



# City Research Online

## City St George's, University of London

**Citation:** Lazaridis, G., Lorenzi, M., Mohamed-Noriega, J., Aguilar-Munoz, S., Suzuki, K., Nomoto, H., Ourselin, S., Garway-Heath, D. F. & United Kingdom Glaucoma Treatment Study Investigators (2021). OCT Signal Enhancement with Deep Learning. *Ophthalmology Glaucoma*, 4(3), pp. 295-304. doi: 10.1016/j.ogla.2020.10.008

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/26980/>

**Link to published version:** <https://doi.org/10.1016/j.ogla.2020.10.008>

**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](#).

# 1 OCT signal enhancement with deep learning

2 Georgios Lazaridis,<sup>1,2,4</sup> Marco Lorenzi,<sup>3</sup> Jibrán Mohamed-Noriega,<sup>1,5</sup> Soledad Aguilar-Munoz,<sup>1</sup>

3 Katsuyoshi Suzuki,<sup>1</sup> Hiroki Nomoto,<sup>1</sup> Sebastien Ourselin,<sup>4</sup> David F. Garway-Heath,<sup>1</sup> on behalf of the

4 United Kingdom Glaucoma Treatment Study Investigators

5 <sup>1</sup>NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of  
6 Ophthalmology, London, United Kingdom

7 <sup>2</sup>Centre for Medical Image Computing, University College London, London, United Kingdom

8 <sup>3</sup>Université Côte d'Azur, Inria Sophia Antipolis, Epione Research Project, France

9 <sup>4</sup>School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom

10 <sup>5</sup>Departamento de Oftalmología, Hospital Universitario, UANL, México

11

12 **Purpose:** To establish whether deep learning methods are able to improve the signal-to-noise  
13 ratio of time-domain (TD) optical coherence tomography (OCT) images to approach that of  
14 spectral-domain (SD) OCT.

15 **Design:** Method agreement study and progression-detection in a randomized, double-masked,  
16 placebo-controlled, multi-centre trial for open-angle glaucoma (OAG) [UK Glaucoma Treatment  
17 Study (UKGTS)].

18 **Participants:** Cohort for training and validation: 77 stable OAG participants with TDOCT and  
19 SDOCT imaging at up to 11 visits within 3 months. Cohort for testing: 284 newly-diagnosed OAG  
20 patients with TDOCT from a cohort of 516 recruited at 10 UK centres between 2007 and 2010.

21 **Methods:** An ensemble of generative adversarial networks (GANs) was trained on TDOCT and  
22 SDOCT image pairs from the training dataset and applied to TDOCT images from the testing  
23 dataset. TDOCT were converted to synthesized SDOCT images and segmented via Bayesian fusion  
24 on the output of the GANs.

25 **Main Outcome Measures:** 1) Bland-Altman analysis to assess agreement between TDOCT and  
26 synthesized SDOCT average retinal nerve fibre layer thickness (RNFLT) measurements and the  
27 SDOCT RNFLT. 2) Analysis of the distribution of the rates of RNFLT change in TDOCT and  
28 synthesized SDOCT in the two treatments arms of the UKGTS was compared. A Cox model for  
29 predictors of time-to-incident VF progression was computed with the TDOCT and the synthesized  
30 SDOCT.

31 **Results:** The 95% limits of agreement between TDOCT and SDOCT were [26.64, -22.95], between  
32 synthesized SDOCT and SDOCT were [8.11, -6.73], and between SDOCT and SDOCT were [4.16, -  
33 4.04]. The mean difference in the rate of RNFL change between UKGTS treatment and placebo  
34 arms with TDOCT was 0.24 (p=0.11) and with synthesized SDOCT was 0.43 (p=0.0017). The hazard  
35 ratio for RNFLT slope in Cox regression modeling for time to incident VF progression was 1.09  
36 (95% CI 1.02 to 1.21) (p=0.035) for TDOCT and 1.24 (95% CI 1.08 to 1.39) (p=0.011) for synthesized  
37 SDOCT.

38 **Conclusions:** Image enhancement significantly improved the agreement of TDOCT RNFLT  
39 measurements with SDOCT RNFLT measurements. The difference, and its significance, in rates of  
40 RNFLT change in the UKGTS treatment arms was enhanced and RNFLT change became a stronger  
41 predictor of VF progression.

## 42 **Introduction**

43 Open-angle glaucoma is a progressive optic neuropathy in which retinal ganglion cell (RGC) axon  
44 loss, probably as a consequence of damage at the optic disc, causes a loss of vision,  
45 predominantly affecting the mid-peripheral visual field and in the 'macula vulnerability zone'[1].  
46 Glaucoma is the leading cause of irreversible blindness worldwide and the second major cause  
47 for blind registration in the UK[2,3]. The vision loss is associated with restricted mobility[4], falls  
48 and motor vehicle accidents[5]. Evaluating the rate of deterioration of the pathology is crucial in  
49 order to assess the risk of functional impairment and to establish sound treatment strategies.  
50 Therefore, accurately monitoring the efficacy of disease-modifying drugs in glaucoma therapy is  
51 of great importance. Clinically, standard automated perimetry (SAP) is employed to assess the  
52 status of the visual field (VF), whereas optical coherence tomography (OCT) is used as a surrogate  
53 measure to evaluate retinal ganglion cell (RGC) loss by measuring retinal nerve fibre layer (RNFL)  
54 thickness around the optic nerve head (ONH).

55 Evidence that imaging can identify progressive glaucomatous damage has been  
56 extensively reported in literature, recognising the potential of structural measures to support VF  
57 testing[18-25]. Medeiros et al.[26,27] address whether biomarkers, such as IOP and imaging  
58 measurements can be used as valid surrogate endpoints in clinical trials evaluating new therapies  
59 for glaucoma. They suggest that a valid surrogate endpoint must be able to predict a clinically  
60 relevant endpoint, such as loss of vision or decrease in quality of life. Moreover, the authors  
61 propose that the effect of a treatment on the surrogate endpoint must capture the effect of the  
62 treatment on the clinically relevant endpoint. Specifically, imaging biomarkers could potentially  
63 be used in combination with functional outcomes in composite endpoints in glaucoma trials,

64 overcoming weaknesses of using structural or functional endpoints separately. Studies should be  
65 designed and conducted in such a way that proper validation of potential biomarkers in glaucoma  
66 clinical trials could be demonstrated. Whereas spectral-domain (SD) and swept-source (SS)  
67 optical coherence tomography (OCT) are the state-of-the-art technologies for structural imaging  
68 of anatomy relevant to glaucoma, no large-scale clinical trials have yet employed SD or SS OCT to  
69 monitor glaucoma deterioration. The UK Glaucoma Treatment Study (UKGTS)[15] is the only  
70 glaucoma study to assess the vision-preserving efficacy of a disease-modifying drug with both VF  
71 and OCT outcomes. In the UKGTS, time-domain (TD) OCT was used as the imaging outcome since  
72 SD OCT (SDOCT), which offers better measurement precision, was not in widespread clinical use  
73 at the time of trial initiation. In the initial reports of the UKGTS, the rate of RNFL loss, measured  
74 with TD OCT, was unable to distinguish the treatment groups in the UKGTS and combining TD  
75 OCT and VF information did not improve detection of the treatment effect over the use of VF  
76 information alone[33]. This is most likely a result of the poor signal-to-noise ratio (SNR) and  
77 precision of TDOCT[23, 40].

78         Meanwhile, various methods for super resolution (SR) using convolutional neural  
79 networks (CNNs), such as generative adversarial networks (GANs), have been proposed to  
80 transform image quality and appearance[28-32]. In medical imaging, GANs have been  
81 successfully employed to address the ill-posed nature of cross-modal synthesis. For example,  
82 GANs have been proposed to predict computed tomography (CT) and positron emission  
83 tomography (PET) images from magnetic resonance imaging (MRI)[28-30]. Concerning signal  
84 enhancement, synthesis has been achieved at different resolution scales and by enforcing cycle-  
85 consistency, albeit not focusing on medical applications [31, 32]. These works may, however,

86 present important limitations for SR in medical imaging. First, due to the restricted view of GANs'  
87 spatial window, preservation of spatial smoothness and anatomical features in predictions is not  
88 always guaranteed. Second, single GAN predictions are characterized by spatial and intensity  
89 variability. Therefore, in order to extract robust anatomical quantifications from the output of  
90 GANs, principled schemes accounting for prediction uncertainty must be developed. This  
91 requires, for instance, probabilistic modelling of the uncertainty of the underlying signal  
92 distributions on distinct image parts, to preserve anatomical structures and account for spatial  
93 coherency.

94 This paper evaluates whether deep learning 'super resolution' techniques to 'learn'  
95 SDOCT images from TDOCT images can improve the signal-to-noise ratio of TD OCT and improve  
96 the performance of TD OCT to identify glaucomatous RNFL changes over time. The motivation  
97 for the work was to improve the image quality of the only existing OCT data set from a large-scale  
98 clinical trial in glaucoma to enable the further exploration of imaging endpoints in future clinical  
99 trials of glaucoma therapy[ref companion piece by editor].

100

## 101 **Methods**

102 The deep learning algorithm was trained and validated on paired TD and SD OCT images from  
103 one dataset ('RAPID') and then tested on the TD OCT images from the UKGTS.

104

## 105 **RAPID**

106 Eighty-two clinically stable glaucoma patients under standard treatment (intraocular pressure  
107 mean 14.0 mmHg [5<sup>th</sup> to 95<sup>th</sup> percentile 8.0 to 21.0 mmHg] and VF MD -4.17 dB [5<sup>th</sup> to 95<sup>th</sup>

108 percentile -14.22 to 0.88dB]) were recruited to a test–retest study. Seventy seven (148 eyes) of  
109 the participants recruited attended for up to 10 visits within a 3-month period, for a total of 1256  
110 patient-eye visits. This data set was taken to represent a ‘stable glaucoma’ cohort; assumptions  
111 made include that, over such a short length of time, no clinically meaningful changes in the VF or  
112 RNFL structure would occur and that the variability characteristics of the VF and RNFL  
113 measurements are similar to those seen in clinical practice over longer periods of time. The study  
114 was undertaken in accordance with good clinical practice guidelines and adhered to the  
115 Declaration of Helsinki. The study was approved by the North of Scotland National Research  
116 Ethics Service committee on 27 September 2013 (reference no.: 13/NS/0132) and NHS  
117 Permissions for Research was granted by the Joint Research Office at University College London  
118 Hospitals NHS Foundation Trust on 3 December 2013. All patients provided written informed  
119 consent before the screening investigations were carried out. Recruitment criteria were based  
120 on those for the UKGTS. Patients were required to have reproducible VF loss with corresponding  
121 damage to the ONH and no other condition that could lead to VF loss, be aged > 18 years and  
122 have a visual acuity of  $\geq 20/40$ , a refractive error within  $\pm 8$  dioptres and an IOP of  $\leq 30$  mmHg.  
123 The VF MD had to be better than  $-16$  dB in the worse eye and better than  $-12$  dB in the better  
124 eye. VF loss was defined as a reduction in sensitivity at two or more contiguous locations with  $p$   
125  $< 0.01$  loss or more, three or more contiguous locations with  $p < 0.05$  loss or more, or a 10-dB  
126 difference across the nasal horizontal midline at two or more adjacent locations in the total  
127 deviation plot. Participants attended approximately once a week for 10 visits, with VF testing and  
128 OCT imaging carried out twice at the first visit and once at each subsequent visit. VF testing was  
129 undertaken with the Humphrey Field Analyser™ (HFA) and OCT imaging was carried out using

130 Stratus TD OCT™ (Carl Zeiss Meditec Inc., Dublin, CA, USA) and Spectralis SD OCT (Heidelberg  
131 Engineering, Heidelberg, Germany) (software version 5.2.4). RAPID participants had slightly more  
132 advanced glaucoma (VF MD -4.17 compared to -2.65 dB) and lower IOP (14.0 compared to 19.0  
133 mmHg) than UKGTS participants. More details can be found elsewhere [33].

134

### 135 **UKGTS**

136 The UKGTS is a multicentre, randomized, double-masked, placebo-controlled trial assessing  
137 visual function preservation in newly diagnosed open-angle glaucoma (OAG) patients (trial  
138 registration number, ISRCTN96423140). 516 newly-diagnosed (previously untreated)  
139 participants with OAG were prospectively recruited at 10 UK centres between 2007 and 2010.

140 The observation period was 2 years, with subjects monitored by VF testing, quantitative imaging,  
141 optic disc photography and tonometry at 11 scheduled visits. ONH structure was monitored with  
142 Heidelberg Retina Tomograph at all study sites and with Stratus TD OCT™ (Carl Zeiss Meditec Inc.,  
143 Dublin, CA, USA) (software version 5.0) and GDxECC Nerve Fiber Analyzer (Carl Zeiss Meditec Inc.,  
144 Dublin, CA, USA) at study sites with those devices. With respect to the whole UKGTS cohort, the  
145 baseline mean IOP ( $\pm$ SD) was  $18.9 \pm 4$  mmHg in the better mean deviation (MD) eyes (median [IQR]  
146 MD -1.27 dB [-2.37, -0.19]) and  $19.9 \pm 4.6$  mmHg in the worse MD eyes (median [IQR] MD -3.30 dB  
147 [-5.60, -1.98]). The median (interquartile range) VF MD for all eligible eyes was -2.9 dB (-1.6 to -  
148 4.8 dB).

149 The participants were allocated randomly to receive the IOP-reducing prostaglandin analog  
150 latanoprost (0.005%) or placebo eye drops. The UKGTS, and the subsequent analysis of  
151 anonymized data in this study, adhered to the tenets of the Declaration of Helsinki and was

152 approved by local institutional review boards (Moorfields and Whittington Research Ethics  
153 Committee on June 1, 2006, ethics approval reference, 09/H0721/56). Study participants  
154 provided written informed consent. A total of 488 from 516 enrolled participants with post-  
155 baseline data were analysed in the trial (latanoprost, n=244; placebo, n=244). Out of those, a  
156 subset of 284 participants (143 participants in the placebo group and 141 participants in the  
157 latanoprost group) had adequate quality VF and OCT data, with > 6 months of follow-up, and five  
158 or more visits and with data for both VFs and OCT at the baseline visit. For eye-based analysis,  
159 the eye with the worse MD was used. VF deterioration was the primary end point in the trial;  
160 time to VF deterioration within 24 months. Deterioration (progression) analysis was performed  
161 in the Humphrey Field Analyser™ (HFA) II-i Guided Progression Analysis™ (GPA) software  
162 (version 5.1.1) (Carl Zeiss Meditec Inc., Dublin, CA, USA), a sensitive technique that considers  
163 changes at individual test locations in the visual field. Deterioration (progression) criteria and  
164 details of the trial design and trial outcome are published elsewhere[15,33]. In short, the time to  
165 VF deterioration was significantly longer in the treatment group than in the placebo group  
166 (adjusted hazard ratio, 0.44; 95% confidence interval, 0.28 to 0.69).

167

### 168 **Visual Field Measurements**

169 All VF tests were performed with the HFA II (or II-i) and the SITA standard 24-2 program. A reliable  
170 VF was one with a false-positive rate of < 15% and < 20% fixation losses (for fixation losses of >  
171 20%, reliability was based on the subjective judgement of the technician supervising the test and  
172 the clinician reading the test, including an assessment of the eye tracker trace). Unreliable tests  
173 were repeated, either on the same day (with a break of at least 30 minutes) or on a subsequent

174 occasion. The reference standard analysis for VF deterioration was that used for the outcome of  
175 the UKGTS and was undertaken with the HFA II-i GPA software (version 5.1.1)[15].

176

### 177 **Spectralis OCT Retinal Nerve Fiber Layer Measurement**

178 In the RAPID study, the circumpapillary RNFL thickness was measured with a 3.5 mm-diameter  
179 scan circle centred on the optic disc with the eye-tracking system activated with Spectralis SD-  
180 OCT Heidelberg Eye Explorer (Heidelberg Engineering, Heidelberg, Germany) (software version  
181 5.2.4). Automatic real-time (ART) function was activated, thereby allowing multiple frames, i.e.  
182 B-scans, to be averaged for speckle noise reduction.

183

### 184 **Stratus OCT Retinal Nerve Fiber Layer Measurement**

185 In the RAPID and the UKGTS, the fast RNFL 3.4 scan protocol was used to measure the  
186 parapapillary RNFL with TD Stratus OCT™ (Carl Zeiss Meditec Inc., Dublin, CA, USA) (software  
187 version 5.0). A scan circle of 3.4 mm in diameter consisting of 256 A-scans was positioned  
188 manually at the centre of the optic disc.

189 Right-hand orientation was used for documentation of clock hour measurements in  
190 SpectralisOCT and StratusOCT and RNFL measurements are provided as means (average RNFL  
191 around the ONH) and in clock-hour sectors.

192

### 193 **Imaging Analysis Protocol and Quality Control**

194 In the original UKGTS analysis, for TDOCT only, the images used followed the fast RNFL protocol:  
195 the OCT instrument software averages the measurements from three images acquired in quick

196 succession and a signal strength of  $\geq 7$  was required; images were retaken if necessary. Images  
197 of lower quality, or those with a software alert, were not included in the analyses. As a result,  
198 10,633 (21.3%) OCT scans were excluded in the original UKGTS analysis[40]. In the present  
199 analyses, for TD OCT in the UKGTS and SD and TD OCT in the RAPID, images were excluded only  
200 when our pre-processing algorithm failed; this was based on the success of an algorithm to  
201 estimate the retinal pigment epithelium (RPE) location (which is subsequently used to flatten the  
202 images, as the topology around the optic nerve head undulates). As a result, in the RAPID study,  
203 from 4,902 TD OCT scans, 257 (5.2%) were excluded. From 1,789 SD OCT scans, 68 (3.8%) were  
204 excluded. A patient with N TDOCT and M SDOCT can theoretically produce a maximum of  $N \times M$   
205 TD–SD OCT image pairs which can subsequently be used for the learning process on cross-modal  
206 synthesis. For the UKGTS TDOCT images, all the raw intensity OCT data were used, including each  
207 one of the three individual sequential ‘fast’ circular scans which are used for averaging, and  
208 images with any signal strength were accepted for application of our algorithm and further  
209 analysis. As a result, a total of 36,169 (31.6%) TDOCT individual scans failed the RPE detection  
210 algorithm. Note that patients were not excluded because of poor scan quality (as determined by  
211 the OCT software) since those scans could theoretically become scans with good quality after  
212 image enhancement. Analysis was based on participants who had 15 (3 x 5) or more raw images,  
213 i.e. five averaged images.

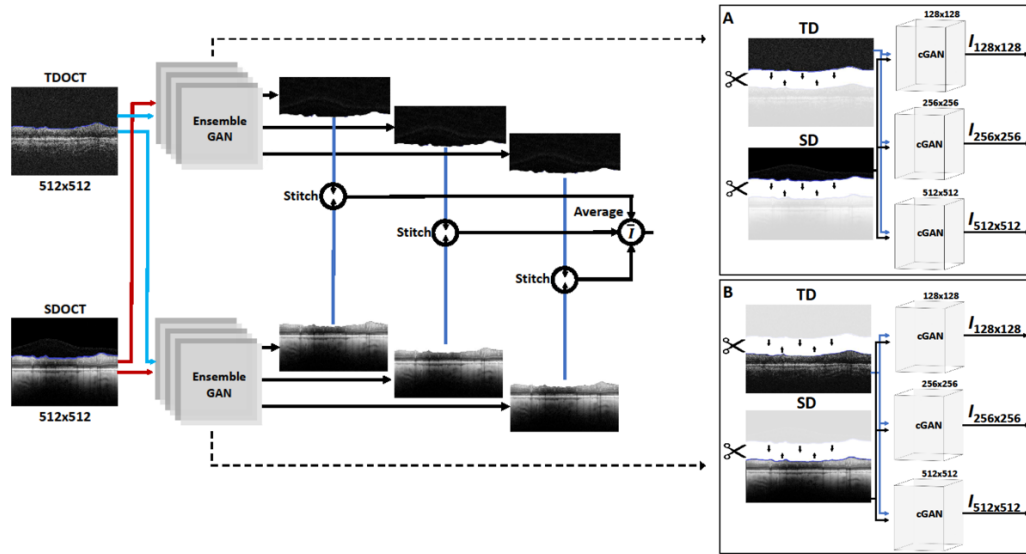
214

## 215 **Analysis**

216 **Image Synthesis.** We use cyclical GANs[32,34] to infer morphological descriptors from low to  
217 high quality anatomical information. OCT images have a very specific geometry where the

218 background, i.e. vitreous cavity, is clearly separated from the retinal layers at the ILM. Thus, we  
219 used image stitching, exploiting the ILM identification, to separate background from layer signal.  
220 Moreover, cycle GANs require a fixed window on which spatial filters and mappings are learned.  
221 However, since OCT signal and noise properties are characterized by different spatial scales, a  
222 modality transfer method based on a fixed spatial window might not be able to capture all the  
223 necessary spatial information needed for synthesis. This reduces the chance for cross-modal  
224 distributions to share supports in latent space. To address this problem, we propose an ensemble  
225 of spatially coherent cycle GANs[32] to learn the TDOCT-to-SDOCT mapping and to translate a  
226 TDOCT into a synthesized SDOCT image. The scheme is the following. Each GAN is trained by  
227 employing a different spatial window size: 128 x 128, 256 x 256 or 512 x 512, learning a mapping  
228 from the observed TDOCT image  $I_{TD}$  and random noise vector  $\mathbf{z}$ , to the target SDOCT image  $I_{SD}$ ,  
229  $G: \{I_{TD}, \mathbf{z}\} \rightarrow I_{SD}$ . As a result, we train six GANs: three with background pairs and three with retinal  
230 layer pairs. The synthesized backgrounds and layers are stitched back according to the window  
231 size, i.e.  $I_{128 \times 128}$ ,  $I_{256 \times 256}$ ,  $I_{512 \times 512}$ , and the average synthesized stitched image  $\bar{I}$  is obtained. To  
232 preserve the morphological correlation between training pairs, cycle GANs were trained with  
233 windows centered at the same geometrical location in both pairs. This deep learning technique  
234 is based on learning the representation between TD and SD OCT using 24,792 paired examples.  
235 The transfer mapping is learned in an independent dataset, i.e. the RAPID dataset, which contains  
236 pairs of both modalities, and the method is applied to the UKGTS dataset, enhancing the TD OCT  
237 images via quality transfer from SD OCT. TD OCT images are converted to 'synthesized SD OCT'  
238 images and segmented via an ensemble of GANs: for each TD OCT, we produce three SD OCT  
239 candidates. Fig. 1 shows the proposed framework for OCT synthesis via the ensemble of GANs.

240 The final RNFL segmentation is obtained on the average synthesized image of the segmented SD  
241 OCT candidates from each of the three GANs in the ensemble via the effective Bayesian label-  
242 propagation of multi-atlas segmentation (MAS)[36]. For segmentation, we adopted the layer  
243 segmentation model of Mayer et al.[37]. For label fusion of the three segmented synthesized SD  
244 OCT candidates, we used, as atlases, their segmented RNFL sections and the original TD OCT RNFL  
245 segmentation. We registered the retinal layers of the atlases, using the method described by Du  
246 *et al.* [38], in the average synthesized image (average of three SD OCT candidates). The Spectralis  
247 SD OCT images were segmented with the same software as that we used for the ‘synthesized SD  
248 OCT’ images. The intuition is that if we can produce realistic SD OCT images, an off-the-shelf  
249 segmentation model should output the same RNFL thickness as obtained with the original data.  
250 Note that the segmentation model of Mayer et al.[37] failed in segmenting TDOCT images. As a  
251 result, the original StratusOCT segmentation was used for TDOCT images. The technical details  
252 of the method are described in Lazaridis et al.[41].



253

Figure 1: SDOCT synthesis via ensemble of GANs. Box A: Backgrounds are painted black. Box B: Three GANs are trained with layer pairs. Synthesized images are stitched back with the backgrounds and the average synthesized stitched image is obtained. Separation of layers and background is illustrated with scissors.

254

255 **Statistical Analysis and Evaluation.** We quantified the quality improvement of the ‘synthesized

256 SD OCT’ images over the original TD OCT images in both the RAPID and UKGTS data sets. Fig. 2

257 shows an example of a SDOCT image synthesized from a TDOCT image. Fig. 2a and Fig. 2b

258 constitute the original TDOCT-SDOCT pair of images, whereas Fig. 2c is the synthesized SDOCT

259 after modality transfer and synthesis. To compare the performance of the Cox models, i.e. Cox

260 model before and after TDOCT image enhancement, we calculate the rank-based Somers’ D

261 between predicted risk scores and observed survival times. We compare the rankings of rate of

262 RNFL loss and time-to-VF progression per patient across the dataset and we assess their

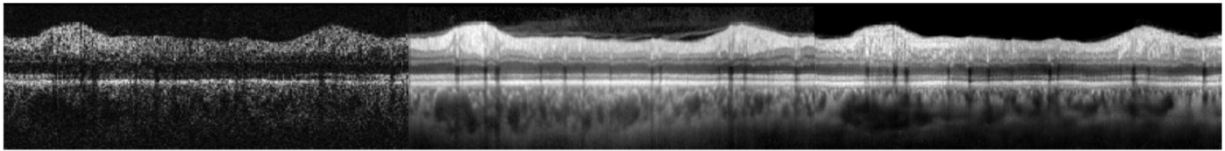
263 agreement. Somers’ D takes values between -1 when all ranking pairs disagree and 1 when all

264 pairs agree. To estimate the standardized effect size for the same population before and after

265 TDOCT image enhancement, we calculate Cohen’s D using the difference in the rates of loss

266 between the treatment groups. Although there are no reference values for Cohen’s standardized

267 effect size measures,  $d = 0.2, 0.5$  and  $0.8$  provide a conventional reference frame, corresponding  
268 to small, medium and large effects [43].



269 (a) TDOCT (b) SDOCT (c) Synthesized SDOCT

Figure 2: OCT synthesis results via fusion of GANs. (a) and (b) illustrate a pair of TDOCT and SDOCT images. (c) Synthesized SDOCT from (a).

270

271 RAPID data set: we compared the agreement of the average RNFL thickness derived from i) the  
272 Stratus TD OCT software and ii) the ‘synthesized SDOCT’ (described above) with the paired  
273 Spectralis SD OCT average RNFL thickness with Bland Altman plots. To give context, we also  
274 present the agreement between SD OCT RNFL thickness measurements acquired on different  
275 days – this represents the ‘ceiling’ one would expect to see if synthesized SD OCT images were  
276 exactly the same as real SD OCT images.

277 UKGTS data set: we compare the ability of the rate of RNFL loss measured with Stratus TD OCT  
278 and synthesized SD OCT to distinguish the treatment arms of the trial (Mann Whitney test). The  
279 effect size is estimated with Cohen’s  $D$ . We also present the respective strength of association of  
280 the rate of RNFL change with time to VF progression in a Cox proportional hazards model.

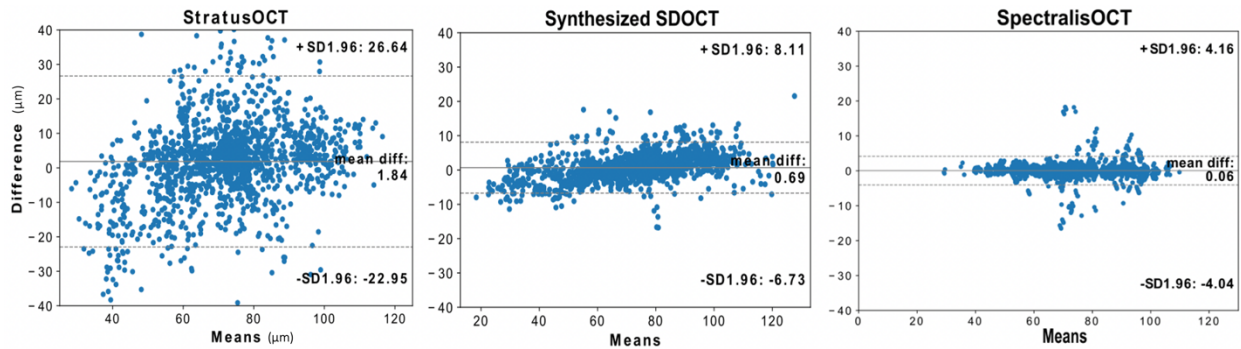
281

## 282 Results

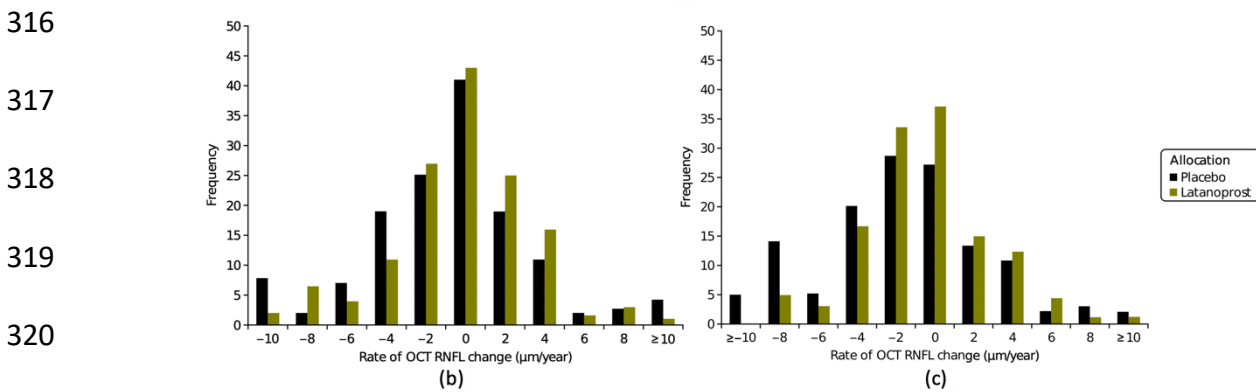
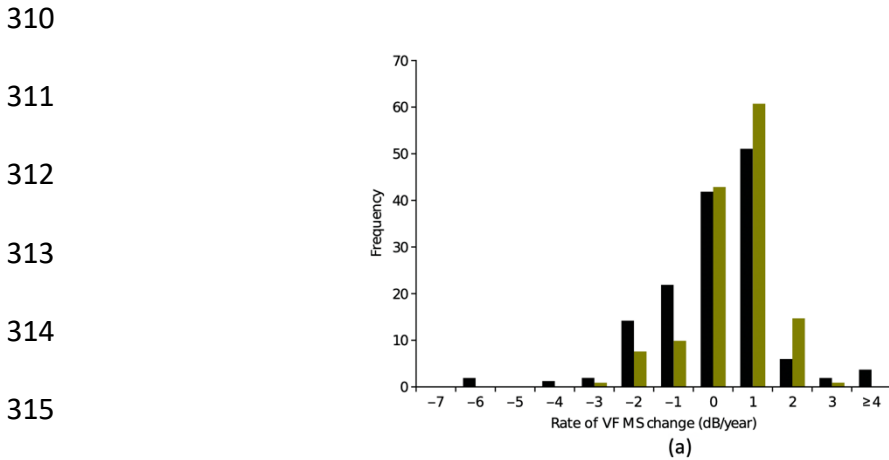
283 Test-retest variability, summarized by the standard deviation of repeat measurements over the  
284 first three visits across all subjects of the RAPID study, was lower for the Synthesized SDOCT than  
285 for the original TDOCT data (Table 1). Table 1 also shows the 95% limits of agreement (LOA) and

286 the mean difference between RNFL measurements. The 95% limits of agreement between TDOCT  
287 and SDOCT were [26.64, -22.95], between synthesized SDOCT and SDOCT were [8.11, -6.73], and  
288 between SDOCT and SDOCT were [4.16, -4.04]. Fig. 3 illustrates the corresponding Bland-Altman  
289 agreement plots of the RNFL measurements made from the segmented synthesized OCT images  
290 with respect to the 'ground truth' Spectralis SD OCT RNFL measurements derived with the same  
291 segmentation algorithm (RAPID data set). Table 2 presents the mean and the range of RNFL loss  
292 rates for TDOCT and synthesized SDOCT images. Table 3 and Table 4 illustrate the Cox  
293 proportional hazards model fitted to the time to VF progression for TD OCT and synthesized SD  
294 OCT. The hazard ratio for RNFLT slope in Cox regression modeling for time to incident VF  
295 progression was 1.09 (95% CI 1.02 to 1.19) ( $p=0.035$ ) for TDOCT and 1.24 (95% CI 1.11 to 1.39)  
296 ( $p=0.011$ ) for synthesized SDOCT. Fig. 4 illustrates the VF mean sensitivity (MS) change in decibels  
297 per year and the distribution of rate of RNFL thickness change for the subset of UKGTS  
298 participants with OCT images. Fig. 4b is generated from the original TD OCT whereas Fig. 4c from  
299 the synthesized SDOCT data. The placebo group had faster rates of deterioration than the  
300 latanoprost group in both cases. For the original TD OCT UKGTS data, the difference in  
301 distribution of slopes was not statistically significant (Mann-Whitney U Test,  $p = 0.08$ ). For the  
302 synthesized SD OCT, the difference was statistically significant (Mann-Whitney U Test,  $p =$   
303  $0.0017$ ). Table 5 illustrates the corresponding effect sizes (Cohen's D), with confidence intervals.  
304 It can be seen that Cohen's D for synthesized SD OCT is closer to Cohen's D for VFs than that for  
305 TD OCT, indicating a modest improvement in effect size. Table 6 compares the predictive power  
306 of the two Cox models; we calculate the rank order statistic Somers' D with confidence limits[42].

307 It can be observed that Somer's D is higher for the Cox model with synthesized SD OCT, indicating  
 308 a stronger predictive power between the rankings of predicted risk and time-to-VF progression.



309  
 Figure 3: Bland-Altman plots on the agreement between time domain and synthesized spectral domain OCT RNFL measurements versus the 'real' spectral domain OCT RNFL measurements on the RAPID dataset. The proposed method leads to significantly better agreement.



310  
 311  
 312  
 313  
 314  
 315  
 316  
 317  
 318  
 319  
 320  
 Figure 4: (a) Distribution of the rate of VF mean sensitivity (MS) change in decibels per year for the subset of UKGTS participants with OCT images (placebo, n = 131 participants; latanoprost, n = 127 participants). Bottom: Distribution of the rate of OCT RNFL thickness change for the subset of UKGTS participants with OCT images. (b) Original UKGTS TDOCT data (placebo, n = 131 participants; latanoprost, n = 127 participants). (c) Synthesized UKGTS SDOCT data (placebo, n = 131 participants; latanoprost, n = 127 participants).

Table 1: Limits of agreement and mean difference between time domain, synthesized spectral domain, 'real' spectral domain OCT RNFL measurements versus the 'real' spectral domain OCT RNFL measurements. The mean SD gives the standard deviation of the first three test-retest visits for both eyes. SDOCT = spectral domain optical coherence tomography; TDOCT = time domain optical coherence tomography

321

322

323

324

Method	Synthesized SDOCT	StratusOCT	SpectralisOCT
95% LOA	<b>[8.11, -6.73]</b>	[26.64, -22.95]	[4.16, -4.04]
Mean Diff.	<b>0.69</b>	1.84	0.06
Mean SD	<b>1.29</b>	2.67	0.77

Table 2: Comparison of rate of RNFL change in Stratus OCT and synthesized spectral domain OCT in the UKGTS data set. The significance of the difference between treatment and placebo progression rates was calculated with the Mann Whitney U test. SDOCT = spectral domain optical coherence tomography; TDOCT = time domain optical coherence tomography

325

Method	StratusOCT		Synthesized SD OCT	
	Treatment	Placebo	Treatment	Placebo
Mean (SD) ( $\mu\text{m}/\text{year}$ )	-0.15 (3.971)	-0.39 (4.139)	-0.83 (2.6116)	-1.26 (2.6720)
Diff. in mean rate (95% CI)	0.24 (-0.837 to 0.672)		0.43* (0.0279 to 0.8321)	
<i>p</i> -value	0.08		0.0017	

326

Table 3: Cox proportional hazards model for time to incident VF progression in the UKGTS with the original TD OCT images. Note  $b$  = regression coefficient, Wald statistic =  $(b/SE)^2$ ,  $p$  =  $p$ -value associated with the Wald statistic and  $\text{Exp}(b)$  = the hazard ratio. (placebo,  $n = 131$  participants; latanoprost,  $n = 127$  participants).

327

Covariate	$b$	SE	Wald	$p$	$\text{Exp}(b)$	95% CI of $\text{Exp}(b)$
Age	0.018	0.014	1.748	0.186	1.018	0.991 to 1.045
Allocation	-0.770	0.287	7.226	0.007	0.463	0.264 to 0.812
Baseline IOP	0.050	0.029	2.972	0.085	1.051	0.993 to 1.113
Baseline VF MD	0.086	0.048	3.123	0.077	1.089	0.991 to 1.198
OCT RNFL slope	0.086	0.041	4.430	0.035	1.089	1.031 to 1.412
Disc haemorrhage	0.576	0.283	4.143	0.042	1.779	1.022 to 3.099

328

329

Table 4: Cox proportional hazards model for time to incident VF progression in the UKGTS with the synthesized SD OCT images. Note b = regression coefficient, Wald statistic =  $(b/SE)^2$ , p = p-value associated with the Wald statistic and Exp(b) = the hazard ratio. (placebo, n = 131 participants; latanoprost, n = 127 participants).

330

Covariate	b	SE	Wald	p	Exp(b)	95% CI of Exp(b)
Age	0.021	0.009	5.444	0.113	1.021	0.922 to 1.152
Allocation	-0.586	0.195	9.030	0.001	0.608	0.315 to 0.901
Baseline IOP	0.106	0.089	1.418	0.109	1.111	0.811 to 1.429
Baseline VF MD	0.041	0.022	3.473	0.062	1.041	0.883 to 1.312
OCT RNFL slope	0.218	0.008	7.425	0.011	1.244	1.105 to 1.394
Disc haemorrhage	0.251	0.109	5.302	0.027	1.285	1.126 to 2.836

331

Table 5: Comparison of treatment groups effect size for each modality. Cohen’s D is calculated as measure of parametric group testing, measuring the effect size. SDOCT = spectral domain optical coherence tomography; TDOCT = time domain optical coherence tomography; CI = confidence interval.

332

Modality	Synthesized SDOCT	StratusOCT	Visual Fields
Cohen’s D	0.256	0.223	0.491
95% CI	[0.126, 0.487]	[0.076, 0.535]	[0.289, 0.652]
p-value	0.03	0.05	0.002

333

334

Table 6: Comparison of the predictive power of Cox models. Somers’ D is calculated as measure of the ordinal predictive power of each model. Confidence intervals and p-values for the predictive powers of each model are also computed. SDOCT = spectral Table 6: Comparison of the predictive power of Cox models. Somers’ D is calculated between predicted risk scores and observed survival times. Confidence

335

336

Model	Synthesized SDOCT	StratusOCT
Somers’ D	0.326	0.289
95% CI	[0.113, 0.581]	[0.129, 0.448]
p-value	0.019	0.009

337

338

339

340

341

## 342 Discussion

343 In this work, we demonstrate that a super resolution deep learning method applied to TD OCT  
344 images significantly improves the signal-to-noise ratio of the images, as quantified by the  
345 agreement of segmented RNFL thickness measurements with SD OCT measurements, and  
346 significantly reduces test-retest variability (Table 1, Figure 3) and the improves the ability of rates  
347 of RNFL loss to separate the treatment arms of the UKGTS. When the rate of RNFL loss in the  
348 UKGTS data set is calculated from the 'synthesized SD OCT' images (Table 2), the difference in  
349 RNFL slope measurements is able to distinguish the treatment groups (Mann-Whitney U Test,  $p$   
350 = 0.0017).

351 The ensemble of GANs approach produced segmented RNFL thickness values more consistent  
352 with the ground truth SD OCT values than the TD OCT, as demonstrated by narrower limits of  
353 agreement (Figure 3, Table 1), and reduced the test retest variability in the measurements by  
354 half, as demonstrated by the smaller standard deviation of repeat measurements (Table 1). The  
355 Bland–Altman plots revealed proportional biases in the evaluation of agreement between SD OCT  
356 and TD OCT, and between SD OCT and synthesized SD OCT RNFL measurements in the RAPID  
357 study data set, suggesting that there may be a calibration difference, possibly related to the  
358 inherent characteristics of the OCT instruments. These findings are in agreement with Leung et  
359 al.[22], where the same proportional bias was reported between Cirrus SD-OCT and Stratus TD  
360 OCT.

361 When the super resolution method was applied to an independent test data set, from the UKGTS,  
362 the better separation of the treatment arms evidenced the data quality improvement. The  
363 analysis of the capability of TD OCT images to distinguish the UKGTS treatment arms showed

364 that, although the rate of RNFLT loss was faster in the placebo-treated eyes, the difference from  
365 the latanoprost-treated eyes did not reach statistical significance (Table 2; Figure 4b). In contrast,  
366 the same analysis with the synthesized SD OCT images demonstrated a statistically significant  
367 difference between treatment and placebo progression rates (MannWhitney U Test,  $p = 0.0017$   
368 (Table 2; Figure 4c). The difference between treatment groups in the rate of RNFL thinning  
369 (synthesized SD OCT) is closer to the difference between groups for the rate of VF MD  
370 deterioration (Figure 4) than for the TD OCT analysis (Table 5). Our analysis further illustrates  
371 that the SD OCT imaging of RNFL may provide a sufficiently high precision for longitudinal  
372 assessment of RNFL changes, as low measurement variability is a prerequisite for detecting  
373 change during longitudinal analysis (Table 6); improving the longitudinal SNR.

374 Further evidence for the improvement in data quality comes from the Cox proportional hazards  
375 model which was fitted to the time to VF progression original UKGTS data (Table 3). This  
376 demonstrated that treatment allocation, the occurrence of a disc haemorrhage during follow-up  
377 (either eye) and the rate of TD OCT RNFL change were significantly associated with survival. Pre-  
378 treatment IOP and baseline VF MD approached statistical significance ( $p$  between 0.077 and  
379 0.085); the overall model fit was significant ( $p = 0.0007$ ). The same model was fitted after TD OCT  
380 signal enhancement (Table 4) and showed a greater level of significance in the overall fit of the  
381 model ( $p = 0.0001$ ). The significance of the association of treatment allocation, occurrence of a  
382 disc haemorrhage during follow-up (either eye) and rate of OCT RNFL change with time to VF  
383 deterioration also improved, with a larger hazard ratio for RNFL change.

384

385

386 **Study weaknesses and further work**

387 In this work, we have used randomised controlled trial data coming from the first large scale  
388 glaucoma trial with OCT data, i.e. the UKGTS. We further presented a super resolution approach  
389 to translate a TD OCT image into a synthesized SD OCT image. The image-enhancement approach  
390 is based on state-of-the-art image synthesis and semi-automated segmentation of the resulting  
391 synthesized SDOCT images, integrating label fusion and deep learning. The proposed  
392 methodology appears robust and flexible both in terms of architecture and label fusion. Since the  
393 training dataset is large and of high resolution, training of each individual model takes a lot of  
394 time, making the method computationally expensive for training. This, limitation, is however a  
395 negligible problem in practice as the algorithm can be run offline. As the agreement of  
396 synthesized SD OCT RNFL measurements with real SD OCT RNFL measurement did not reach the  
397 level of agreement indicated by the limits of agreement for repeat real SD OCT RNFL  
398 measurements, this study likely underestimates the potential utility of SD OCT imaging in future  
399 trials.

400 The TD OCT images were segmented with the proprietary instrument software and the real and  
401 synthesized SD OCT images with a publicly-available algorithm; we did not have access to the  
402 proprietary algorithm to apply to SD OCT images and the publicly-available algorithm failed on  
403 the TD OCT images. Therefore, the results we report relate to comparisons of the compound  
404 ‘image + segmentation algorithm’.

405

406 Future work will focus on combining SD OCT RNFL rates of change to VF rates of change, in a  
407 similar way as that done for TD OCT[40], to see whether the addition of the imaging data

408 improves study power over the use of VF data alone. The motivation is that although the signal-  
409 to-noise ratio in the TD OCT UKGTS data is too poor to draw conclusions with respect to disease  
410 deterioration, the synthesized SD OCT data provided some evidence that imaging outcomes  
411 capture the effect of treatment on the VF outcome.

412

413

414

415

416

417

418

## 419 **Conclusion**

420 In clinical trials with a vision function outcome, variability in measurements results in the  
421 requirement for large numbers of patients observed over long intervals. As a result, new  
422 beneficial treatments to patients may be delayed and may not be evaluated as trials become  
423 more costly. It is well established that imaging measurements of structural damage to the ONH  
424 are associated with VF loss in glaucoma. Furthermore, imaging measurements are often  
425 considered more precise than VF measurements, making them attractive as potential surrogate  
426 outcomes for clinical trials and clinical practice. The OCT data available in the UKGTS were from  
427 the TD OCT, with poor signal-to-noise characteristics. Previous analysis of the OCT data failed to  
428 distinguish the treatment arms[40]. Here, we show that a super resolution deep learning method  
429 was able to considerably improve data quality, demonstrated by better agreement of RNFL

430 measurements from synthesized SD OCT images, compared with their source TD OCT images,  
431 with RNFL measurements from actual SD OCT images. When applied to an independent data set  
432 from the UKGTS, the data quality improved to the extent that imaging measurements were able  
433 distinguish treatment groups. These findings suggest that a benefit to trial power can be achieved  
434 by a) further increase the resolution of SDOCT using SR methods b) ensemble methods to  
435 segment more efficiently SDOCT images.

436

437

438

439

440

#### 441 **References**

- 442 1. Hood DC, Raza AS, de Moraes CG, Liebmann JM, Ritch R. Glaucomatous damage of the macula.  
443 Prog Retin Eye Res. 2013;32:1-21. doi:10.1016/j.preteyeres.2012.08.003
- 444 2. Resnikoff S Pascolini D Etya'ale D et al. Global data on visual impairment in the year 2002. Bull  
445 World Health Organ. 2004; 82: 844-851
- 446 3. Bunce C, Wormald R. Causes of blind certifications in England and Wales: April 1999–March  
447 2000. Eye. 2008; 22: 905-911
- 448 4. Friedman DS, Freeman E, Munoz B, Jampel HD, West SK. Glaucoma and mobility performance:  
449 the Salisbury Eye Evaluation Project. Ophthalmology. 2007; 114: 2232-2237
- 450 5. Haymes SA, Leblanc RP, Nicolela MT, Chiasson LA, Chauhan BC. Risk of falls and motor vehicle  
451 collisions in glaucoma. Invest Ophthalmol Vis Sci. 2007; 48: 1149-1155

452 deterioration. The AGIS Investigators. *Am J Ophthalmol*. 2000; 130: 429-440

453 7. Garway-Heath DF, Crabb DP, Bunce C, et al. Latanoprost for open-angle glaucoma (UKGTS): a  
454 randomised, multicentre, placebo-controlled trial. *Lancet*. 2015;385(9975):1295-1304.  
455 doi:10.1016/S0140-6736(14)62111-5

456 8. Leung CK, Yu M, Weinreb RN, Lai G, Xu G, Lam DS. Retinal nerve fiber layer imaging with  
457 spectral-domain optical coherence tomography: patterns of retinal nerve fiber layer progression.  
458 *Ophthalmology*. 2012;119(9):1858-1866. doi:10.1016/j.ophtha.2012.03.044

459 9. Leung CK, Ye C, Weinreb RN, Yu M, Lai G, Lam DS. Impact of age-related change of retinal  
460 nerve fiber layer and macular thicknesses on evaluation of glaucoma progression. *Ophthalmology*  
461 2013;120:2485–92.

462 10. Leung CK. Diagnosing glaucoma progression with optical coherence tomography. *Curr Opin*  
463 *Ophthalmol* 2014;25:104–11.

464 11. Leung CK, Cheung CY, Lin D, Pang CP, Lam DS, Weinreb RN. Longitudinal variability of optic  
465 disc and retinal nerve fiber layer measurements. *Invest Ophthalmol Vis Sci* 2008;49:4886–92

466 12. Leung CK, Chiu V, Weinreb RN, Liu S, Ye C, Yu M, et al. Evaluation of retinal nerve fiber layer  
467 progression in glaucoma: a comparison between spectral-domain and time-domain optical  
468 coherence tomography. *Ophthalmology* 2011;118:1558–62

469 13. Leung CK, Cheung CY, Weinreb RN, Qiu Q, Liu S, Li H, et al. Retinal nerve fiber layer imaging  
470 with spectral-domain optical coherence tomography: a variability and diagnostic performance  
471 study. *Ophthalmology* 2009;116:1257–63.

472 14. Daga FB, Gracitelli CPB, Diniz-Filho A, et al Is vision-related quality of life impaired in patients  
473 with preperimetric glaucoma? *British Journal of Ophthalmology* 2019;103:955-959.

- 474 15. Felipe A. Medeiros, Linda M. Zangwill, Luciana M. Alencar, Christopher Bowd, Pamela A.  
475 Sample, Remo Susanna, Robert N. Weinreb; Detection of Glaucoma Progression with Stratus OCT  
476 Retinal Nerve Fiber Layer, Optic Nerve Head, and Macular Thickness Measurements. *Invest.*  
477 *Ophthalmol. Vis. Sci.* 2009;50(12):5741-5748. doi: 10.1167/iovs.09-3715.
- 478 16. Medeiros FA. Biomarkers and surrogate endpoints in glaucoma clinical trials. *Br J Ophthalmol.*  
479 2015;99(5):599–603. doi:10.1136/bjophthalmol-2014-305550
- 480 17. Medeiros FA. Biomarkers and Surrogate Endpoints: Lessons Learned From Glaucoma. *Invest*  
481 *Ophthalmol Vis Sci.* 2017;58(6):BIO20–BIO26. doi:10.1167/iovs.17-21987
- 482 18. Nie D, Trullo R, Lian J, et al. Medical Image Synthesis with Context-Aware Generative  
483 Adversarial Networks. *Med Image Comput Comput Assist Interv.* 2017;10435:417-425.  
484 doi:10.1007/978-3-319-66179-7\_48
- 485 19. Wolterink JM, Dinkla AM, Savenije MHF, Seevinck PR, van den Berg CAT, Išgum I. Deep MR to  
486 CT synthesis using unpaired data. *Med Image Comput Comput Assist Interv.* 2017;10557:14–23.  
487 10.1007/978-3-319-68127-6\_2
- 488 20. Ben-Cohen A, Klang E, Raskin SP, Amitai MM, Greenspan H. Virtual PET Images from CT  
489 Data Using Deep Convolutional Networks: Initial Results. *Med Image Comput Comput Assist*  
490 *Interv.* 2017;10557:49-57. 10.1007/978-3-319-68127-6\_6
- 491 21. Wang TC, Liu, MY, et al. High-Resolution Image Synthesis and Semantic Manipulation with  
492 Conditional GANs. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.  
493 2018:8798-8807, doi: 10.1109/CVPR.2018.00917.

494 22. Zhu, JY, Park, T., Isola, P., Efros, AA: Unpaired Image-to-Image Translation Using Cycle-  
495 Consistent Adversarial Networks. IEEE International Conference on Computer Vision (ICCV).  
496 2017:2242-2251, doi: 10.1109/ICCV.2017.244.

497 23. Garway-Heath DF, Quartilho A, Prah P, Crabb DP, Cheng Q, Zhu H. Evaluation of Visual Field  
498 and Imaging Outcomes for Glaucoma Clinical Trials (An American Ophthalmological Society  
499 Thesis). Trans Am Ophthalmol Soc. 2017;115:T4.

500 24. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, and Bengio  
501 Y. Generative adversarial nets. In Proceedings of the 27th International Conference on Neural  
502 Information Processing Systems, 2014;2:2672-2680.

503 25. Zhu JY, Park T, Isola P, Efros, AA: Unpaired Image-to-Image Translation Using Cycle-Consistent  
504 Adversarial Networks. IEEE International Conference on Computer Vision (ICCV).  
505 2017:2242:2251.

506 26. Sabuncu, M.R., Yeo, B.T.T., Van Leemput, K., Fischl, B., Golland, P.: A Generative Model for  
507 Image Segmentation Based on Label Fusion. IEEE Trans. Med. Imaging. 2010;29:1714–1729.

508 27. Mayer MA, Hornegger J, Mardin CY, Tornow RP: Retinal Nerve Fiber Layer Segmentation  
509 on FD-OCT Scans of Normal Subjects and Glaucoma Patients. Biomed. Opt. Express.  
510 2010;1:1358-1383.

511 28. Du X, Gong L et al.: Non-rigid Registration of Retinal OCT Images Using Conditional Correlation  
512 Ratio. Med Image Comput Comput Assist Interv. 2017:159–167.

513 29 Leung CK, Carol Yim-lui Cheung, Robert N. Weinreb, Gary Lee, Dusheng Lin, Chi PP, Dennis SC  
514 Lam; Comparison of Macular Thickness Measurements between Time Domain and Spectral

515 Domain Optical Coherence Tomography. *Invest. Ophthalmol. Vis. Sci.* 2008;49(11):4893-4897.  
516 doi: 10.1167/iovs.07-1326.

517 30. Garway-Heath DF, Zhu H, Cheng Q, Morgan K, Frost C, Crabb DP, et al. Combining optical  
518 coherence tomography with visual field data to rapidly detect disease progression in glaucoma:  
519 a diagnostic accuracy study. *Health Technol Assess* 2018;22(4)

520 31. Lazaridis G, Lorenzi M, Ourselin S, Garway-Heath DF. Enhancing OCT Signal by Fusion of GANs:  
521 Improving Statistical Power of Glaucoma Clinical Trials. *Med Image Comput Comput Assist Interv.*  
522 2019;11764;1–9. doi: 10.1007/978-3-030-32239-7\_1

523 32. Newson RB. Comparing the Predictive Powers of Survival Models Using Harrell’s C or Somers’  
524 D. *The Stata Journal.* 2010;10(3):339-358. doi:10.1177/1536867X1001000303

525 33. Cohen J. *Statistical power analysis for the behavioral sciences.* Hillsdale, N.J.: L. Erlbaum  
526 Associates; 1988