Can Automated Imaging for Optic Disc and Retinal Nerve Fiber Layer Analysis Aid Glaucoma Detection?

Katie Banister, PhD,1 Charles Boachie, BSc (Hons),2 Rupert Bourne, FRCoPhth,3 Jonathan Cook, PhD,4 Jennifer M. Burr, MD,5 Craig Ramsay, PhD,1 David Garway-Heath, MD,6 Joanne Gray, MSc,7 Peter McMeekin, PhD,7 Rodolfo Hernández, MSc,8 Augusto Azuara-Blanco, FRCoPhth9

Purpose: To compare the diagnostic performance of automated imaging for glaucoma.

Design: Prospective, direct comparison study.

Participants: Adults with suspected glaucoma or ocular hypertension referred to hospital eye services in the United Kingdom.

Methods: We evaluated 4 automated imaging test algorithms: the Heidelberg Retinal Tomography (HRT; Heidelberg Engineering, Heidelberg, Germany) glaucoma probability score (GPS), the HRT Moorfields regression analysis (MRA), scanning laser polarimetry (GDx enhanced corneal compensation; Glaucoma Diagnostics (GDx), Carl Zeiss Meditec, Dublin, CA) nerve fiber indicator (NFI), and Spectralis optical coherence tomography (OCT; Heidelberg Engineering) retinal nerve fiber layer (RNFL) classification. We defined abnormal tests as an automated classification of outside normal limits for HRT and OCT or NFI ≥ 56 (GDx). We conducted a sensitivity analysis, using borderline abnormal image classifications. The reference standard was clinical diagnosis by a masked glaucoma expert including standardized clinical assessment and automated perimetry. We analyzed 1 eye per patient (the one with more advanced disease). We also evaluated the performance according to severity and using a combination of 2 technologies.

Main Outcome Measures: Sensitivity and specificity, likelihood ratios, diagnostic odds ratio, and proportion of indeterminate tests.

Results: We recruited 955 participants, and 943 were included in the analysis. The average age was 60.5 years (standard deviation, 13.8 years); 51.1% were women. Glaucoma was diagnosed in at least 1 eye in 16.8%; 32% of participants had no glaucoma-related findings. The HRT MRA had the highest sensitivity (87.0%; 95% confidence interval [CI], 80.2%–92.1%), but lowest specificity (63.9%; 95% CI, 60.2%–67.4%); GDx had the lowest sensitivity (35.1%; 95% CI, 27.0%–43.8%), but the highest specificity (97.2%; 95% CI, 95.6%–98.3%). The HRT GPS sensitivity was 81.5% (95% CI, 73.9%–87.6%), and specificity was 67.7% (95% CI, 64.2%–71.2%); OCT sensitivity was 76.9% (95% CI, 69.2%–83.4%), and specificity was 78.5% (95% CI, 75.4%–81.4%). Including only eyes with severe glaucoma, sensitivity increased: HRT MRA, HRT GPS, and OCT would miss 5% of eyes, and GDx would miss 21% of eyes. A combination of 2 different tests did not improve the accuracy substantially.

Conclusions: Automated imaging technologies can aid clinicians in diagnosing glaucoma, but may not replace current strategies because they can miss some cases of severe glaucoma. Ophthalmology 2016;123:930-938 Crown Copyright © 2016 Published by ELSEVIER Inc. on behalf of American Academy of Ophthalmology. This is an open access article under the Open Government Licence (OGL) (http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/).

Supplemental material is available at www.aaojournal.org.

See Editorial on page 926.
Moorfields regression analysis (MRA) and the glaucoma probability score (GPS). The RNFL can be assessed using either scanning laser polarimetry, currently available as the GDx-PRO (Glaucoma Diagnostics [GDx] Carl Zeiss Meditec, Dublin, CA) or spectral-domain optical coherence tomography (OCT), with several commercial devices available. These imaging tests are user friendly and provide automated quantitative classifications.

Although many published data describe the diagnostic performance of imaging techniques in cohorts of retrospectively selected glaucoma patients or glaucoma-free normal subjects, there is no high-quality evidence of the comparative accuracy of current imaging techniques for identifying glaucoma in consecutive patients with unknown status screened for possible glaucoma. Existing data from case-control studies may not be applicable to the clinically relevant population who undergo assessment and diagnosis. We aimed to assess and compare the performance of these commercially available technologies to detect glaucoma in a prospective cohort. This work was conducted as part of a wider publicly funded study (National Institute for Health Research Health Technology Assessment [HTA], 09/22/111) that also evaluated cost-effectiveness of these imaging techniques in a triage setting in the United Kingdom.

### Methods

#### Study Design and Participants

We conducted a pragmatic multicenter, within-patient, comparative evaluation of the diagnostic accuracy of automated imaging techniques for diagnosis of glaucoma, the Glaucoma Automated Tests Evaluation (GATE). We selected participants prospectively, and they underwent imaging with all technologies under evaluation and then had the reference standard diagnosis (clinical assessment by a glaucoma expert, including examination of the fundus by biomicroscopy and visual field testing with Humphrey 24-2 Swedish interactive threshold algorithm testing, masked to the imaging test results). The study was approved by the North of Scotland Research Ethics Committee (reference, 10/S0801/58) and was conducted according to the tenets of the Declaration of Helsinki. The full study protocol is publicly available. We sought patient views on the design, conduct, and analysis of the study through representatives from the International Glaucoma Society.

The study was coordinated from a central study office in the Health Services Research Unit, University of Aberdeen, and was conducted in 5 National Health Service hospital eye services in the United Kingdom (Aberdeen, Bedford, Hinchinbrooke, Liverpool, and Moorfields). We identified eligible patients from their referral letter as being adults (age ≥18 years) who were newly referred from primary care to the department of ophthalmology of the recruiting hospital with a possible glaucoma diagnosis or glaucoma-related finding. This included high intraocular pressure; possible abnormalities in the optic disc, visual field test results, or both; and possible narrow anterior chamber angle. Patients were ineligible if they had a previous diagnosis of glaucoma or had already been seen by an ophthalmologist.

#### Table 1. Possible Diagnoses by the Clinician Performing the Reference Standard Measurement, Ranked in Order of Severity

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glaucoma</td>
<td>Evidence of glaucomatous optic neuropathy* and a characteristic visual field loss with MD of −12.01 dB or worse with MD of between −6.01 dB and −12 dB</td>
</tr>
<tr>
<td>Glaucoma suspect</td>
<td>Appearance suggestive of glaucomatous optic neuropathy, but also may represent a variation of normality, with normal visual fields (with or without high IOP)</td>
</tr>
<tr>
<td>Visual field suspect</td>
<td>Visual field loss suggestive of glaucoma, but also may represent a variation of normality, with normal appearance of the optic disc (with or without high IOP)</td>
</tr>
<tr>
<td>Ocular hypertension</td>
<td>Both the visual field and optic nerve appear normal in the presence of elevated pressure &gt;21 mmHg</td>
</tr>
<tr>
<td>Primary angle closure</td>
<td>Closed anterior chamber angle (appositionally or synechial) in at least 270° and at least 1 of the following: IOP &gt;21 mmHg and presence of peripheral anterior synechiae; both visual field and optic nerve appear normal</td>
</tr>
<tr>
<td>Primary angle-closure suspect</td>
<td>Closed anterior chamber angle (appositionally without any synechiae) in at least 270°, with IOP ≤21 mmHg; both visual field and optic nerve appear normal</td>
</tr>
<tr>
<td>No glaucoma-related findings</td>
<td>Absence of any of the above diagnoses</td>
</tr>
</tbody>
</table>

IOP = intraocular pressure; MD = mean deviation.

*Any of the following: optic disc or retinal nerve fiber layer structural abnormalities; diffuse thinning, focal narrowing, or notchting of the optic disc rim, especially at the inferior or superior poles; documented, progressive thinning of the neuroretinal rim with an associated increase in cupping of the optic disc; diffuse or localized abnormalities of the peripapillary retinal nerve fiber layer, especially at the inferior or superior poles; disc rim or peripapillary retinal nerve fiber layer hemorrhages; and optic disc neural rim asymmetry of the 2 eyes consistent with loss of neural tissue.

**Reliable visual field abnormality considered a valid representation of the subject’s functional status. Visual field damage consistent with retinal nerve fiber layer damage (e.g., nasal step, arcuate field defect, or paracentral depression in clusters of test sites). Visual field loss in 1 hemifield that is different from the other hemifield, that is, across the horizontal midline (in early or moderate cases). Absence of other known explanations.

### Participant Recruitment Process

We sent information about the study to potential eligible patients at each recruiting hospital, before their first hospital appointment. At their first clinic appointment, we then approached patients, and those patients who agreed to participate and signed the consent
Figure 1. Flow diagram showing diagnosis of the cohort. FN = false negative; FP = false positive; GPS = glaucoma probability score; HRT = Heidelberg retinal tomography; MRA = Moorfields regression analysis; OCT = optical coherence tomography; TN = true negative; TP = true positive.
form were enrolled. We recorded the demographics (age, gender) of those patients who declined to take part.

Testing Regimen
Before seeing the ophthalmologist, participants were imaged using all imaging tests in a random order followed by visual field measurement with standard automated perimetry with the Humphrey 24-2 Swedish interactive threshold algorithm strategy. An ophthalmologist with glaucoma expertise (reference standard) who was masked to the imaging results then assessed participants. We conducted all tests on the same day in 2 centers, and in 3 centers, the clinician assessment was on a separate date, within 2 weeks.

Index Tests
The following imaging technologies were compared.

Heidelberg Retina Tomograph III. A confocal laser scanning ophthalmoscope (Heidelberg Retina Tomograph III; Heidelberg Engineering) measures quantitative structural information of the optic disc anatomy and creates a topographic map of the retinal surface. Images are given a quality index (the mean topography standard deviation), which the manufacturer recommends should be less than 40 μm. We assessed 2 classification tools: the MRA and the GPS. Both methods produced an overall classification of within normal limits, borderline, or outside normal limits by comparison with normative data. Abnormal index test results for MRA and GPS were defined as an overall classification of outside normal limits.

Scanning Laser Polarimetry. Scanning laser polarimetry (the GDx with enhanced corneal compensation or the GDx PRO; Carl Zeiss Meditec) analyzes polarized light reflected from the fundus to measure the RNFL thickness. Images are given a quality index Q, which the manufacturer recommends should be 7 or more. The software generates an automated discriminating classifier of glaucoma, the nerve fiber indicator (NFI) value. The manufacturers report that 95% of the normal population has an NFI value of 35 or less and that 99% of the normal population has an NFI value of 55 or less. An abnormal index test result for the GDx was taken as an NFI value of 56 or more. An NFI value between 36 and 55 was considered to be similar to the borderline category of the other imaging tests. Measurements were obtained using either the GDx PRO or GDx variable corneal compensation (VCC) with an updated enhanced corneal compensation module.

Optical Coherence Tomography. Spectral-domain OCT was performed using the Spectralis (Heidelberg Engineering), which produces a detailed cross-sectional image of the retina using reflected light in a similar manner to B-mode ultrasound. Images are given a quality index Q, which the manufacturer recommends should be more than 15. The software (RNFL software version 5) automatically compared an average RNFL thickness with a normative database and produced an overall classification of within normal limits, borderline, or outside normal limits. Abnormal index test results for OCT were taken as a classification of outside normal limits.

We classified images as good quality if they met the manufacturer quality threshold as described above. Pupils were not dilated routinely unless image quality was inadequate. Between 1 and 3 experienced technicians performed the imaging tests at each center. We did not have any restriction on the same technician performing all imaging tests on an individual. The imaging technologies automatically generated a classification measurement without user input, except the HRT MRA, which required manual identification of the margin of the optic disc.

Reference Standard
An ophthalmologist with glaucoma expertise performed a clinical assessment that formed the reference standard (Table 1). All glaucoma experts participating in the study were trained in the procedures and used a preagreed definition of glaucoma (Table 1). 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 Swedish interactive threshold algorithm 24-2. This represents currently recommended clinical practice in the United Kingdom. Imaging test results were not available to the ophthalmologist. If a clinical diagnosis could not be established, (e.g., unreliable visual field measurement), an inconclusive diagnosis was recorded. To ensure a valid and consistent application of the agreed reference standard, a limited number of consultant ophthalmologists provided the assessment (n = 11) according to the diagnosis chart shown in Table 1.

Outcomes
Primary outcomes were sensitivity and specificity. Secondary outcomes were diagnostic odds ratio, likelihood ratio, and proportion of indeterminate test results.

Sample Size
We calculated the sample size using a 5% significance level based on a 2-sided test using standard diagnostic accuracy study methods. A study of 897 individuals would have 90% power to detect a 6% difference in accuracy for the primary outcome of diagnosis of glaucoma based on conservative assumptions (a probability of disagreement of 0.18, a glaucoma rate of 25%, and a sensitivity of 86%). This sample size also would yield 80% power for detecting a 6% difference in accuracy should the sensitivity be 93% (the current best estimate from meta-analyses of high-quality diagnostic studies). For specificity, there would be more than 90% power to detect a 5% difference. We assumed that 6% of test results would be indeterminate, and this increased the total sample size to 954.

Statistical Analysis
We included a single eye per patient in the analysis. We ranked the clinical diagnosis in order of decreasing severity according to Table 1. We used the worst eye based on this ranking for our analyses. When both eyes had the same diagnosis, we chose an eye at random. We calculated diagnostic measures (sensitivity, specificity, likelihood ratios, and diagnostic odds ratio) for each test for detection of glaucoma. We compared the diagnostic performance (sensitivity and specificity) of the imaging tests using McNemar’s test and generated corresponding 95% paired confidence intervals (CIs). We did not impute any missing data. We reported an indeterminate result when no automatic classification was generated, imaging quality was low, or an artifact was present. We explored the following using sensitivity analyses: performance of the tests across the spectrum of disease, classifying borderline test results as abnormal, combining results of more than one imaging test, and considering glaucoma suspects as glaucoma. In combining test results, if either test gave abnormal results, we classified this as abnormal. We only classified combined results as normal if results from both tests were normal. We calculated the area under the receiver operating characteristic curve (ROC) for each test using the global parameter for HRT or OCT and the NFI parameter for GDx, and we compared these using the method of DeLong et al. We...
have reported the results according to the Standards for Reporting Diagnostic accuracy studies (STARD) guidelines. We conducted a post hoc analysis, suggested by a peer reviewer, to explore informally the difference in test performance between tests by calculating individual recruiting center sensitivity and specificity (results not reported here).

### Results

We recruited 966 participants (48% of those approached) between April 2011 and July 2013. After obtaining informed consent, we excluded 11 participants: 10 were ineligible (preexisting glaucoma, n = 4; referred from other ophthalmologist, n = 4; or not referred for glaucoma, n = 2), and 1 person withdrew from the study. Therefore, 955 participants were available for the index test comparison. Imaging was not implemented for all index tests in 12 participants. Table 2 shows the distribution of test results, including indeterminate rates, is shown in Table 5 (available at www.aaojournal.org). Indeterminate rates were low (<10%) for all index tests; OCT had the lowest rate of indeterminate test results (4.2%). Indeterminate test results (including those with low quality) were not included in the primary analysis.

Figure 1 shows the categorization of the test results according to the reference standard finding (true and false positive, true and false negative). Of the 943 participants for whom all 4 tests were performed, 158 were classified by the clinician as glaucoma and 770 as not glaucoma. No conclusive diagnosis could be made for 4 participants. Table 6 (available at www.aaojournal.org) shows the diagnostic performance of the 4 tests, including CIs. The HRT MRA had the highest sensitivity (87.0%) but the lowest specificity (63.9%); GDx had the lowest sensitivity (35.1%) but the highest specificity (97.2%); and the other 2 tests provided intermediate results (HRT GPS values were very similar to the HRT MRA results, and OCT had very similar sensitivity and specificity values). The diagnostic odds ratio ranged from 9.24 for HRT GPS to 18.48 for GDx. Table 7 (available at www.aaojournal.org) shows the paired comparison of the 4 imaging tests.

### Table 2. Participant Demographics and Ocular Characteristics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
<th>Glaucoma</th>
<th>Nonglaucoma</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.</td>
<td>943</td>
<td>158</td>
<td>770</td>
</tr>
<tr>
<td>Mean age (SD), yrs</td>
<td>60.5 (13.8)</td>
<td>67.4 (12.7)</td>
<td>59.2 (13.6)</td>
</tr>
<tr>
<td>Female gender, no. (%)</td>
<td>482 (51.1)</td>
<td>74 (46.8)</td>
<td>401 (52.1)</td>
</tr>
<tr>
<td>Ethnicity, no. (%)</td>
<td>Black or blackCARIBBEAN</td>
<td>25 (2.7)</td>
<td>4 (2.5)</td>
</tr>
<tr>
<td></td>
<td>Black or blackBRITISH-AFRICAN</td>
<td>20 (2.1)</td>
<td>6 (3.8)</td>
</tr>
<tr>
<td></td>
<td>Asian or AsianBRITISH-INDIAN</td>
<td>18 (1.9)</td>
<td>5 (3.2)</td>
</tr>
<tr>
<td></td>
<td>Asian or AsianBRITISH-PAKISTANI</td>
<td>4 (0.4)</td>
<td>0 (0)</td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td>1 (0.1)</td>
<td>1 (0.6)</td>
</tr>
<tr>
<td></td>
<td>Other Asian</td>
<td>4 (0.4)</td>
<td>1 (0.6)</td>
</tr>
<tr>
<td></td>
<td>Mixed white and blackAfrican</td>
<td>1 (0.1)</td>
<td>1 (0.6)</td>
</tr>
<tr>
<td></td>
<td>White British</td>
<td>826 (89.2)</td>
<td>140 (88.6)</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>29 (3.1)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

### Table 3. Diagnosis of Participants’ Worse Eye by Clinician in Secondary Care (Reference Standard)

<table>
<thead>
<tr>
<th>Diagnosis by Clinician</th>
<th>Worse Eye</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No.</strong></td>
<td><strong>%</strong></td>
</tr>
<tr>
<td>Number of eyes</td>
<td>932</td>
</tr>
<tr>
<td>Glaucoma (reference standard positive test results)</td>
<td>158</td>
</tr>
<tr>
<td>Mild</td>
<td>78</td>
</tr>
<tr>
<td>Moderate</td>
<td>45</td>
</tr>
<tr>
<td>Severe</td>
<td>26</td>
</tr>
<tr>
<td>Severity not recorded</td>
<td>9</td>
</tr>
<tr>
<td>Disc suspect</td>
<td>170</td>
</tr>
<tr>
<td>VF suspect</td>
<td>36</td>
</tr>
<tr>
<td>VF + disc suspect</td>
<td>36</td>
</tr>
<tr>
<td>OHT</td>
<td>115</td>
</tr>
<tr>
<td>PAC</td>
<td>31</td>
</tr>
<tr>
<td>PAC suspect</td>
<td>83</td>
</tr>
<tr>
<td>No glaucoma-related findings</td>
<td>299</td>
</tr>
<tr>
<td>Inconclusive diagnosis</td>
<td>4</td>
</tr>
</tbody>
</table>

OHT = ocular hypertension; PAC = primary angle closure; VF = visual field.

---

*Self-reported ethnicity; no ethnicity was recorded for 15 participants.
We explored with sensitivity analyses the performance of the tests across the spectrum of disease by excluding mild or moderate glaucoma diagnoses, or both, as shown in Figure 2. Restricting the reference standard to a diagnosis of severe glaucoma yielded very high sensitivity (95%) for HRT MRA, HRT GPS, and OCT; of these 3, OCT demonstrated the highest specificity (71%).

We also explored the effect of including a borderline classification as an abnormal index test in addition to outside normal limits results, shown in Figure 3. Figure 4 (available at www.aaojournal.org) shows the categorization flow diagram for the analysis in Figure 3. Sensitivity was higher for all tests than under the default analysis, but with correspondingly lower
specificity. Likelihood ratios (and 95% CIs) showed evidence of both being able to rule in and out the presence of glaucoma for all 4 imaging tests (CIs did not contain 1.0). The positive likelihood ratio for GDx was substantially higher than for the other tests (less than 5). This was because of the very high specificity that more than offset the lower relative sensitivity performance. The diagnostic odds ratios ranged from 7.36 for GDx to 14.62 for HRT MRA. Table 8 (available at www.aaojournal.org) shows a further analysis of the diagnostic performance of the imaging tests when glaucoma suspect cases as well as glaucoma cases were classified as cases of disease for the reference standard.

Figure 5 (available at www.aaojournal.org) shows the receiver operating characteristics (ROC) curves for each test under the default analysis. The areas under the ROC curve were similar, with OCT having the highest (0.84) and MRA the lowest (0.79). There was no statistical evidence of a difference between them ($P = 0.077$).

We further explored the diagnostic accuracy of using combinations of tests (Table 9; available at www.aaojournal.org). The HRT MRA was combined with the other 3 tests to form a combined test. Combining tests in this way increased the sensitivity as expected, but only marginally, and at the expense of a greater decrease in specificity. Finally, the post hoc informal exploration of center differences showed some indication of MRA specificity varying between centers.

**Discussion**

The GATE study was a large, prospective, within-patient, comparative diagnostic accuracy study that provided the sensitivity and specificity data of 4 diagnostic imaging tests for glaucoma. In our study, most participants had successful testing and all 4 imaging tests had some value in terms of diagnosing or ruling out glaucoma. The HRT MRA had the highest sensitivity, but lower specificity, compared with other tests. By contrast, GDx had the best specificity, although the lowest sensitivity. The HRT GPS results were similar to those of the HRT MRA, as might be anticipated given that their analysis is based on imaging the same structure (i.e., the optic disc), and this finding has been reported previously in different populations. Sensitivity for OCT was of very similar magnitude to its specificity. Most tests conducted yielded good-quality automated test results, between 92% (GDx) and 97% (OCT). Optical coherence tomography had the lowest percentage of low-quality imaging results and GDx had the highest, according to the image quality classification provided in the device software. There also was a small percentage of cases (<5%) in which an automated imaging test result could not be produced.

A number of sensitivity analyses were carried out to assess the robustness of the findings of the primary analysis. The HRT MRA had consistently the highest sensitivity across most analyses, but at a cost of lower specificity compared with other tests. When excluding cases of mild glaucoma, HRT GPS had higher sensitivity than HRT MRA. For severe glaucoma, OCT had the highest sensitivity and specificity. By contrast, GDx had the highest specificity, but the lowest sensitivity, across all analyses. None of the imaging tests had a sensitivity of 100% in identifying severe glaucoma.

Varying the test definition of abnormal imaging test results by including the borderline category as positive results had the anticipated impact of improving the detection of glaucoma, although at the expense of more nonglaucoma cases being classified falsely as being cases of glaucoma. Using combinations of 2 imaging tests led to a marginal improvement in the detection of glaucoma and a reduction in the number of indeterminate test results. Because the HRT GPS and HRT MRA analyses consistently gave the best performing sensitivity and both methods are available on the same machine (HRT), an increase in HRT performance in terms of sensitivity could be obtained by assigning a diagnosis of glaucoma when either the HRT GPS or HRT MRA classifications were outside normal limits. However, the improvement was smaller than the loss in performance in terms of correctly identifying nondisease cases. Considering the additional cost and practical implications, including training and the purchase of additional equipment, suggests that the use of a single technology is to be preferred. The informal exploration of test performance by center suggested that HRT MRA may vary slightly more in this regard. However, there were likely to be some differences between centers in terms of population, which may explain this, or alternatively, it may be a chance finding.

Although the automated imaging classifications performed well at identifying disease (sensitivity), they did not perform as well in identifying normal cases (specificity). The exception to this was GDx, which had very poor sensitivity but moderate specificity. The choice of which imaging test is to be preferred to aid diagnosis reflects the inherent trade-off regarding diagnostic testing, where the desire not to miss glaucoma when present must be balanced against the desire to identify correctly those who are without disease.

Among the strengths of this study is the fact that the enrolled population included consecutive recruitment of subjects without a known history of disease, which reflects a potential clinical application of a diagnostic test. Patients were referred from community optometrists, and this gives a representative sample of current UK practice. The recruited cohort included a large percentage of patients without disease as well as those across the spectrum of glaucoma, including glaucoma suspects and those with ocular hypertension. Prevalence of glaucoma in this cohort was nearly 20%. Other reported studies evaluating the performance of these diagnostic technologies often have used patients already diagnosed with glaucoma, with patients already identified as normal, which has a risk of selection bias. A previous head-to-head comparison of the best-performing parameters of previous versions of the imaging devices the Heidelberg Retina Tomograph II, GDx VCC, and Stratus OCT found similar performance across techniques among a prospective sample of glaucoma patients compared with healthy volunteers. However, the sample did not represent routine practice (i.e., when the disease status is unknown), and the technologies have been upgraded since.

The tests that we compared are all routinely available in clinical practice, and each provided an automated classification of glaucoma status. By using the automated outputs, there is no need to interpret the large number of images and parameters that are produced by the technologies, and this therefore may be attractive to clinicians who are not glaucoma experts. An additional strength of our study is that we
used a paired design in which participants underwent all 4 tests, allowing a head-to-head comparison. We chose as the reference standard the clinical assessment provided by different ophthalmologists with glaucoma expertise. All ophthalmologists were trained in the study protocols and agreed to a common set of criteria to define glaucoma and normality and were masked to the imaging test results when making their diagnosis. By using different ophthalmologists working at different units, the results of the study are more likely to be generalizable than results from studies performed in a single unit.

Among the limitations, we recognize that diagnosing glaucoma during the very early stage of disease is challenging. Although consensus was sought through structured discussions and agreement, some assessor differences may have remained between the centers. Ideally, a longitudinal follow-up would provide the best possible reference standard. This was proposed by Medeiros et al. who used optic nerve head progression on stereophotographic examination as the criterion for glaucoma diagnosis. We used a single eye (worse diagnosis) per patient, and assessment of the presence of glaucoma in either eye may have a slightly different diagnostic performance. A limitation of the HRT MRA analysis is that it requires a trained technician to identify manually the margins of the optic disc on the acquired image, and this may affect the diagnostic performance if it is not carried out appropriately by a trained examiner. There are other OCT instruments in clinical use with glaucoma diagnostic capabilities, and the results of this study using the Spectralis device may not be applicable fully to other OCT technologies. In this study, we analyzed only a small amount of potential data generated. In addition, OCT hardware and software are evolving rapidly, and new developments may improve their diagnostic performance. Also, current OCT technology now is able to quantify macular parameters that may be useful to diagnose glaucoma, for example, in myopia.

The results of this study have implications for clinical practice. No test demonstrated or could have been expected to demonstrate 100% sensitivity and specificity, and less-than-perfect results reflect a combination of factors that include an imperfect reference standard and normative databases, at least in part because of the variability of normal anatomic features and instrument operator-dependent factors. Thus, relying solely on the automatic classification to diagnose glaucoma or using this as a replacement for visual field testing is not recommended. We suggest automated imaging technologies are best suited as an adjunct to the ophthalmologist’s assessment. It should also be noted that the clinical biomicroscopic examination of the optic nerve and nerve fiber layer yields a true stereoscopic view of the structures that may reveal glaucomatous features that are not visible in the computer-generated images from the imaging technologies used in this study, such as optic disc hemorrhages. These imaging instruments probably have their greatest usefulness when assisting in the diagnosis of more subtle glaucomatous optic neuropathy, particularly in situations where the nerve fiber layer and optic disc are indistinct clinically. Additionally, there is a role to assist clinicians who may not be glaucoma specialists or in the case of trainees where the additional information provided by these devices can be a valuable component of training. From a UK perspective, using automated imaging classification of imaging technologies as a triage test may prove to be cost-effective.22

In conclusion, this study offers valuable insight into the comparative diagnostic performance of automated diagnostic tests using some of their outputs (disc classification for HRT and RNFL classification for GDx enhanced corneal compensation and Spectralis OCT) that are easy to read and interpret and are encountered in a real-life setting. Automated testing is helpful in and aids glaucoma diagnosis. However, reliance solely on imaging (as used in this study) as a means of diagnosis is not recommended because some patients with severe glaucoma may be missed.

Acknowledgments. The authors thank all the GATE study participants and the staff at each of our recruiting centers for taking part in this study: NHS Grampian, Hinchingbrooke Hospital NHS Trust, Bedford Hospital NHS Trust, Moorfields Eye Hospital NHS Foundation Trust, The Royal Liverpool, and Broadgreen University Hospitals NHS Trust; the members of our independent steering committee: Colm O’Brien (chair), Anthony King, Anja Tuulonen, Russell Young, and David Wright; the staff at the GATE study office; and Gladys McPherson and the programming team based in the Centre for Healthcare Randomised Trials within the Health Services Research Unit, University of Aberdeen.

References

Footnotes and Financial Disclosures

Originally received: July 24, 2015.
Final revision: January 26, 2016.
Accepted: January 27, 2016.
Manuscript no. 2015-1245.

1 Health Services Research Unit, University of Aberdeen, Aberdeen, United Kingdom.
2 Robertson Centre for Biostatistics, University of Glasgow, Glasgow, United Kingdom.
3 Vision & Eye Research Unit, Postgraduate Institute, Anglia Ruskin University, Cambridge, United Kingdom.
4 Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom.
5 School of Medicine, University of St. Andrews, St. Andrews, United Kingdom.
6 National Institute for Health Research Biomedical Research Centre, Moorfields Eye Hospital and University College London Institute of Ophthalmology, London, United Kingdom.
7 Faculty of Health and Life Sciences, Northumbria University, Newcastle-upon-Tyne, United Kingdom.
8 Health Economics Research Unit, University of Aberdeen, Aberdeen, United Kingdom.
9 Centre for Experimental Medicine, Queen’s University, Belfast, United Kingdom.

Financial Disclosure(s):
The author(s) have made the following disclosure(s): D.G.-H.: Lecturer — Heidelberg Engineering, Heidelberg, Germany; Personal fee — National Institute for Health Research, London, UK (for imaging device evaluation); Patent – ANSWERS (for combining imaging and visual function measures); Equipment loan — Carl Zeiss Meditec, Germany; Heidelberg Engineering, Heidelberg, Germany.

Supported by the National Institute for Health Research (NIHR), Health Technology Assessment (HTA) program (grant no.: 09/22/111). The Health Services Research Unit is core funded by the Chief Scientist Office of the Scottish Government Health and Social Care Directorates. The views and opinions expressed in this article are those of the authors and do not necessarily reflect those of the NIHR HTA program, the NIHR, the National Health Service or the Department of Health, or the funders that provided institutional support for the study.

Author Contributions:

Conception and design: Banister, Cook, Burr, Ramsay, Hernández, Azuara-Blanco

Data collection: Banister, Boachie, Bourne, Cook, Ramsay, Garway-Heath, Hernández, Azuara-Blanco

Obtained funding: Bourne, Cook, Burr, Ramsay, Garway-Heath, Hernández, Azuara-Blanco

Overall responsibility: Banister, Boachie, Bourne, Cook, Burr, Ramsay, Garway-Heath, Gray, McMeekin, Hernández, Azuara-Blanco

Abbreviations and Acronyms:

CI = confidence interval; GDx = Glaucoma Diagnostics; GPS = glaucoma probability score; HRT = Heidelberg retina tomograph; MRA = Moorfields regression analysis; NFI = nerve fiber indicator; OCT = optical coherence tomography; RNFL = retinal nerve fiber layer; ROC = receiver operating characteristic.

Correspondence:
Augusto Azuara-Blanco, FRCOphth, Centre for Experimental Medicine, Institute of Clinical Sciences - Block A, Queen’s University, Grosvenor Road, Belfast, BT12 6BA United Kingdom. E-mail: a.azuara-blanco@qub.ac.uk.