OCT signal enhancement with deep learning

Georgios Lazaridis,[1,2,4] Marco Lorenzi,[3] Jibran Mohamed-Noriega,[1,5] Soledad Aguilar-Munoa,[1] Katsuyoshi Suzuki,[1] Hiroki Nomoto,[1] Sebastien Ourselin,[4] David F. Garway-Heath,[1] on behalf of the United Kingdom Glaucoma Treatment Study Investigators

[1]NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
[2]Centre for Medical Image Computing, University College London, London, United Kingdom
[3]Université Côte d'Azur, Inria Sophia Antipolis, Epione Research Project, France
[4]School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom
[5]Departamento de Oftalmología, Hospital Universitario, UANL, México

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

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

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

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

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

Results: The 95% limits of agreement between TDOCT and SDOCT were [26.64, -22.95], between synthesized SDOCT and SDOCT were [8.11, -6.73], and between SDOCT and SDOCT were [4.16, -4.04]. The mean difference in the rate of RNFL change between UKGTS treatment and placebo arms with TDOCT was 0.24 (p=0.11) and with synthesized SDOCT was 0.43 (p=0.0017). The hazard ratio for RNFLT slope in Cox regression modeling for time to incident VF progression was 1.09 (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 SDOCT.

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

Introduction

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

Evidence that imaging can identify progressive glaucomatous damage has been extensively reported in literature, recognising the potential of structural measures to support VF testing[18-25]. Medeiros et al.[26,27] address whether biomarkers, such as IOP and imaging measurements can be used as valid surrogate endpoints in clinical trials evaluating new therapies for glaucoma. They suggest that a valid surrogate endpoint must be able to predict a clinically relevant endpoint, such as loss of vision or decrease in quality of life. Moreover, the authors propose that the effect of a treatment on the surrogate endpoint must capture the effect of the treatment on the clinically relevant endpoint. Specifically, imaging biomarkers could potentially be used in combination with functional outcomes in composite endpoints in glaucoma trials, overcoming weaknesses of using structural or functional endpoints separately. Studies should be designed and conducted in such a way that proper validation of potential biomarkers in glaucoma clinical trials could be demonstrated. Whereas spectral-domain (SD) and swept-source (SS) optical coherence tomography (OCT) are the state-of-the-art technologies for structural imaging of anatomy relevant to glaucoma, no large-scale clinical trials have yet employed SD or SS OCT to monitor glaucoma deterioration. The UK Glaucoma Treatment Study (UKGTS)[15] is the only glaucoma study to assess the vision-preserving efficacy of a disease-modifying drug with both VF and OCT outcomes. In the UKGTS, time-domain (TD) OCT was used as the imaging outcome since SD OCT (SDOCT), which offers better measurement precision, was not in widespread clinical use at the time of trial initiation. In the initial reports of the UKGTS, the rate of RNFL loss, measured with TD OCT, was unable to distinguish the treatment groups in the UKGTS and combining TD OCT and VF information did not improve detection of the treatment effect over the use of VF information alone[33]. This is most likely a result of the poor signal-to-noise ratio (SNR) and precision of TDOCT[23, 40].

Meanwhile, various methods for super resolution (SR) using convolutional neural networks (CNNs), such as generative adversarial networks (GANs), have been proposed to transform image quality and appearance[28-32]. In medical imaging, GANs have been successfully employed to address the ill-posed nature of cross-modal synthesis. For example, GANs have been proposed to predict computed tomography (CT) and positron emission tomography (PET) images from magnetic resonance imaging (MRI)[28-30]. Concerning signal enhancement, synthesis has been achieved at different resolution scales and by enforcing cycle-consistency, albeit not focusing on medical applications [31, 32]. These works may, however, present important limitations for SR in medical imaging. First, due to the restricted view of GANs' spatial window, preservation of spatial smoothness and anatomical features in predictions is not always guaranteed. Second, single GAN predictions are characterized by spatial and intensity variability. Therefore, in order to extract robust anatomical quantifications from the output of GANs, principled schemes accounting for prediction uncertainty must be developed. This requires, for instance, probabilistic modelling of the uncertainty of the underlying signal distributions on distinct image parts, to preserve anatomical structures and account for spatial coherency.

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

Methods

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

RAPID

Eighty-two clinically stable glaucoma patients under standard treatment (intraocular pressure mean 14.0 mmHg [5th to 95th percentile 8.0 to 21.0 mmHg] and VF MD −4.17 dB [5th to 95th percentile -14.22 to 0.88dB]) were recruited to a test–retest study. Seventy seven (148 eyes) of the participants recruited attended for up to 10 visits within a 3-month period, for a total of 1256 patient-eye visits. This data set was taken to represent a 'stable glaucoma' cohort; assumptions made include that, over such a short length of time, no clinically meaningful changes in the VF or RNFL structure would occur and that the variability characteristics of the VF and RNFL measurements are similar to those seen in clinical practice over longer periods of time. The study was undertaken in accordance with good clinical practice guidelines and adhered to the Declaration of Helsinki. The study was approved by the North of Scotland National Research Ethics Service committee on 27 September 2013 (reference no.: 13/NS/0132) and NHS Permissions for Research was granted by the Joint Research Office at University College London Hospitals NHS Foundation Trust on 3 December 2013. All patients provided written informed consent before the screening investigations were carried out. Recruitment criteria were based on those for the UKGTS. Patients were required to have reproducible VF loss with corresponding damage to the ONH and no other condition that could lead to VF loss, be aged > 18 years and have a visual acuity of $\geq$ 20/40, a refractive error within $\pm$ 8 dioptres and an IOP of $\leq$ 30 mmHg. The VF MD had to be better than –16 dB in the worse eye and better than –12 dB in the better eye. VF loss was defined as a reduction in sensitivity at two or more contiguous locations with $p < 0.01$ loss or more, three or more contiguous locations with $p < 0.05$ loss or more, or a 10-dB difference across the nasal horizontal midline at two or more adjacent locations in the total deviation plot. Participants attended approximately once a week for 10 visits, with VF testing and OCT imaging carried out twice at the first visit and once at each subsequent visit. VF testing was undertaken with the Humphrey Field Analyser^{TM} (HFA) and OCT imaging was carried out using

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

UKGTS

The UKGTS is a multicentre, randomized, double-masked, placebo-controlled trial assessing visual function preservation in newly diagnosed open-angle glaucoma (OAG) patients (trial registration number, ISRCTN96423140). 516 newly-diagnosed (previously untreated) participants with OAG were prospectively recruited at 10 UK centres between 2007 and 2010. The observation period was 2 years, with subjects monitored by VF testing, quantitative imaging, optic disc photography and tonometry at 11 scheduled visits. ONH structure was monitored with Heidelberg Retina Tomograph at all study sites and with Stratus TD OCT^{TM}(Carl Zeiss Meditec Inc., Dublin, CA, USA) (software version 5.0) and GDxECC Nerve Fiber Analyzer (Carl Zeiss Meditec Inc., Dublin, CA, USA) at study sites with those devices. With respect to the whole UKGTS cohort, the baseline mean IOP (±SD) was 18.9±4mmHg in the better mean deviation (MD) eyes (median [IQR] MD -1.27dB [-2.37, -0.19]) and 19.9±4.6mmHg in the worse MD eyes (median [IQR] MD -3.30dB [-5.60, -1.98]). The median (interquartile range) VF MD for all eligible eyes was -2.9 dB (-1.6 to -4.8 dB).

The participants were allocated randomly to receive the IOP-reducing prostaglandin analog latanoprost (0.005%) or placebo eye drops. The UKGTS, and the subsequent analysis of anonymized data in this study, adhered to the tenets of the Declaration of Helsinki and was approved by local institutional review boards (Moorfields and Whittington Research Ethics Committee on June 1, 2006, ethics approval reference, 09/H0721/56). Study participants provided written informed consent. A total of 488 from 516 enrolled participants with post-baseline data were analysed in the trial (latanoprost, n=244; placebo, n=244). Out of those, a subset of 284 participants (143 participants in the placebo group and 141 participants in the latanoprost group) had adequate quality VF and OCT data, with > 6 months of follow-up, and five or more visits and with data for both VFs and OCT at the baseline visit. For eye-based analysis, the eye with the worse MD was used. VF deterioration was the primary end point in the trial; time to VF deterioration within 24 months. Deterioration (progression) analysis was performed in the Humphrey Field Analyser$^{\text{TM}}$ (HFA) II-i Guided Progression Analysis$^{\text{TM}}$ (GPA) software (version 5.1.1) (Carl Zeiss Meditec Inc., Dublin, CA, USA), a sensitive technique that considers changes at individual test locations in the visual field. Deterioration (progression) criteria and details of the trial design and trial outcome are published elsewhere[15,33]. In short, the time to VF deterioration was significantly longer in the treatment group than in the placebo group (adjusted hazard ratio, 0.44; 95% confidence interval, 0.28 to 0.69).

Visual Field Measurements

All VF tests were performed with the HFA II (or II-i) and the SITA standard 24-2 program. A reliable VF was one with a false-positive rate of < 15% and < 20% fixation losses (for fixation losses of > 20%, reliability was based on the subjective judgement of the technician supervising the test and the clinician reading the test, including an assessment of the eye tracker trace). Unreliable tests were repeated, either on the same day (with a break of at least 30 minutes) or on a subsequent occasion. The reference standard analysis for VF deterioration was that used for the outcome of the UKGTS and was undertaken with the HFA II-i GPA software (version 5.1.1)[15].

Spectralis OCT Retinal Nerve Fiber Layer Measurement

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

Stratus OCT Retinal Nerve Fiber Layer Measurement

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

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

Imaging Analysis Protocol and Quality Control

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

Analysis

Image Synthesis. We use cyclical GANs[32,34] to infer morphological descriptors from low to high quality anatomical information. OCT images have a very specific geometry where the background, i.e. vitreous cavity, is clearly separated from the retinal layers at the ILM. Thus, we used image stitching, exploiting the ILM identification, to separate background from layer signal. Moreover, cycle GANs require a fixed window on which spatial filters and mappings are learned. However, since OCT signal and noise properties are characterized by different spatial scales, a modality transfer method based on a fixed spatial window might not be able to capture all the necessary spatial information needed for synthesis. This reduces the chance for cross-modal distributions to share supports in latent space. To address this problem, we propose an ensemble of spatially coherent cycle GANs[32] to learn the TDOCT-to-SDOCT mapping and to translate a TDOCT into a synthesized SDOCT image. The scheme is the following. Each GAN is trained by employing a different spatial window size: 128 x 128, 256 x 256 or 512 x 512, learning a mapping from the observed TDOCT image $\mathbf{I}_{\mathbf{TD}}$ and random noise vector **z**, to the target SDOCT image $\mathbf{I}_{\mathbf{SD}}$, G: $\{\mathbf{I}_{\mathbf{TD}}, \mathbf{z}\} \rightarrow \mathbf{I}_{\mathbf{SD}}$. As a result, we train six GANs: three with background pairs and three with retinal layer pairs. The synthesized backgrounds and layers are stitched back according to the window size, i.e. $I_{128x128}$, $I_{256x256}$, $I_{512x512}$, and the average synthesized stitched image $\bar{\boldsymbol{I}}$ is obtained. To preserve the morphological correlation between training pairs, cycle GANs were trained with windows centered at the same geometrical location in both pairs. This deep learning technique is based on learning the representation between TD and SD OCT using 24,792 paired examples. The transfer mapping is learned in an independent dataset, i.e. the RAPID dataset, which contains pairs of both modalities, and the method is applied to the UKGTS dataset, enhancing the TD OCT images via quality transfer from SD OCT. TD OCT images are converted to ‘synthesized SD OCT’ images and segmented via an ensemble of GANs: for each TD OCT, we produce three SD OCT candidates. Fig. 1 shows the proposed framework for OCT synthesis via the ensemble of GANs.

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

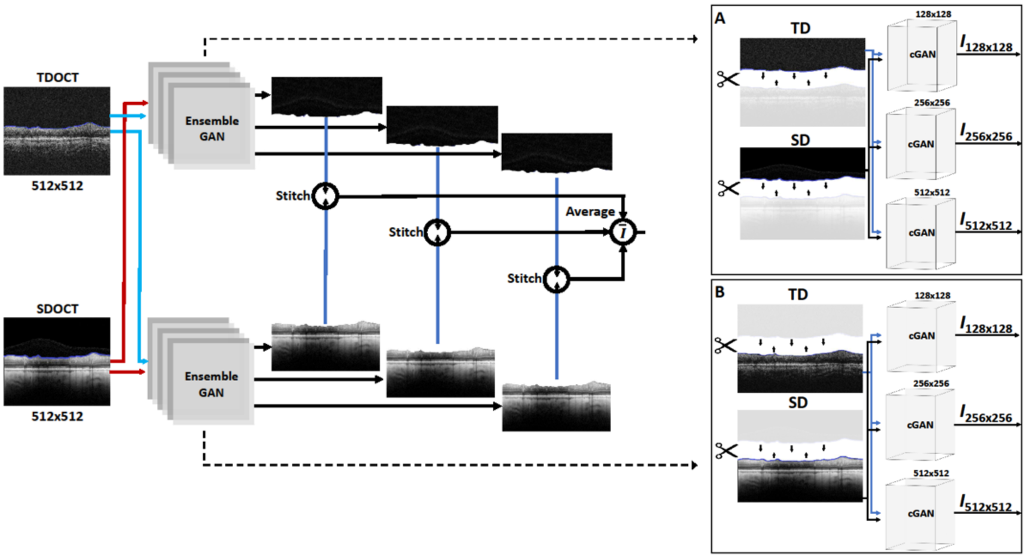

Figure 1: SDOCT synthesis via ensemble of GANs. Box A: Backgrounds are painted black. Box B: Three GANs are trained with layer pairs. Synthesized images are stitched back with the backgrounds and the average synthesized stitched image is obtained. Separation of layers and background is illustrated with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

Statistical Analysis and Evaluation. We quantified the quality improvement of the ‘synthesized SD OCT’ images over the original TD OCT images in both the RAPID and UKGTS data sets. Fig. 2 shows an example of a SDOCT image synthesized from a TDOCT image. Fig. 2a and Fig. 2b constitute the original TDOCT-SDOCT pair of images, whereas Fig. 2c is the synthesized SDOCT after modality transfer and synthesis. To compare the performance of the Cox models, i.e. Cox model before and after TDOCT image enhancement, we calculate the rank-based Somers’ D between predicted risk scores and observed survival times. We compare the rankings of rate of RNFL loss and time-to-VF progression per patient across the dataset and we assess their agreement. Somers’ D takes values between -1 when all ranking pairs disagree and 1 when all pairs agree. To estimate the standardized effect size for the same population before and after TDOCT image enhancement, we calculate Cohen’s D using the difference in the rates of loss between the treatment groups. Although there are no reference values for Cohen’s standardized effect size measures, d = 0.2, 0.5 and 0.8 provide a conventional reference frame, corresponding to small, medium and large effects [43].

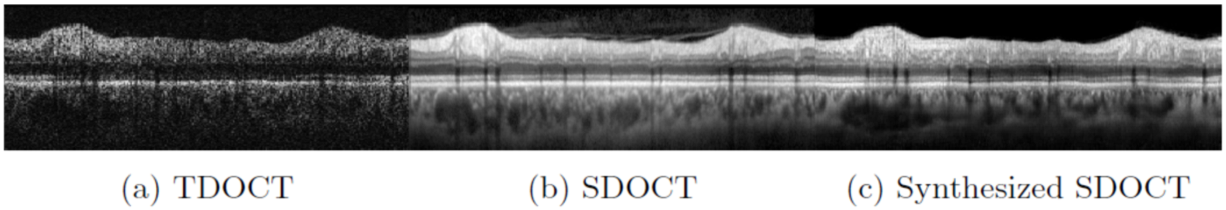
(a) TDOCT (b) SDOCT (c) Synthesized SDOCT

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).

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

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

Results

Test-retest variability, summarized by the standard deviation of repeat measurements over the first three visits across all subjects of the RAPID study, was lower for the Synthesized SDOCT than for the original TDOCT data (Table 1). Table 1 also shows the 95% limits of agreement (LOA) and the mean difference between RNFL measurements. The 95% limits of agreement between TDOCT and SDOCT were [26.64, -22.95], between synthesized SDOCT and SDOCT were [8.11, -6.73], and between SDOCT and SDOCT were [4.16, -4.04]. Fig. 3 illustrates the corresponding Bland-Altman agreement plots of the RNFL measurements made from the segmented synthesized OCT images with respect to the 'ground truth' Spectralis SD OCT RNFL measurements derived with the same segmentation algorithm (RAPID data set). Table 2 presents the mean and the range of RNFL loss rates for TDOCT and synthesized SDOCT images. Table 3 and Table 4 illustrate the Cox proportional hazards model fitted to the time to VF progression for TD OCT and synthesized SD OCT. The hazard ratio for RNFLT slope in Cox regression modelling for time to incident VF 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) (p=0.011) for synthesized SDOCT. Fig. 4 illustrates the VF mean sensitivity (MS) change in decibels per year and the distribution of rate of RNFL thickness change for the subset of UKGTS participants with OCT images. Fig. 4b is generated from the original TD OCT whereas Fig. 4c from the synthesized SDOCT data. The placebo group had faster rates of deterioration than the latanoprost group in both cases. For the original TD OCT UKGTS data, the difference in distribution of slopes was not statistically significant (Mann-Whitney U Test, *p* = 0.08). For the synthesized SD OCT, the difference was statistically significant (Mann-Whitney U Test, *p* = 0.0017). Table 5 illustrates the corresponding effect sizes (Cohen's D), with confidence intervals. It can be seen that Cohen's D for synthesized SD OCT is closer to Cohen's D for VFs than that for TD OCT, indicating a modest improvement in effect size. Table 6 compares the predictive power of the two Cox models; we calculate the rank order statistic Somers' D with confidence limits[42].

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

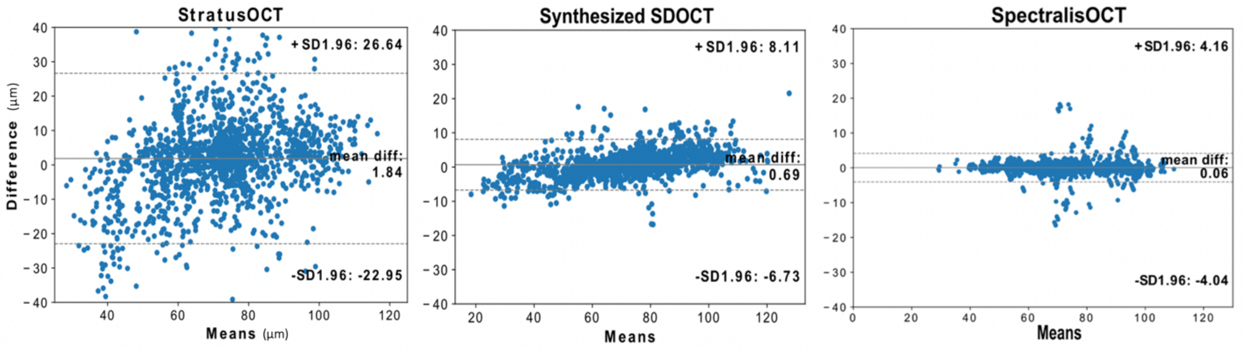

Figure 3: Bland-Altman plots on the agreement between time domain and synthesized spectral domain OCT RNFL measurements versus the 'real' spectral domain OCT RNFL measurements on the RAPID dataset. The proposed method leads to significantly better agreement.

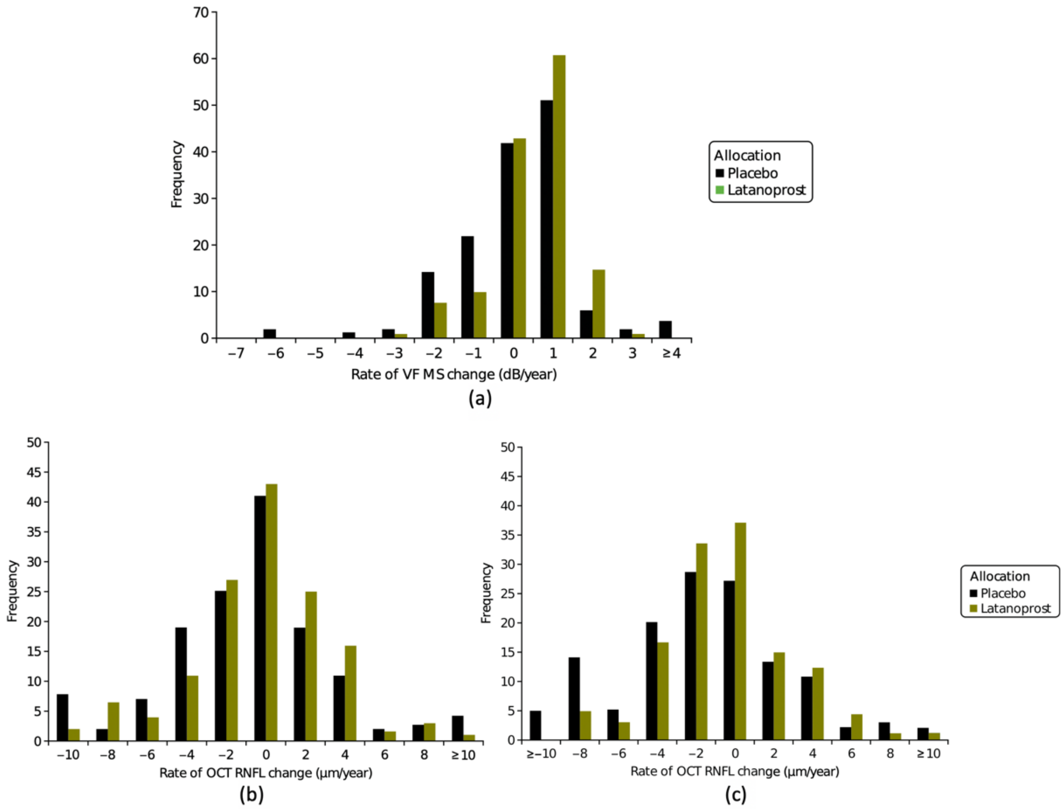

Figure 4: (a) Distribution of the rate of VF mean sensitivity (MS) change in decibels per year for the subset of UKGTS participants with OCT images (placebo, n = 131 participants; latanoprost, n = 127 participants). Bottom: Distribution of the rate of OCT RNFL thickness change for the subset of UKGTS 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

Method	**Synthesized SDOCT**	StratusOCT	SpectralisOCT
95% LOA	[**8.11**, **-6.73**]	[26.64, -22.95]	[4.16, -4.04]
Mean Diff.	**0.69**	1.84	0.06
Mean SD	**1.29**	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

Method	StratusOCT		Synthesized SD OCT	
	Treatment	Placebo	Treatment	Placebo
Mean (SD) (μm/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)	
p-value	0.08		0.0017	

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 Exp(b) = the hazard ratio. (placebo, n = 131 participants; latanoprost, n = 127 participants).

Covariate	b	SE	Wald	*p*	Exp(b)	95% CI of 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

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).

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

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.

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

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

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

Discussion

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

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

When the super resolution method was applied to an independent test data set, from the UKGTS, the better separation of the treatment arms evidenced the data quality improvement. The analysis of the capability of TD OCT images to distinguish the UKGTS treatment arms showed that, although the rate of RNFLT loss was faster in the placebo-treated eyes, the difference from the latanoprost-treated eyes did not reach statistical significance (Table 2; Figure 4b). In contrast, the same analysis with the synthesized SD OCT images demonstrated a statistically significant difference between treatment and placebo progression rates (MannWhitney U Test, p = 0.0017 (Table 2; Figure 4c). The difference between treatment groups in the rate of RNFL thinning (synthesized SD OCT) is closer to the difference between groups for the rate of VF MD deterioration (Figure 4) than for the TD OCT analysis (Table 5). Our analysis further illustrates that the SD OCT imaging of RNFL may provide a sufficiently high precision for longitudinal assessment of RNFL changes, as low measurement variability is a prerequisite for detecting change during longitudinal analysis (Table 6); improving the longitudinal SNR.

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

Study weaknesses and further work

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

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

Future work will focus on combining SD OCT RNFL rates of change to VF rates of change, in a similar way as that done for TD OCT[40], to see whether the addition of the imaging data improves study power over the use of VF data alone. The motivation is that although the signal-to-noise ratio in the TD OCT UKGTS data is too poor to draw conclusions with respect to disease deterioration, the synthesized SD OCT data provided some evidence that imaging outcomes capture the effect of treatment on the VF outcome.

Conclusion

In clinical trials with a vision function outcome, variability in measurements results in the requirement for large numbers of patients observed over long intervals. As a result, new beneficial treatments to patients may be delayed and may not be evaluated as trials become more costly. It is well established that imaging measurements of structural damage to the ONH are associated with VF loss in glaucoma. Furthermore, imaging measurements are often considered more precise than VF measurements, making them attractive as potential surrogate outcomes for clinical trials and clinical practice. The OCT data available in the UKGTS were from the TD OCT, with poor signal-to-noise characteristics. Previous analysis of the OCT data failed to distinguish the treatment arms[40]. Here, we show that a super resolution deep learning method was able to considerably improve data quality, demonstrated by better agreement of RNFL measurements from synthesized SD OCT images, compared with their source TD OCT images, with RNFL measurements from actual SD OCT images. When applied to an independent data set from the UKGTS, the data quality improved to the extent that imaging measurements were able distinguish treatment groups. These findings suggest that a benefit to trial power can be achieved by a) further increase the resolution of SDOCT using SR methods b) ensemble methods to segment more efficiently SDOCT images.

References

1. Hood DC, Raza AS, de Moraes CG, Liebmann JM, Ritch R. Glaucomatous damage of the macula. Prog Retin Eye Res. 2013;32:1-21. doi:10.1016/j.preteyeres.2012.08.003

2. Resnikoff S Pascolini D Etya'ale D et al. Global data on visual impairment in the year 2002. Bull World Health Organ. 2004; 82: 844-851

3. Bunce C, Wormald R. Causes of blind certifications in England and Wales: April 1999–March 2000. Eye. 2008; 22: 905-911

4. Friedman DS, Freeman E, Munoz B, Jampel HD, West SK. Glaucoma and mobility performance: the Salisbury Eye Evaluation Project. Ophthalmology. 2007; 114: 2232-2237

5. Haymes SA, Leblanc RP, Nicolela MT, Chiasson LA, Chauhan BC. Risk of falls and motor vehicle collisions in glaucoma. Invest Ophthalmol Vis Sci. 2007; 48: 1149-1155

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

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

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

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

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

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

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

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

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

15. Felipe A. Medeiros, Linda M. Zangwill, Luciana M. Alencar, Christopher Bowd, Pamela A. Sample, Remo Susanna, Robert N. Weinreb; Detection of Glaucoma Progression with Stratus OCT Retinal Nerve Fiber Layer, Optic Nerve Head, and Macular Thickness Measurements. Invest. Ophthalmol. Vis. Sci. 2009;50(12):5741-5748. doi: 10.1167/iovs.09-3715.

16. Medeiros FA. Biomarkers and surrogate endpoints in glaucoma clinical trials. Br J Ophthalmol. 2015;99(5):599–603. doi:10.1136/bjophthalmol-2014-305550

17. Medeiros FA. Biomarkers and Surrogate Endpoints: Lessons Learned From Glaucoma. Invest Ophthalmol Vis Sci. 2017;58(6):BIO20–BIO26. doi:10.1167/iovs.17-21987

18. Nie D, Trullo R, Lian J, et al. Medical Image Synthesis with Context-Aware Generative Adversarial Networks. Med Image Comput Comput Assist Interv. 2017;10435:417-425. doi:10.1007/978-3-319-66179-7_48

19. Wolterink JM, Dinkla AM, Savenije MHF, Seevinck PR, van den Berg CAT, Išgum I. Deep MR to CT synthesis using unpaired data. Med Image Comput Comput Assist Interv. 2017;10557:14–23. 10.1007/978-3-319-68127-6_2

20. Ben-Cohen A, Klang E, Raskin SP, Amitai MM, Greenspan H. Virtual PET Images from CT Data Using Deep Convolutional Networks: Initial Results. Med Image Comput Comput Assist Interv. 2017;10557:49-57. 10.1007/978-3-319-68127-6_6

21. Wang TC, Liu, MY, et al. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018:8798-8807, doi: 10.1109/CVPR.2018.00917.

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

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

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

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

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

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

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

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

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

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

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

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

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