subjects, whose calculation is based on the independent internal normative database

416 built in the device, we found that our sample did not show important deviations from the

417 normative values. Indeed, the average MD was -1.12 ± 1.64 dB (Median: -0.91 dB, IQR: 1.97

418 dB).

419 Finally, the design of this study only allowed for a relative comparison of discriminative

420 power. Evaluation of the actual diagnostic accuracy would need a further validation on an

421 independent dataset, to assess how much these findings can be extracted on the general

422 population. Furthermore, such an evaluation should be conducted on a set of subjects before

423 the reference test (the clinical diagnosis of GON) is performed, as case-control scenarios are

424 known to produce biased estimates in discrimination analyses. One option might be to test

425 glaucoma suspects with the CMP before they are diagnosed as healthy or as having glaucoma.

426

References

428 1.   Glen FC, Baker H, Crabb DP. A qualitative investigation into patients' views on visual

429 field testing for glaucoma monitoring. BMJ Open 2014;4(1):e003996.

430 2.   Bellmann C, Feely M, Crossland MD, et al. Fixation stability using central and

431 pericentral fixation targets in patients with age-related macular degeneration. Ophthalmology

432 2004;111(12):2265-70.

433 3.   Hanout M, Horan N, Do DV. Introduction to microperimetry and its use in analysis of

434 geographic atrophy in age-related macular degeneration. Curr Opin Ophthalmol

435 2015;26(3):149-56.

436 4.   Rossetti L, Digiuni M, Rosso A, et al. Compass: clinical evaluation of a new instrument

437 for the diagnosis of glaucoma. PLoS One 2015;10(3):e0122157.

438 5.   Fogagnolo P, Modarelli A, Oddone F, et al. Comparison of Compass and Humphrey

439 perimeters in detecting glaucomatous defects. Eur J Ophthalmol 2016;26(6):598-606.

440   6.   Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting

441   diagnostic accuracy studies: explanation and elaboration. BMJ Open 2016;6(11):e012799.

442   7.   Fidalgo BM, Crabb DP, Lawrenson JG. Methodology and reporting of diagnostic

443   accuracy studies of automated perimetry in glaucoma: evaluation using a standardised

444   approach. Ophthalmic Physiol Opt 2015;35(3):315-23.

445   8.   Heijl A, Krakau CE. An automatic static perimeter, design and pilot study. Acta

446   Ophthalmol (Copenh) 1975;53(3):293-310.

447   9.   Turpin A, McKendrick AM, Johnson CA, Vingrys AJ. Properties of perimetric threshold

448   estimates from full threshold, ZEST, and SITA-like strategies, as determined by computer

449   simulation. Invest Ophthalmol Vis Sci 2003;44(11):4787-95.

450   10.   King-Smith PE, Grigsby SS, Vingrys AJ, et al. Efficient and unbiased modifications of the

451   QUEST threshold method: theory, simulations, experimental evaluation and practical

452   implementation. Vision Res 1994;34(7):885-912.

453   11.   Bryan SR, Vermeer KA, Eilers PH, et al. Robust and censored modeling and prediction

454   of progression in glaucomatous visual fields. Invest Ophthalmol Vis Sci 2013;54(10):6694-

455   700.

456   12.   Artes PH, Crabb DP. Estimating normative limits of Heidelberg Retina Tomograph optic

457   disc rim area with quantile regression. Invest Ophthalmol Vis Sci 2010;51(1):355-61.

458   13.   Heijl A, Lindgren G, Olsson J. Normal variability of static perimetric threshold values

459   across the central visual field. Arch Ophthalmol 1987;105(11):1544-9.

460   14.   Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to

461   analyze and compare ROC curves. BMC Bioinformatics 2011;12(1):77.

462   15.   Acton JH, Bartlett NS, Greenstein VC. Comparing the Nidek MP-1 and Humphrey field

463   analyzer in normal subjects. Optom Vis Sci 2011;88(11):1288-97.

464   16.   Heijl A, Lindgren A, Lindgren G. Test-retest variability in glaucomatous visual fields.

465   Am J Ophthalmol 1989;108(2):130-5.

466   17.   Brusini P, Filacorda S. Enhanced Glaucoma Staging System (GSS 2) for classifying

467   functional damage in glaucoma. J Glaucoma 2006;15(1):40-6.

468   18.   Heijl A, Lindgren G, Olsson J. The effect of perimetric experience in normal subjects.

469   Arch Ophthalmol 1989;107(1):81-6.

470   19.   Werner EB, Petrig B, Krupin T, Bishop KI. Variability of automated visual fields in

471   clinically stable glaucoma patients. Invest Ophthalmol Vis Sci 1989;30(6):1083-9.

472   20.   Kutzko KE, Brito CF, Wall M. Effect of instructions on conventional automated

473   perimetry. Invest Ophthalmol Vis Sci 2000;41(7):2006-13.

474   21.   Kulze JC, Stewart WC, Sutherland SE. Factors associated with a learning effect in

475   glaucoma patients using automated perimetry. Acta Ophthalmol (Copenh) 1990;68(6):681-6.

476   22.   Rubinstein NJ, McKendrick AM, Turpin A. Incorporating Spatial Models in Visual Field

477   Test Procedures. Transl Vis Sci Technol 2016;5(2):7.

478   23.   Bengtsson B, Heijl A, Olsson J. Evaluation of a new threshold visual field strategy, SITA,

479   in normal subjects. Swedish Interactive Thresholding Algorithm. Acta Ophthalmol Scand

480   1998;76(2):165-9.

481   24.   Artes PH, Iwase A, Ohno Y, et al. Properties of perimetric threshold estimates from Full

482   Threshold, SITA Standard, and SITA Fast strategies. Invest Ophthalmol Vis Sci

483   2002;43(8):2654-9.

484   25.   Russell RA, Crabb DP, Malik R, Garway-Heath DF. The relationship between variability

485   and sensitivity in large-scale longitudinal visual field data. Invest Ophthalmol Vis Sci

486   2012;53(10):5985-90.

487 26.   Gardiner SK, Swanson WH, Goren D, et al. Assessment of the reliability of standard

488 automated perimetry in regions of glaucomatous damage. Ophthalmology 2014;121(7):1359-

489 69.

490 27.   Gardiner SK, Mansberger SL. Effect of Restricting Perimetry Testing Algorithms to

491 Reliable Sensitivities on Test-Retest Variability. Invest Ophthalmol Vis Sci 2016;57(13):5631-

492 6.

493 28.   Pathak M, Demirel S, Gardiner SK. Reducing Variability of Perimetric Global Indices

494 from Eyes with Progressive Glaucoma by Censoring Unreliable Sensitivity Data. Transl Vis Sci

495 Technol 2017;6(4):11.

496 29.   Gardiner SK, Swanson WH, Demirel S. The Effect of Limiting the Range of Perimetric

497 Sensitivities on Pointwise Assessment of Visual Field Progression in Glaucoma. Invest

498 Ophthalmol Vis Sci 2016;57(1):288-94.

499 30.   Wyatt HJ, Dul MW, Swanson WH. Variability of visual field measurements is correlated

500 with the gradient of visual sensitivity. Vision Res 2007;47(7):925-36.

501 31.   Newkirk MR, Gardiner SK, Demirel S, Johnson CA. Assessment of false positives with

502 the Humphrey Field Analyzer II perimeter with the SITA Algorithm. Invest Ophthalmol Vis Sci

503 2006;47(10):4632-7.

504 32.   Garway-Heath DF, Hitchings RA. Sources of bias in studies of optic disc and retinal

505 nerve fibre layer morphology. Br J Ophthalmol 1998;82(9):986.

506

507 **Figure Legends**

508 **Figure 1.** GSS2[17] plot showing the distribution of the 499 subjects with glaucomatous optic

509 neuropathy in the different stages of the classification. The light grey lines indicate the

510 boundaries for the different stages. Subjects are classified based on their MD and PSD values

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

512   different stages.

513

514   **Figure 2.** The two panels show the agreement of MD (on the left) and MS (on the right) values

515   between CMP (vertical axis) and HFA (horizontal axis). The black solid line indicates the ideal

516   perfect agreement. The red dots represent the healthy subjects while the green dots indicate

517   glaucoma subjects. Differently from MS, MD values did not show important differences

518   between the two devices.

519

520   **Figure 3.** Average sensitivity (dB) for each of the 52 locations considered in this analysis for

521   CMP (A) and HFA (B). The bottom panels report the average pairwise difference per location

522   in the healthy subjects (C) and for glaucoma patients (D).

523

524   **Figure 4.** Average total deviation value (dB) for each of the 52 locations considered in this

525   analysis for CMP (A) and HFA (B). Panel C reports the average pairwise difference (CMP –

526   HFA) in Total Deviation per location in the glaucoma subjects (in bold all differences

527   exceeding 1 dB).

528

529   **Figure 5.** Partial ROC curves built using the MD (in the leftmost panel) as a classifier. The

530   middle and rightmost panels depict partial ROC curves built using the number of abnormal

531   locations at two different cut-offs, 5% and 1%, on the probability maps for TD values. There

532   was no significant difference in either the TD 5% or the TD 1%. MD = Mean Deviation; TD =

533   Total Deviation.

534

535 **Figure 6.** Bland – Altman plots for MS. Red dots represent MS measurements from the HFA,

536 blue dots from the CMP. The shaded grey area indicates the 95 % limits of agreement on the

537 test-retest difference. The black solid line indicates the mean difference between test-retest

538 MS measurements. A small offset in the mean difference can be detected in the glaucoma

539 group with the CMP (bottom – left panel).

540

541 **Figure 7.** Bland – Altman plots for all sensitivities. Red dots represent MS measurements from

542 the HFA, blue dots from the CMP. The shaded grey area indicates the 95 % limits of agreement

543 on the test-retest difference. 95% Limits of agreement were narrower for sensitivities above

544 or equal to 15 dB, larger between 11 dB and 14 dB and equivalent below 10 dB. The larger

545 range in the differences was at 14 dB (-27 dB, 27 dB) for CMP and at 12 dB (-24 dB, 25 dB) for

546 HFA.