

Title: Diagnostic accuracy of Optical Coherence Tomography for diagnosing glaucoma: secondary analyses of the GATE study

Authors: Gianni Virgili, MD¹, Manuele Michelessi, MD², Jonathan Cook, PhD,³ Charles Boachie, PhD,⁴ Jennifer Burr.MD,⁵ Katie Banister, PhD,⁶ David Garway-Heath.MD,⁷ Rupert Bourne, MD,⁸ Almudena Asorey, MD,⁹ Craig Ramsay, PhD,⁶ Augusto Azuara-Blanco, MD, PhD¹⁰

1. Department of Translational Surgery and Medicine, Eye Clinic, University of Florence, Florence, Italy
2. Ophthalmology, Fondazione G.B. Bietti per lo studio e la ricerca in Oftalmologia-IRCCS, Rome, Italy
3. Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, UK
4. Robertson Centre for Biostatistics, University of Glasgow, UK
5. School of Medicine, University of St. Andrews, UK
6. Health Services Research Unit, University of Aberdeen, UK
7. NIHR Biomedical Research Centre, Moorfields Eye Hospital and UCL Institute of Ophthalmology, UK
8. Vision & Eye Research Unit, Postgraduate Institute, Anglia Ruskin University, UK
9. Department of Ophthalmology, Hospital San Carlos, Madrid, Spain
10. Centre for Experimental Medicine, Queen's University, Belfast, UK

Corresponding author:

Augusto Azuara-Blanco, MD, PhD
Professor of Ophthalmology, Queen's University Belfast
Institute of Clinical Sciences - Block A, Grosvenor Road, Belfast, BT12 6BA, UK
Phone: 02890976460
Fax: 02890632699
E-mail: a.azuara-blanco@qub.ac.uk

Financial Support:

- The GATE study was funded by the National Institute of Health Research HTA programme (09/22/111). The views and opinions expressed are those of the authors and do not necessarily reflect those of the HTA programme, NIHR, NHS or the Department of Health.
- The contribution of Manuele Michelessi of the IRCCS Fondazione Bietti in this paper was supported by the Italian Ministry of Health and by Fondazione Roma

Conflict of Interest: no conflicting relationship exists for any author

Introduction

In the last decade, a large number of studies has investigated the accuracy of imaging tests that assess the optic nerve head (ONH) and retinal nerve fiber layer (RNFL) thickness for diagnosing glaucoma.¹⁻⁹ Despite the amount of information available, a recent Cochrane review found suboptimal study design in nearly all diagnostic accuracy studies, since they adopted a case-control design which is known to overestimate accuracy.¹⁰

The GATE study is a large, multicenter, prospective test accuracy study that investigated the performance of automated classifications of three common imaging technologies for diagnosing manifest glaucoma, the Heidelberg Retina Tomography (Heidelberg Engineering, Heidelberg, Germany), scanning laser polarimetry (GDx-ECC, Carl Zeiss Meditec, Dublin, CA) and Spectralis optical coherence tomography (OCT, Heidelberg Engineering, Heidelberg, Germany). The performance of a triage test including imaging, intraocular pressure (IOP) and visual acuity was evaluated. In triage the new test is used before the existing testing pathway (e.g., clinical examination and visual field testing for glaucoma). The target population was selected in a well-defined step of the clinical pathway: patients referred by optometrists to hospital eye services in UK after prior testing identified increased risk for, or suspicion of, glaucoma.^{11,12} Although the use of an imaging technology as part of a triage test was found to be efficient, the GATE study described suboptimal performance of OCT when the standard automated classification of the global RNFL thickness was used, which included the categories 'within normal limits', 'borderline' or 'outside normal limits'.

OCT is the most widespread imaging technology. In this in-depth analysis of GATE study OCT data, we wished to explore further objectives: what is the performance of OCT at a

pre-defined high sensitivity level (needed for a triage setting, such as in the GATE study) or at high specificity level (needed for a screening setting, for which GATE would only provide indirect evidence), and what RNFL thickness cut-offs could be used for triage or screening, respectively? Other questions included: can glaucoma diagnosis be improved using additional information relating to the subject's between- and within-eye variation and other data available such as OCT scan quality, intraocular pressure (IOP), patient's age and refractive error?

Materials and Methods

Study design and participants have been described before.¹¹ In brief, in the GATE study we conducted a pragmatic multicentre within-patient comparative evaluation of the diagnostic accuracy of automated imaging techniques for diagnosis of glaucoma.¹¹

Institutional Review Board (IRB)/Ethics Committee approval was obtained. We recruited 966 participants prospectively and consecutively among 2088 participants referred to five UK hospital eye services from community optometrists and invited to attend the study. An ophthalmologist with glaucoma expertise performed a clinical assessment which formed the reference standard. Clinical examination included Goldmann applanation tonometry, gonioscopy, biomicroscopic examination of the optic disc (pupils dilated unless contraindicated) and evaluation of the visual field test with Humphrey SITA 24-2.

Glaucoma experts participating in the study met before starting the study and used a pre-agreed definition of glaucoma.^{11,12} Manifest glaucoma was classified with visual field assessment as mild, moderate or severe using mean deviation cut-offs of -6 dB and -12 dB. Imaging test results were not available to the ophthalmologist and the imaging

technician was also masked to glaucoma status assessment. All participants in the GATE study who underwent OCT were included in this study.

The primary analysis included data using from both eyes. A participant was defined as having glaucoma if the manifest glaucoma was diagnosed in one or both eyes, i.e. considering glaucoma-suspects as case-negatives. We conducted a secondary analysis including suspect and manifest glaucoma cases as positive diagnosis. We used receiver operator characteristics (ROC) curve analysis to assess the performance of RNFL thickness for diagnosing manifest glaucoma.¹³ Diagnostic factors were analyzed and then sequentially combined as follows. Based on a recent systematic review and primary research,^{10,14} we selected RNFL average and inferior quadrant thickness as OCT parameters of interest as they have previously been shown to have the highest diagnostic accuracy. We also extracted the absolute value of the difference between the inferior-temporal and the superior-temporal sector as a measure of within-eye difference, since this topographic comparison resembles one of the perimetric criteria for diagnosing glaucoma (i.e. Glaucoma Hemifield Test).¹⁵ The absolute value of the interocular difference in RNFL thickness was used as a measure of between-eye variation.¹⁶ We compared the medians of patient-level relevant variables in manifest glaucoma versus other groups using a nonparametric k-sample test on the equality of medians. We used logistic regression-based ROC analysis to build a set of models adding one parameter at a time with the following prespecified sequence: RNFL thickness in the eye with the lower value, either average or inferior quadrant; absolute interocular difference, average or inferior quadrant; absolute difference between inferior and superior temporal sectors; and the average quality of the two scans in the subject's eyes. Then, we considered the following variables in the model as they can be obtained at any optometrists' practice:

intraocular pressure (IOP) as the higher measure in either eye; refractive error as spherical equivalent, higher absolute value of the two eyes; and age. Sensitivity at 95% specificity and specificity at 95% sensitivity were calculated from ROC curves based on percentile value calculations according to Pepe et al.¹⁶ using 1000 bootstrap samples.

In a second set of exploratory sensitivity analyses, we considered an identical set of measures with data measured at the eye level, i.e. using two records per patient, adjusting for within-subject correlation in ROC analyses using logistic regression models with a clustered sandwich estimator.

The effect measure to select the best model was initially the AUC. We also obtained the specificity at 0.95 sensitivity and *a priori* considered that a specificity of at least 0.50 was acceptable for a SpPin test (specificity, positive result good for ruling in disease), reducing the burden of referrals with minimal loss of glaucoma cases. Vice versa, a SnNout test (sensitivity, negative result good for ruling out disease) was deemed acceptable at a minimum 0.95 sensitivity and 0.50 specificity.¹⁷ For univariate analyses of RNFL thickness, we recorded the cut-off at which such high sensitivity and specificity values were obtained.

We computed Diagnostic Odds Ratios (DORs) at 0.95 specificity or 0.95 sensitivity for models including a single variable, either average or inferior RNFL thickness.

No imputation of missing data was applied since data were complete at over 95%.

Participants were excluded from analyses only if the scan quality score was 15 or less in either eye, as stated in the protocol, since the likelihood of a correct glaucoma classification is reduced in this case

All calculations were made using Stata 14.1 (StataCorp, College Station, TX).

Results

A complete description of the patients' enrollment and flow in the GATE study has been previously published.¹¹ A flow diagram of participants included in this study is shown in figure 1. Out of 2088 subjects who were invited to attend, 966 participants (48%) consented into GATE study, of whom 11 were not referred for glaucoma and 1 withdrew, leaving 955 eligible patients. Of these, 56 were excluded (see Figure 1). Mean age was 62 years (standard deviation: 14 years) and 49% of participants were male. After clinician evaluation, manifest glaucoma was found in 55 participants in the right eye only, in 43 participants in the left eye only, while in 55 participants it was diagnosed in both eyes, totaling 153 participants with glaucoma in at least one eye (17%); 235 (26.1%) participants were suspect for glaucoma in at least one eye

Figure 2 shows the distribution of all variables in glaucoma and non-glaucoma cases as boxplots. There were significant differences ($p \leq 0.001$) between non-glaucoma and glaucoma participants in median values of all RNFL thickness parameters and other variables, as defined in the Methods: average RNFL thickness (89 versus 69 micron) or inferior RNFL thickness (111 versus 74 micron), the interocular difference of the average RNFL thickness (1.5 versus 5 micron) or inferior quadrant RNFL thickness (4 versus 11.5 micron), the largest difference in RNFL thickness between superior-temporal and inferior-temporal sectors (10 versus 17.5 micron), scan quality (28.5 versus 25.8), spherical equivalent refractive error (0.75 D versus 2.38 D) and age (62 versus 71 years), while median IOP did not differ (21.5 vs 22 mmHg, $p=0.740$). Despite these significant differences a substantial overlap is seen for all parameters between glaucoma and non-

glaucoma participants (Figure 2), and particularly the upper quartile of glaucoma participants generally overlapped with the middle quartiles of non-glaucoma participants.

ROC curve analysis

Figure 3 presents ROC curves of all variables of interest in groups of three variables to improve clarity. The ROC curve of the average and inferior RNFL thickness overlapped, though inferior thickness seemed to be better than average thickness at high specificity levels in the lower-left corner of the ROC plane. With reference to the AUC of the average thickness (0.83), that of the inferior was similar (0.83, $p=0.668$), while all others were significantly lower than average RNFL thickness ($p\leq 0.001$): interocular difference in average (0.74) or in inferior quadrant thickness (0.72), difference of superior to inferior temporal sectors (0.71), scan quality (0.68), age (0.67), spherical equivalent (0.61), IOP (0.55).

Table 1 shows the performance of a series of models in which variables are added one at a time based on the sequence explained in the Methods. AUCs were between 0.83 and 0.88. Sensitivity at 0.95 specificity was between 0.36 and 0.55 for all models, the highest values were reached by models including the inferior quadrant rather than the average RNFL thickness. Specificity at 0.95 sensitivity was also found to range between about 0.36 and 0.58.

Figure 3 shows that, at high specificity, the better performance of the inferior quadrant compared to the average was due to the irregular shape of the ROC curve for the inferior quadrant compared to the average at high specificity in the left and lower corner of the ROC space. A sensitivity analysis using a parametric (binormal) ROC model gave almost overlapping curves and sensitivity estimates of 0.45 and 0.43 for average and inferior

parameters, respectively. An alternative analysis at 0.90 specificity found that sensitivity was 0.57 and 0.56, respectively.

For the average, DORs were 10.7 and 11.6, whereas for inferior thickness they were 23.2 and 12.1, respectively, which can be considered a modest accuracy performance.

Threshold analysis and secondary analyses

ROC data tabulation found that a high specificity of 0.95 was found at a threshold of ≤ 62 micron for the average (sensitivity 0.38) and ≤ 77 micron for the inferior quadrant (sensitivity 0.55) RNFL thickness. The thresholds for reaching a high sensitivity of 0.95 were ≤ 92 micron and ≤ 117 micron, resulting in a specificity of 0.36 and 0.39 for the average and inferior RNFL thickness, respectively (Table 1).

Using the highly sensitive cut-off in a cohort of 1000 screened subjects containing 170 with glaucoma, triage with OCT would result in the referral of 692 participants with average RNFL thickness ≤ 92 micron or 667 participants with inferior quadrant thickness ≤ 117 micron, in order to detect 161 participants with glaucoma in at least one eye, while missing 9 patients. Data on the severity of glaucoma suggested that 6 missed participants would have mild glaucoma (out of 93 with mild glaucoma in the hypothetical cohort, 6% missed) and 3 would have moderate or severe glaucoma (out of 77 with moderate or severe glaucoma, 4% missed).

Pooling manifest glaucoma (n. 153) with suspect glaucoma (n. 233) in either eye to form the target condition, for average and inferior quadrant RNFL thickness the accuracy figures would be, respectively: AUC 0.73 (95%CI: 0.70 – 0.77) and 0.72 (95%CI: 0.68 – 0.75); sensitivity at 0.95 specificity 0.28 (95%CI: 0.19 – 0.38) and 0.27 (95%CI: 0.20 – 0.32); specificity at 0.95 sensitivity 0.13 (95%CI: 0.08 – 0.22) and 0.10 (95%CI: 0.06 –

0.19). Using the same target condition, at 0.90 specificity, sensitivity would be 0.44 (95%CI 0.36 – 0.50) and 0.37 (95%CI 0.28 – 0.43) for average and inferior quadrant RNFL thickness, respectively.

Figure 4 presents the positivity fraction for each diagnosis category using the highly sensitive cut-off of 117 micron or less for the inferior sector, where the detection rate of manifest glaucoma was 95% by design. Estimates are displayed separately for right and left eyes, which show consistent patterns. The positivity fraction for OHT was 63-65% and was similar to that of eyes with no glaucoma-related findings, whereas glaucoma-suspects had a higher referral rate, particularly for disc/VF suspects (Figure 4).

In sensitivity analyses on statistical modelling, models using an identical set of measures with data measured at the eye level (i.e. using one record per eye, adjusting for within-subject correlation in ROC analyses) either had poor convergence or did not change accuracy estimates (data available upon request).

Discussion

This in depth analysis of GATE data aimed at measuring the accuracy of OCT parameters at high levels of either sensitivity or specificity that would be relevant for a triage or screening setting, respectively, and to explore the potential for improvement in accuracy by adding parameters that are easily available to primary eye care professionals, such as within-subject symmetry data, OCT scan quality, age, refractive error and IOP. We selected the global and inferior RNFL based on a recent systematic review that found these parameters to have the highest diagnostic accuracy. We found the improvement in diagnostic performance obtained in this way is modest. OCT accuracy is likely better using inferior sector RNFL thickness at high sensitivity, reaching about 0.50 specificity.

The addition of data on between-eye and within-eye asymmetry, as well as age, refractive error and IOP did not improve substantially the diagnostic performance of the simple, easy-to-use automated classification that we used in GATE.

Since a large Cochrane review has shown that almost all studies on ONH and RNFL imaging tests for diagnosing glaucoma were case-control studies,¹⁰ only recently have large diagnostic studies been conducted that measures the accuracy of OCT in a cohort of prospectively enrolled subjects according to a single and well-defined question made in a screening¹⁴ or primary care referral^{11,12} setting. Different study designs and settings can lead to measuring different test performances. Particularly, it is well known that case-control studies overestimate accuracy.^{18,19}

In a case-detection context, where glaucoma prevalence was about 5%, Dabasia et al¹⁴ found that iVue OCT (*Optovue, Fremont, CA*) sensitivity at 90% specificity for detection of glaucoma suspect/POAG combined was 0.57 (0.44-0.70) and for POAG, it was 0.83 (0.68-0.98) for best-performing OCT parameter, i.e. inferior quadrant RNFL thickness.

Combining different tests was found to be useful to increase post-test probability. This study cannot be compared with ours because of the community-based triage setting and the adoption of a different OCT device.

The GATE study¹¹ found OCT sensitivity was 0.77 and specificity was 0.79 for diagnosing manifest glaucoma using the Spectralis RNFL classification of outside normal limits.

Combination of two different imaging tests did not improve substantially the diagnostic accuracy. The DOR was 12.2 for the automated classification in the primary GATE analysis, which was achieved about at the diagonal of the ROC curve, where sensitivity equals specificity. This DOR value is similar to those found for average and inferior RNFL thickness in our analyses, which is consistent with the fact that we sought to obtain

accuracy estimates at very high sensitivity or specificity. Though such DOR values are usually regarded as modest accuracy, these secondary analyses of GATE data aimed to define RNFL thickness threshold values at the boundary of the ROC curve that may be useful in clinical settings (e.g., triage or case detection) to prioritize either case detection or the reduction of referrals when accuracy is suboptimal.

The strengths of the GATE study are its prospective design and multicenter data collection on a large number of participants for whom a well-defined question was made: whether imaging tests could be used as triage test on referrals by optometrists to hospital eye services. However, there are some potential limitations. The analyses conducted in this secondary publication were only partly pre-specified, thus exploratory in nature. Two of the participating units recruited a relatively small number of potential participants thus the recruits were not necessarily consecutive. Although consensus among clinicians at different sites was sought through structured discussion and agreement, some assessor differences may have remained. In addition, the population included in GATE would not be representative of the general population and thus the potential applicability of this study for informing the value of OCT for screening needs to be interpreted with caution. In fact, GATE provides accuracy estimates that can be generalized to triage settings with a similar profile of referrals, whereas it can offer at best indirect evidence for researchers planning population-based screening or case-detection studies. The reliability of our estimates of sensitivity may be limited by the small absolute number of false positives, but was still sufficient to make values above 0.50-0.60 implausible when the specificity was fixed at 0.95. Moreover, our estimates of cut-off values to achieve desired accuracy levels should be confirmed in further research in order to be generalizable to other settings. However it

is uncertain whether the results of this study may be generalizable to other OCT instruments.

Diagnosing glaucoma during the very early stage of disease is challenging. It is possible OCT may correctly detect pre-perimetric disease and thus a positive test may be wrongly considered to be a false positive. Ideally a longitudinal follow-up would provide the best possible reference standard.

The analyses presented in this secondary publication provide further information on Spectralis OCT accuracy at high sensitivity or high specificity, as well as retinal thickness cut-offs to achieve them. When using OCT for triaging glaucoma, the overall RNFL, the inferior RNFL thickness data and the addition of between-eye and within-eye asymmetry data, as well as age, refractive error and IOP did not improve substantially the diagnostic performance of the simple, easy to use Spectralis automated classification as used in GATE.

References

1. Ferreras A, Pablo LE, Pajarín AB, Larrosa JM, Polo V, Pueyo V. Diagnostic ability of the Heidelberg Retina Tomograph 3 for glaucoma. *Am J Ophthalmol*. 2008 Feb;145:354-359.
2. Reus NJ, Lemij HG. Diagnostic accuracy of the GDx VCC for glaucoma. *Ophthalmology*. 2004 Oct;111:1860-1865.
3. Wu H, de Boer JF, Chen TC. Diagnostic capability of spectral-domain optical coherence tomography for glaucoma. *Am J Ophthalmol*. 2012 May;153:815-826.
4. Akashi A, Kanamori A, Nakamura M, Fujihara M, Yamada Y, Negi A. Comparative assessment for the ability of Cirrus, RTVue, and 3D-OCT to diagnose glaucoma. *Invest Ophthalmol Vis Sci*. 2013 Jul 10;54:4478-4484.
5. Oddone F, Centofanti M, Tanga L, et al. Influence of disc size on optic nerve head versus retinal nerve fiber layer assessment for diagnosing glaucoma. *Ophthalmology*. 2011 Jul;118:1340-1347.
6. Medeiros FA, Zangwill LM, Bowd C, Weinreb RN. Comparison of the GDx VCC scanning laser polarimeter, HRT II confocal scanning laser ophthalmoscope, and stratus OCT optical coherence tomograph for the detection of glaucoma. *Arch Ophthalmol*. 2004 Jun;122:827-837.
7. Mwanza JC, Durbin MK, Budenz DL, et al. Glaucoma diagnostic accuracy of ganglion cell-inner plexiform layer thickness: comparison with nerve fiber layer and optic nerve head. *Ophthalmology*. 2012 Jun;119:1151-1158.
8. De León-Ortega JE, Sakata LM, Monheit BE, McGwin G Jr, Arthur SN, Girkin CA. Comparison of diagnostic accuracy of Heidelberg Retina Tomograph II and Heidelberg Retina Tomograph 3 to discriminate glaucomatous and nonglaucomatous eyes. *Am J*

Ophthalmol. 2007 Oct;144:525-532.

9. Badalà F, Nouri-Mahdavi K, Raoof DA, Leeprechanon N, Law SK, Caprioli J. Optic disk and nerve fiber layer imaging to detect glaucoma. *Am J Ophthalmol.* 2007 Nov;144:724-732.
10. Michelessi M, Lucenteforte E, Oddone F, et al. Optic nerve head and fibre layer imaging for diagnosing glaucoma. *Cochrane Database Syst Rev.* 2015 Nov 30;11:CD008803.
11. Azuara-Blanco A, Banister K, Boachie C, et al. Automated imaging technologies for the diagnosis of glaucoma: a comparative diagnostic study for the evaluation of the diagnostic accuracy, performance as triage tests and cost-effectiveness (GATE study). *Health Technol Assess.* 2016 Jan;20:1-168.
12. Banister K, Boachie C, Bourne R, Cook J, Burr JM, Ramsay C, Garway-Heath D, Gray J, McMeekin P, Hernández R, Azuara-Blanco A. Can Automated Imaging for Optic Disc and Retinal Nerve Fiber Layer Analysis Aid Glaucoma Detection? *Ophthalmology.* 2016 May;123:930-8.
13. Pepe M, Longton G, Janes H. Estimation and Comparison of Receiver Operating Characteristic Curves. *Stata J.* 2009 Mar 1;9:1.
14. Dabasia PL, Fidalgo BR, Edgar DF, Garway-Heath DF, Lawrenson JG. Diagnostic Accuracy of Technologies for Glaucoma Case-Finding in a Community Setting. *Ophthalmology.* 2015;122:2407-2415
15. Katz J, Quigley HA, Sommer A. Detection of incident field loss using the glaucoma hemifield test. *Ophthalmology.* 1996 Apr;103:657-663.
16. Levine RA, Demirel S, Fan J, Keltner JL, Johnson CA, Kass MA; Ocular Hypertension Treatment Study Group. Asymmetries and visual field summaries as

predictors of glaucoma in the ocular hypertension treatment study. *Invest Ophthalmol Vis Sci.* 2006 Sep;47:3896-3903.

17. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine. How to practice and teach EBM. New York: Churchill Livingstone, 2000.

18. Medeiros FA, Ng D, Zangwill LM, Sample PA, Bowd C, Weinreb RN. The effects of study design and spectrum bias on the evaluation of diagnostic accuracy of confocal scanning laser ophthalmoscopy in glaucoma. *Invest Ophthalmol Vis Sci.* 2007 Jan;48:214-222.

19. Rao HL, Kumbar T, Addepalli UK, et al. Effect of spectrum bias on the diagnostic accuracy of spectral-domain optical coherence tomography in glaucoma. *Invest Ophthalmol Vis Sci.* 2012 Feb 29;53:1058-1065.

Figure Legends:

Figure 1. Participants' flow.

Figure 2. Boxplots of OCT parameters, spherical equivalent, IOP and age for non-glaucoma (left in each square) and glaucoma (right in each square) cases. Average and inferior RNFL thickness is recorded as the lower value in either eye of each patient, interocular difference as the absolute value, superior-temporal to inferior-temporal difference as the absolute value of the larger difference in either eye of each patient, scan quality as the mean value of the patient's eyes, spherical equivalent as the more positive measure of the patient's eyes, IOP as the higher measure of the patient's eyes.

Figure 3. ROC curves of all parameters for detecting manifest glaucoma.