Screening tests for detecting open angle glaucoma:

systematic review and meta-analysis

Graham Mowatt,1 Jennifer M. Burr,1,2 Jonathan A. Cook,1 M.A. Rehman Siddiqui,2
Craig Ramsay,1 Cynthia Fraser,1 Augusto Azuara-Blanco,2 Jonathan J Deeks,3 for the
OAG Screening Project Group

1Health Services Research Unit, Institute of Applied Health Sciences, College of Life
Sciences and Medicine, University of Aberdeen
2Eye Clinic, Aberdeen Royal Infirmary
3Department of Public Health and Epidemiology, University of Birmingham

Corresponding author: Graham Mowatt, Health Services Research Unit, 3rd Floor,
University of Aberdeen, Health Sciences Building, Foresterhill, Aberdeen AB25 2ZD,
UK; g.mowatt@abdn.ac.uk.

Funding

This paper was developed from a Health Technology Assessment on the clinical and
cost-effectiveness of screening for OAG, and funded by the UK National Institute for
Health Research Health Technology Assessment programme (project number
04/08/02). The authors’ work was independent of the funding source.

Word count: abstract, 250; main text excluding title page, abstract, tables, figures and
references, 3898.
Purpose

To assess the comparative accuracy of potential screening tests for open angle glaucoma (OAG).

Methods

Medline, Embase, Biosis (to November 2005), Science Citation Index (to December 2005) and The Cochrane Library (Issue 4 2005) were searched. Studies assessing candidate screening tests for detecting OAG in people over 40 years of age that reported true and false positives and negatives were included. Meta-analysis was undertaken using the hierarchical summary receiver operating characteristic model.

Results

Forty studies enrolling over 48,000 people reported nine tests. Most tests were reported by only a few studies. Frequency Doubling Technology (FDT) (C-20-1) was significantly more sensitive than ophthalmoscopy (30, 95% credible interval (CrI) 0 to 62) and Goldmann applanation tonometry, (GAT), (45, 95% CrI 17 to 68), while threshold standard automated perimetry (SAP) and Heidelberg Retinal Tomograph (HRT II) were both more sensitive than GAT (41, 95% CrI 14 to 64 and 39, 95% CrI 3 to 64 respectively). GAT was more specific than both FDT C-20-5 (19, 95% CrI 0 to 53) and threshold SAP (14, 95% CrI 1 to 37). Judging performance by diagnostic odds ratio, FDT, Oculokinetic perimetry and HRT II are promising tests. Ophthalmoscopy, SAP, retinal photography and GAT had relatively poor performance as single tests. These findings are based on heterogeneous data of limited quality and as such associated with considerable uncertainty.
Conclusions

No test or group of tests were clearly superior as glaucoma screening tests. Further research is required to evaluate the comparative accuracy of the most promising tests.

Introduction

Glaucoma describes a group of eye diseases in which there is progressive damage to the optic nerve, leading to impaired vision and in some cases blindness if untreated. Glaucoma is the leading cause of irreversible blindness worldwide,\textsuperscript{1,2} with open angle glaucoma (OAG) the most common form.\textsuperscript{1} Late detection is a major risk factor for blindness;\textsuperscript{1,3-5} it is estimated from population surveys that in developed countries, more than 50% of prevalent OAG is undetected,\textsuperscript{6} and this estimate is likely to be higher in developing countries. Recent evidence suggests that treatment is effective at delaying progression,\textsuperscript{7,8} thus population based screening of OAG is under consideration.\textsuperscript{6,9-11} For screening to be considered several criteria need to be met regarding the condition, the test and the screening programme.\textsuperscript{9}

Tests for glaucoma involve an assessment of structural changes at the optic nerve head, functional visual loss by visual field testing, and the level of the intraocular pressure (IOP). There are many potential tests or combinations of tests for detecting glaucoma, however to date no single test or combination of tests has been identified as an optimal screening ‘test’ for glaucoma.

The aim of this study was to assess the comparative accuracy of candidate screening tests.
Methods

Search Strategy
Highly sensitive electronic searches, using both controlled vocabulary and free text terms, were undertaken. We searched the following electronic databases: Medline (1966 – November Week 3 2005), Medline In Process (23 February and 6 December 2005), Embase (1980 – 2005 Week 49), Science Citation Index (1981 – 3 December 2005), Biosis (1985 – 30 November 2005) and Cochrane Central Register of Controlled Trials (CENTRAL) (The Cochrane Library, Issue 4 2005). In addition full text electronic searches of the American Journal of Ophthalmology (1998 – November 2005), Ophthalmology (1998 – November 2005), British Journal of Ophthalmology (1998 – November 2005), Investigative Ophthalmology & Visual Science (1998 – November 2005) and the Journal of Glaucoma (2001 – November 2005) were undertaken. Searches were restricted to English language publications. The reference lists of included studies were scanned to identify additional potentially relevant reports. Full details of the sources searched and search strategies used are available elsewhere or can be obtained by contacting the authors.

Inclusion and Exclusion Criteria
We included studies that assessed the accuracy of tests for detecting OAG in people over 40 years of age who were likely to be representative of a screening situation (i.e. no selection and no previous tests have been done) or of a glaucoma suspect population (i.e. patients identified from prior testing as possibly having glaucoma or having e.g., high IOP, or another risk factor for glaucoma but with an unconfirmed diagnosis) Both randomised (where participants were randomised to one or more tests) and observational (both cohort and case-control) studies were included. The
reference standard was either confirmed OAG on follow-up or ophthalmologist-diagnosed OAG as reported by the study. This latter reference standard required a clinical judgement by an ophthalmologist including an evaluation of the optic nerve and a measure of visual function. In addition the study had to either report or allow the calculation of true and false positives and negatives.

Non-English language reports were excluded, as were conference abstracts. Case reports and studies investigating technical aspects of a test were excluded. Case-control studies where the control group consisted of people with no ocular disease or specifically excluded people with other ocular disease, so that the spectrum of disease and non-disease was unlike that to be encountered in a screening situation, were also excluded. The spectrum of disease expected would be similar to the spectrum of the disease of the general population (e.g., more patients with mild glaucoma, less patients with severe glaucoma).

The candidate tests fell within the three broad categories of (a) structure (ophthalmoscopy; optic disc photography; retinal nerve fibre layer (RNFL) photography; Heidelberg retinal tomography (HRT) version II; GDx VCC retinal nerve fibre layer (RNFL) analyser; optical coherence tomography (OCT); retinal thickness analyser (RTA)), (b) function (oculokinetic perimetry (OKP); white-on-white standard automated perimetry (SAP) including suprathreshold and threshold; short wave-length automated perimetry (SWAP); frequency doubling technology (FDT); motion detection perimetry (MDP)) and (c) IOP (Goldmann applanation tonometry (GAT); non contact tonometry (NCT); Tonopen).

Data Abstraction and Quality Assessment
Two reviewers undertook single data extraction of the included studies. In the event of any uncertainty, the other reviewer provided advice and validated the data extraction.

Two reviewers independently assessed the quality of the included studies using a version of QUADAS adapted for assessing reports of the accuracy of screening tests for OAG. QUADAS is a quality assessment tool for use in systematic reviews of diagnostic studies. Disagreements were resolved by consensus or arbitration by a third reviewer. A ‘higher quality study’ was considered to be one that was checked ‘yes’ to questions 1 (patient spectrum representative), 3 and 4 (partial and differential verification bias avoided) and 6 and 7 (test review bias and diagnostic review bias avoided) of the adapted QUADAS checklist.

**Statistical Methods**

After data extraction a ‘common’ (most frequently reported) cutoff for each test was selected following discussion by two ophthalmologists (JB, RS). Summary receiver operating characteristic (SROC) curves were produced for each test where two or more studies reported estimates of sensitivity and specificity at the common cutoff. Meta-analysis models were fitted using the hierarchical summary receiver operating characteristic (HSROC) model in WinBUGS 1.4. Normally distributed random effects were assumed with non-informative uniform priors. No adjustment was made for the correlation between results from paired studies, as the level of information required is rarely reported. Summary sensitivity, specificity and diagnostic odds ratios (DORs) at the operating point were reported for each model as median and 95% credible interval (CrI). A DOR is a single indicator of test performance and is the ratio of the odds of testing positive in those with the disease...
relative to the odds of testing positive in those without the disease. It can be calculated from the sensitivity and specificity:

\[
DOR = \frac{\text{sensitivity}}{1 - \text{sensitivity}} \div \frac{1 - \text{specificity}}{\text{specificity}}
\]

Credible intervals are the Bayesian equivalent of confidence intervals. A simplified model, which assumed a symmetrical ROC shape, was used where limited data caused convergence problems under the full model. Sensitivity analysis was undertaken by examining separately the results of the higher quality studies, using HSROC analysis where more than one higher quality study reported the same test.

Comparisons between tests were made in two ways. First, studies which directly compared participants who either received all tests or were randomised to different tests were identified, and the direct comparisons inspected. Secondly, an indirect comparison between tests, for all tests reported by two or more studies were modelled together in a single HSROC model to formally compare test performance. Pair-wise differences in sensitivity and specificity between tests were assessed from the median difference and corresponding 95% CrI.

Results

Trial Flow

Figure 1 shows the flow of studies through the review. Out of a total of 5918 titles/abstracts screened, 877 potentially relevant full text articles were obtained, with 40 studies, published in 46 reports, meeting the inclusion criteria.

Study Characteristics and Methodological Quality

The characteristics of the included studies are shown in Table 1. Twenty studies were population-based and representative of a screening setting while 20 studies
were considered representative of a glaucoma-suspect population referred from primary care, of which eight were cohort studies\textsuperscript{40-47} and 12 were case-control studies.\textsuperscript{48-61} Seven studies\textsuperscript{18,34,40,43,44,48,58} used the first and best reference standard of OAG confirmed on longitudinal follow-up while the remainder used ophthalmologist-diagnosed OAG.

The 40 studies enrolled over 48,000 people, with over 39,000 included in the analysis. The studies took place from 1963 to 2004. In 26 studies reporting participant gender, \(51\%\) were women. The median (range) age of participants across studies was 60.5 years (13 to 97 years). The reports included a number of major population-based prevalence surveys, such as the Baltimore Eye Survey,\textsuperscript{25,31} the Blue Mountains Eye Study,\textsuperscript{23} the Crete, Greece Glaucoma Study,\textsuperscript{27} the Dalby Population Survey,\textsuperscript{17} the Egna-Neumarkt Study,\textsuperscript{18} the Framingham Eye Study,\textsuperscript{43} the Glaucoma Screening Study (GLASS),\textsuperscript{24,26} the Groningen Longitudinal Glaucoma Study,\textsuperscript{53,54,59} the Rhondda Valley Study,\textsuperscript{22} the Rotterdam Study,\textsuperscript{38} the Segovia Study\textsuperscript{16} and the Visual Impairment Project.\textsuperscript{37}

The included studies reported the following tests: ophthalmoscopy (seven studies); optic disc photography (six studies); retinal nerve fibre layer (RNFL) photography (four studies); Heidelberg retinal tomograph (HRT) II (three studies); oculokinetic perimetry (OKP) (four studies); standard automated perimetry (SAP) (14 studies); frequency doubling technology (FDT) (eight studies); Goldmann applanation tonometry (GAT) (nine studies); non contact tonometry (NCT) (one study). No reports of GDx VCC, OCT, RTA, SWAP, MDP or Tonopen were identified that met our inclusion criteria.

Figure 2 summarises the results of the quality assessment for the 40 included studies. Study quality was variable, only eight studies\textsuperscript{20,21,30,34,38,39,45,46} met the specified criteria for higher quality studies.
Quantitative Data Synthesis

- Individual tests

The sensitivity and specificity of the individual tests included in the HSROC meta-analysis models are shown in Figure 3 and Appendix 1, which also includes DORs.

DORs ranged from 10 for FDT C-20-5 to 181 for FDT C-20-1, with higher DORs indicating a better ability to differentiate between disease and non-diseased. There was statistical heterogeneity (variability in outcome beyond what would be expected by chance) across studies for most tests. Ophthalmoscopy, retinal photography (optic disc photography and RNFL photography), GAT, standard automated perimetry (threshold and suprathreshold) and FDT C20-5 were all relatively poorly performing tests based on lower DORs (range 10-30).

Eight studies met the criteria for higher quality studies, including six population-based studies and two cohort studies, and test accuracy data are detailed in Table 2. For both SAP threshold and FDT C-20-5, higher quality studies reported lower values for both sensitivity and specificity when compared with all studies, while two FDT C-20-5 studies not meeting the criteria for higher quality reported very high sensitivity values (98% and 100% respectively). For optic disc photography, compared with all studies, the higher quality studies reported similar sensitivity (74% versus 73%) but lower specificity (82% versus 89%). For HRT II, compared with all studies, the higher quality studies reported higher sensitivity (93% versus 86%) but slightly lower specificity (85% versus 89%).

Seven studies reported test accuracy in different stages of glaucoma. Of those reporting the same tests for different stages of glaucoma, Leong and colleagues reported a sensitivity of 72% for SAP (suprathreshold) for early stage
glaucoma while Enger and colleagues\cite{51} and Katz and colleagues\cite{24} both reported a sensitivity of 97% for SAP (threshold) for early/moderate stage glaucoma.

- Studies directly comparing tests

Six studies directly compared two or more of the following tests for detection of OAG: optic disc photography, HRT II, SAP, FDT, GAT.\cite{23,30,34,36,46,55} Table 3 shows the common cut-off selected, sensitivity, specificity, DORs and relative DORs for these studies. In each study SAP (either suprathreshold or threshold) was included as a comparator. DORs for the tests ranged from 4 for SAP threshold\cite{46} to 75 for HRT II\cite{30} (Table 3). In terms of relative DORs, compared with SAP, GAT performed better in one study\cite{36} but worse in another\cite{23} (statistically significant), HRT II performed better than SAP in one study\cite{30} (statistically significant) but worse in another,\cite{55} FDT C-20-5\cite{30} and FDT C-20 matrix\cite{46} performed better than SAP, while optic disc photography\cite{34} showed a broadly similar performance.

- Indirect comparisons in a single HSROC model

The results of the indirect comparisons in a single HSROC model are shown in Table 4. From the large number of comparisons undertaken, six showed a statistically significant difference between tests (four in terms of sensitivity and two in terms of specificity). There was evidence that, at the common cut-off, FDT C-20-1 was significantly more sensitive than ophthalmoscopy (30, 95% CrI 0 to 62) and GAT (45, 95% CrI 17 to 68), and that both SAP threshold (41, 95% CrI 14 to 64) and HRT II (39, 95% CrI 3 to 64) were significantly more sensitive than GAT. There was also evidence that GAT was significantly more specific than both FDT C-20-5 (19, 95% CrI 0 to 53) and SAP threshold (14, 95% CrI 1 to 37). Other differences in accuracy
between tests may well exist which could not be detected due to the high level of uncertainty. The wide credible intervals reflected the small number of studies reporting each test and the generally high level of heterogeneity. Due to the imprecision in the estimates, no test (or even a group of tests) was clearly more accurate, based upon a 5% significance level. Further analysis, at 10% and 20% levels of significance, identified additional statistically significant comparisons (Table 4). For example, in terms of sensitivity, at a 10% significance level FDT C-20-1 was better than SAP suprathreshold and at a 20% level better than optic disc photography, RNFL photography and FDT C-20-5. OKP was better than GAT at a 10% level and HRT II better than opthalmoscopy at a 20% level. In terms of specificity, at a 20% level FDT C-20-1 was better than SAP threshold and FDT C-20-5.

**Discussion**

To our knowledge, this is the first systematic review of screening and diagnostic tests in glaucoma and includes 40 studies enrolling over 48,000 people and reporting nine tests. Most tests were reported by only a few, mostly heterogeneous, studies. The included studies reported tests of structure (ophthalmoscopy; optic disc photography, RNFL photography, HRT II), visual function (FDT, OKP, SAP) and IOP (GAT, NCT). Other tests were considered, including those of structure (GDx VCC, OCT, RTA), visual function (SWAP; MDP), or using Tonopen to measure IOP. However, no studies using these tests met our inclusion criteria in terms of reporting of test accuracy outcomes.

A systematic review of test accuracy is unlikely to identify the best test but can identify more promising tests. It is difficult to rank tests on paired values of sensitivity and specificity, as a highly specific test may be associated with a low sensitivity and vice versa. The choice of test depends on the importance of the trade
off between missed cases, and false positives. OAG affects an estimated 2% of the adult population; a test of low specificity would be likely to overburden a health service with people who do not have glaucoma and cause unnecessary anxiety for a many individuals, equally a test of low sensitivity would miss treatable disease which might be unacceptable to society. The DOR, a single measure of test accuracy, is a useful measure for comparing accuracy of several tests in a meta-analysis. Based on a DOR ≥50, FDT C-20-1 and OKP (both tests of visual function) and HRT II (a test of glaucomatous optic neuropathy) merit further evaluation as to their performance as screening tests for glaucoma. It should be noted that these findings are relevant to the common cutoff point selected for each test; selection was based on the most frequently reported cut-off and when several cut-offs were reported the cut-off most likely to represent early glaucoma. Furthermore these findings are based on heterogeneous data of limited quality and as such are associated with considerable uncertainty.

Methods of meta-analysis of diagnostic accuracy which combine studies where both sensitivity and specificity vary have been available since 1990 and are continuing to evolve. These methods are based on the idea of a trade-off relationship between sensitivity and specificity, as occurs when studies vary in threshold, and aim to estimate the shape and position of the underlying Receiver Operating Curve. From the estimate of this curve it is possible to identify “operating points”. The approach adopted in our review identifies the average operating point for each test, and makes comparisons between them, based upon those studies reporting each test that share a common cut point. The Cochrane Collaboration are commencing publication of systematic reviews of diagnostic test accuracy and the analytical approach we have followed is the one that they are recommending. Estimation of a summary point specific to a test being used at a common threshold
obtains the best estimate of test accuracy in parameters that are clinically meaningful. The trade-off between sensitivity and specificity is important in judging the performance of a test and is best depicted by a ROC curve across different cutoff points. However, the included studies did not usually provide information across the whole range of cutoff points to allow such analyses to be undertaken.

We used a Bayesian Hierarchical SROC model as standard methods for meta-analysis do not address the issue of threshold effect and are therefore not appropriate. A number of different levels of analyses were undertaken, including an analysis where all tests were modelled simultaneously using this Bayesian approach. This allowed indirect comparison of sensitivities and specificities to be made, in addition to allowing DORs to be calculated, which is one of the advantages of the Bayesian method adopted. To produce results that are comparable to those from standard methods of meta-analysis we did not use informative priors.

In addition to providing sensitivity and specificity estimates we also reported the DOR results. Some meta-analysis models can only provide the DOR estimate and therefore we included this measure for comparability. A strength of the DOR is that it is a mathematically robust measure, (like the standard odds ratio) and represents diagnostic accuracy as a single value. However, a disadvantage is that different combinations of sensitivity and specificity values can lead to the same DOR value.

To be included studies had to meet specific inclusion criteria. The validity of indirect comparisons does depend upon assumptions regarding the characteristics of the included studies; however the indirect method is formally performing the comparison that users of the report are likely to make when assessing the pooled results for the individual tests. As such this method of indirect comparisons serves
an important purpose and reaffirms the lack of certainty about which test is indeed the best.

There are many potential sources of bias in primary diagnostic accuracy studies. Despite the huge volume of literature, no good quality studies were found providing a positive response to all questions on the modified QUADAS checklist. Based on limited evidence, of tests reported by higher quality studies, including the three tests that were considered to merit further evaluation, estimates of sensitivity and specificity varied according to study quality.

There is no universally agreed optimal reference standard for the diagnosis of OAG, although progressive structural optic neuropathy has been proposed as the best possible reference standard. In this review either of two reference standards were considered. There was no obvious pattern in terms of the sensitivity and specificity of the tests in the seven studies that used the first and best reference standard of OAG confirmed on longitudinal follow-up compared with the remainder that used ophthalmologist-diagnosed OAG. Although the latter is suboptimal compared with the former, it is the accepted reference standard in clinical practice. However establishing a reference standard in glaucoma is problematic, as in some people optic disc damage precedes visual field loss while in others the reverse is the case.

The accuracy of a test may vary according to the population in which it is performed. Samples with higher prevalence often arise through preferential inclusion of suspect cases, which shifts the disease severity to include more moderate and severe disease, and since it is easier to differentiate between severely diseased and non-diseased people, a test would be expected to report improved (apparent) sensitivity and specificity. Therefore studies with a significantly higher prevalence than expected in a screening population should be interpreted with this limitation in
These studies, including two that met the criteria for higher quality studies, tended to recruit their participants through media advertising rather than contacting individuals in a predefined population and can be considered to be more representative of screening in higher-risk populations.

Twenty of the 40 studies included were hospital-based, which enriched population, and likely to include a disproportionate number of participants with high IOP, and with previous experience of tests, potentially leading to over optimistic performance estimates. The majority of the case-control studies identified applied stringent criteria for inclusion such as visual acuity of 6/9, or no other ocular disease and as such were highly prone to bias. To minimise this spectrum bias, case-control studies (n=57) where the participants were considered unrepresentative of a case-mix found in a general population where OAG screening would be carried out were excluded from the review.

In the meta-analysis models for the individual tests, statistical heterogeneity was evident across most studies. Empirically, there was no obvious single cause for the heterogeneity, but potential contributory factors include differences in populations, study design, setting, prevalence and severity of glaucoma within studies. Other factors include differences in reference standard, and in tests included within the same category (e.g. different types of perimetry and ophthalmoscopy have a large number of variants, potentially leading to heterogeneity in discriminatory power across studies reporting those tests), and the extent to which studies were affected by other potential biases (e.g. partial and differential verification bias, incorporation bias, test and diagnostic review bias).

**Limitations**
Relatively few studies were identified for each test and it was not possible to perform sensitivity analysis based on study design. The common cutoff chosen for each test was the one most frequently reported across the included studies for that test, although this may not be the most appropriate. The majority of the studies were poorly reported, an issue that has been highlighted in recent literature. Only six of the 40 studies directly compared two or more tests. It was not possible to provide summary results of studies that directly compared tests because of small numbers. Studies not providing sufficient information to allow the calculation of 2 x 2 tables were excluded, although they may have contributed information in terms of sensitivity and specificity.

Systematic reviews provide a robust and rigorous evaluation of the available evidence, but by their nature as new studies are published the review requires updating. Since the completion of our meta-analysis further studies have been published on the performance of the tests included in this review. These include population-based studies in the USA, UK, Hungary, Japan and China. These studies provide additional information on the performance of FDT perimetry alone, in combination with GDx VCC, and combined with an IOP measurement and data on the performance of HRT II in an elderly population in the UK and in a community screening programme in Japan, comparing HRT II with non-mydriatic fundus photography. Although systematic reviews rapidly become out of date, which is a limitation, a strength of a systematic review is that the methods are transparent and reproducible such that the review can be updated as further data become available in the future. Priorities for future research and optimal study designs can also be identified.

Implications for Practice and Recommendations for Research
Ideally, a screening test for OAG should be safe, easy to administer and interpret, portable, quick, and acceptable to the people who are to be tested, and sufficiently valid to distinguish between those who do and do not have OAG. Many potential screening tests for glaucoma are available. Of the many candidate tests, no one test or group of tests was clearly more accurate. Based on limited data, relatively poorly performing tests, ophthalmoscopy, standard automated perimetry, retinal photography, and Goldmann applanation tonometry, were identified.

Frequency doubling technology, (C 20-1), Heidelberg Retinal Tomography and oculokinetic perimetry were identified as having better diagnostic performance than other candidate tests, although these findings were based on poor quality evidence. Further studies should evaluate the most promising tests in directly comparative studies in a relevant population.

Acknowledgements
This paper was developed from a Health Technology Assessment on the clinical and cost-effectiveness of screening for OAG, and funded by the UK National Institute for Health Research Health Technology Assessment programme (project number 04/08/02). The Health Services Research Unit and the Health Economics Research Unit are both core funded by the Chief Scientist Office of the Scottish Government Health Directorates. The views expressed in this report are those of the authors and not necessarily those of the funders. The authors are grateful to members of the OAG project group for assistance and comments on all stages of the review. Other members of the OAG Screening Project Group were: Rodolfo Hernandez, Luke Vale, Tania Lourenco, Cynthia Fraser, John Cairns, Richard Wormald, Stephen McPherson, David Wright, Kannaiyan Rabindranath, Rod Taylor, Norman Waugh and Adrian Grant.
References


11 *Screening for glaucoma: an update - non-systematic review (project)*. Agence d’Evaluation des Technologies et des Modes d’Intervention en Sante (AETMIS);


Figure 1  Flow of studies through review process.

Titles and abstracts screened (n=5918)

Not relevant (n=5041)

Full articles screened (n=877)

Excluded studies (n=831):
Case-control studies with participants not representative of a screening situation or of a glaucoma suspect population referred from GP or optometric practice (n=57)
Failed to meet one or more inclusion criteria in terms of study design, participants, index tests, reference standard or outcomes reported (n=774)

Met inclusion criteria (40 studies, 46 publications)
Figure 2  Results of the quality assessment of the 40 included studies.

Notes:
1. Where a study reported two or more tests and they differed in meeting any one QUADAS criterion, e.g. one test checked ‘Yes’ and one checked ‘No’, then the ‘No’ answer was taken to represent the study for that item. This applied to the following studies for the following items:
   (i) Robin 2005.30 Question 5 was Yes for FDT, No for HRT II, No for SAP.
   (ii) Wang 1998.36 Question 5 was Yes for ophthalmoscopy, Yes for RNFL photography, No for SAP, No for GAT. Question 9 was Yes for RNFL photography, No for ophthalmoscopy, No for SAP, No for GAT.
   (iii) Marraffa 1989.44 Question 11 was Yes for Henson, Unclear for the other perimetry tests.
   (iv) Spry 2005.46 Questions 5 and 7 were Yes for FDT, No for SAP.
   (v) Harper 1994.52 Questions 5 and 12 were Yes for OKP, No for SAP.
   (vi) Quigley 1980.57 Question 5 was Yes for RNFL, No for optic disc photography.
2. Unclear means insufficient information was provided to determine whether the item should be checked Yes or No.
Figure 3 Summary of sensitivity and specificity of tests included in the HSROC meta-analysis models.

Notes:
1. Number of studies: ophthalmoscopy (n = 5), optic disc photography (n = 6), RNFL photography (n = 4), HRT II (n = 3), FDT C-20-1 (n = 3), FDT C-20-5 (n = 5), OKP (n = 4), SAP suprathreshold (n = 9), SAP threshold (n = 5), GAT (n = 9).
Table 1  Characteristics of the included studies.

<table>
<thead>
<tr>
<th>Study id</th>
<th>Index test(s)</th>
<th>Test(s) carried out and interpreted by</th>
<th>Reference Standard</th>
<th>Enrolled (people)</th>
<th>Analysed</th>
<th>Mean age (range)</th>
<th>Gender</th>
<th>Country</th>
<th>Time period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anton 2004</td>
<td>GAT</td>
<td>Ophthalmologists</td>
<td>Ophthalmic examination</td>
<td>569</td>
<td>510</td>
<td>(40 to 79)</td>
<td>M: 232; F: 278</td>
<td>Spain (Segovia Study)</td>
<td>N/S</td>
</tr>
<tr>
<td>Bengtsson 1980</td>
<td>GAT</td>
<td>Ophthalmologists</td>
<td>Ophthalmic examination</td>
<td>1938</td>
<td>1511</td>
<td>(55 to 69)</td>
<td>N/S</td>
<td>Sweden (Dalby Population Survey)</td>
<td>1977 - 1978</td>
</tr>
<tr>
<td>Bonomi 2001</td>
<td>GAT</td>
<td>Ophthalmologists</td>
<td>Follow-up confirmation</td>
<td>5816</td>
<td>4297 eyes of 4297 people</td>
<td>(40 to 80+)</td>
<td>M: 1882; F: 2415</td>
<td>Italy (Egna-Neumarkt Study)</td>
<td>N/S</td>
</tr>
<tr>
<td>Detry-Morel 2004</td>
<td>FDT C-20-5</td>
<td>Residents in training, paramedical staff</td>
<td>Ophthalmic examination</td>
<td>1802</td>
<td>3211 eyes of 1620 people</td>
<td>63 (22 to 97)</td>
<td>M: 680; F: 940</td>
<td>Belgium</td>
<td>October 1999</td>
</tr>
<tr>
<td>Harasymowycz 2005</td>
<td>HRT II</td>
<td>Ophthalmic photographer</td>
<td>Ophthalmic examination</td>
<td>303</td>
<td>264 right eyes, 265 left eyes of 271 people</td>
<td>62.2 (SD 11.6)</td>
<td>M: 90; F: 179</td>
<td>Canada</td>
<td>August 2003 – February 2004</td>
</tr>
<tr>
<td>Hollows 1966</td>
<td>GAT</td>
<td>Ophthalmologists</td>
<td>Ophthalmic examination</td>
<td>4608</td>
<td>4231</td>
<td>55 (40 to 74)</td>
<td>Approx: M: 3639; F: 592</td>
<td>UK (Rhondda Valley Study)</td>
<td>Summer 1963</td>
</tr>
<tr>
<td>Ivers 2001</td>
<td>SAP suprathreshold; GAT</td>
<td>N/S</td>
<td>Ophthalmic examination</td>
<td>4433</td>
<td>3654 (both tests)</td>
<td>(49 to 97)</td>
<td>M: 1582; F: 2072</td>
<td>Australia (Blue Mountains Eye Study)</td>
<td>1992 - 1994</td>
</tr>
<tr>
<td>Study id</td>
<td>Index test(s)</td>
<td>Test(s) carried out and interpreted by</td>
<td>Reference Standard</td>
<td>Enrolled (people)</td>
<td>Analysed</td>
<td>Mean age (range)</td>
<td>Gender</td>
<td>Country</td>
<td>Time period</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------------</td>
<td>----------------------------------------</td>
<td>--------------------</td>
<td>-------------------</td>
<td>----------</td>
<td>------------------</td>
<td>--------</td>
<td>---------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Mansberger 2005²⁸</td>
<td>FDT C-20-5</td>
<td>N/S</td>
<td>Ophthalmic examination</td>
<td>296</td>
<td>251 eyes of 251 people</td>
<td>45 (30 to 65)</td>
<td>M: 117; F: 174</td>
<td>India</td>
<td>N/S</td>
</tr>
<tr>
<td>Mundorf 1989²⁹</td>
<td>SAP suprathreshold</td>
<td>N/S</td>
<td>Ophthalmic examination</td>
<td>145</td>
<td>145</td>
<td>71</td>
<td>M: 40; F: 105</td>
<td>USA</td>
<td>N/S</td>
</tr>
<tr>
<td>Robin 2005³⁰</td>
<td>Ophthalmoscopy;</td>
<td>Appropriately trained staff</td>
<td>Ophthalmic examination</td>
<td>704</td>
<td>261 eyes of 261 people (all tests)</td>
<td>65</td>
<td>M: 281; F: 378</td>
<td>Australia</td>
<td>Nov 2001</td>
</tr>
<tr>
<td>Weih 2001³⁴</td>
<td>Ophthalmoscopy</td>
<td>N/S</td>
<td>Consensus by panel of ophthalmologists, based on results of ophthalmic examination</td>
<td>4744</td>
<td>4636</td>
<td>59 (SD 12)</td>
<td>M: 2230; F: 2514</td>
<td>Australia (Visual Impairment Project)</td>
<td>1992 – 1996</td>
</tr>
<tr>
<td>Wolfs 1999³⁸</td>
<td>Optic disc</td>
<td>Technicians</td>
<td>Ophthalmic examination</td>
<td>6777</td>
<td>5143 eyes of 5143 people</td>
<td>(55 and over)</td>
<td>N/S</td>
<td>Netherlands (Rotterdam Study)</td>
<td>N/S</td>
</tr>
<tr>
<td>Yamada 1999³⁹</td>
<td>OKP; FDT C-20-1</td>
<td>Technicians</td>
<td>Decision of glaucoma specialists, based on ophthalmic records</td>
<td>259</td>
<td>175 eyes of 175 people (OKP); 240 eyes of 240 people (FDT)</td>
<td>FDT: 59.6 (SD 14.7); OKP: 58.8 (SD 15.6)</td>
<td>M: 108; F: 135</td>
<td>USA</td>
<td>N/S</td>
</tr>
<tr>
<td>Study id</td>
<td>Index test(s)</td>
<td>Test(s) carried out and interpreted by</td>
<td>Reference Standard</td>
<td>Enrolled (people)</td>
<td>Analysed</td>
<td>Mean age (range)</td>
<td>Gender</td>
<td>Country</td>
<td>Time period</td>
</tr>
<tr>
<td>----------------</td>
<td>----------------------------------------</td>
<td>---------------------------------------------------</td>
<td>--------------------</td>
<td>-------------------</td>
<td>----------</td>
<td>------------------</td>
<td>--------</td>
<td>---------------</td>
<td>-------------</td>
</tr>
</tbody>
</table>
| **Christoffersen 1995**<sup>19</sup>  
Patient source: general practice | OKP | GPs, medical secretaries | Ophthalmic examination | 195 | 187 | 57 (40 to 84) | M: 51; F: 136 | Norway | N/S |
| **Vernon 1990**<sup>32</sup>  
Patient source: general practice | Ophthalmoscopy; SAP suprathreshold; NCT | Ophthalmoscopy: experienced ophthalmologists; NCT/SAP: non-ophthalmological trained staff | Ophthalmic examination | 988 | 854 (ophth); 855 (SAP); 874 (NCT) | 65 | M: 374; F: 500 | UK | N/S |
| **Wang 1998**<sup>36</sup>  
Patient source: general practice | Ophthalmoscopy; SAP suprathreshold; GAT [RNFL photography] | N/S | Ophthalmic examination | 530 (from primary care clinic) | 400 (ophth); 214 (SAP); 357 (GAT) [136 (RNFL photo)] | (40 to 65+) | M: 111; F: 294 | USA | Jul 1991 – Feb 1992 |

**Population-based studies (case-control)**

<table>
<thead>
<tr>
<th>Study id</th>
<th>Index test(s)</th>
<th>Test(s) carried out and interpreted by</th>
<th>Reference Standard</th>
<th>Enrolled (people)</th>
<th>Analysed</th>
<th>Mean age (range)</th>
<th>Gender</th>
<th>Country</th>
<th>Time period</th>
</tr>
</thead>
</table>
| **Vitale 2000**<sup>34</sup>  
Patient source: Cases and controls: sample of patients with and without glaucoma from the Baltimore Eye Study Follow-up | Optic disc photography; SAP suprathreshold | Experienced technicians | Follow-up confirmation | 249 | 182 (disc photo); 228 (SAP); | 68 | M: 100; F: 149 | USA (Baltimore Eye Study Follow-up Study) | 1994 |
<table>
<thead>
<tr>
<th>Study id</th>
<th>Index test(s)</th>
<th>Test(s) carried out and interpreted by</th>
<th>Reference Standard</th>
<th>Enrolled (people)</th>
<th>Analysed</th>
<th>Mean age (range)</th>
<th>Gender</th>
<th>Country</th>
<th>Time period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ekstrom 1993</td>
<td>GAT</td>
<td>N/S</td>
<td>Follow-up confirmation</td>
<td>760</td>
<td>413</td>
<td>(65 to 74)</td>
<td>M: 364; F: 396</td>
<td>Sweden (Tierp Glaucoma Survey)</td>
<td>Mar 1984 - Mar 1986</td>
</tr>
<tr>
<td>Hammond 1979</td>
<td>Ophthalmoscopy</td>
<td>Nurses skilled in use of the ophthalmoscope</td>
<td>Ophthalmic examination</td>
<td>219</td>
<td>188</td>
<td>(21 and over)</td>
<td>N/S</td>
<td>USA</td>
<td>N/S</td>
</tr>
<tr>
<td>Khong 2001</td>
<td>FDT C-20-5</td>
<td>N/S</td>
<td>Ophthalmic examination</td>
<td>228</td>
<td>113</td>
<td>68.5</td>
<td>M: 104; F: 119</td>
<td>Australia</td>
<td>Dec 1999 – Jan 2000</td>
</tr>
<tr>
<td>Leibowitz 1980</td>
<td>GAT</td>
<td>Generally performed by 2nd or 3rd year residents in ophthalmology</td>
<td>Follow-up confirmation</td>
<td>2631</td>
<td>574</td>
<td>(&lt;65 to 75+)</td>
<td>M: 272; F: 302</td>
<td>USA (Framingham Eye Study)</td>
<td>Feb 1973 – Feb 1975</td>
</tr>
<tr>
<td>Marraffa 1989</td>
<td>SAP suprathreshold</td>
<td>Ophthalmologists</td>
<td>Follow-up confirmation</td>
<td>104</td>
<td>182 eyes of 104 people</td>
<td>54.3</td>
<td>M: 45; F: 59</td>
<td>Italy</td>
<td>N/S</td>
</tr>
<tr>
<td>Schultz 1995</td>
<td>Optic disc photography</td>
<td>Carried out: N/S Interpreted: 3rd year ophthalmology</td>
<td>Ophthalmic examination</td>
<td>258</td>
<td>365 eyes of ? people</td>
<td>(&lt;40 to &gt;70)</td>
<td>M: 112; F: 144; Unknown: 2</td>
<td>USA</td>
<td>N/S</td>
</tr>
<tr>
<td>Study id</td>
<td>Index test(s)</td>
<td>Test(s) carried out and interpreted by</td>
<td>Reference Standard</td>
<td>Enrolled (people)</td>
<td>Analysed</td>
<td>Mean age (range)</td>
<td>Gender</td>
<td>Country</td>
<td>Time period</td>
</tr>
<tr>
<td>---------------------</td>
<td>---------------------------------------------------</td>
<td>----------------------------------------------</td>
<td>--------------------</td>
<td>-------------------</td>
<td>----------------</td>
<td>-----------------</td>
<td>--------</td>
<td>-------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Spry 2005[^46]</td>
<td>SAP threshold; FDT C-20 matrix</td>
<td>clinic staff trained in visual field testing; FDT: N/S</td>
<td>Ophthalmic examination</td>
<td>48</td>
<td>48 (both tests)</td>
<td>67.3 (SD 13.5)</td>
<td>M: 24; F: 24</td>
<td>UK</td>
<td>Oct 2003–Jan 2004</td>
</tr>
<tr>
<td>Patient source:</td>
<td>hospital eye service</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Theodossiades 2001[^47]</td>
<td>Ophthalmoscopy</td>
<td>Optometrists</td>
<td>Ophthalmic examination</td>
<td>50</td>
<td>50 eyes of 50 people</td>
<td>N/S</td>
<td>N/S</td>
<td>UK</td>
<td>N/S</td>
</tr>
<tr>
<td>Patient source:</td>
<td>glaucoma clinics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Airaksinen 1984[^48]</td>
<td>RNFL photography</td>
<td>N/S</td>
<td>Follow-up confirmation</td>
<td>142</td>
<td>132 eyes of 132 people</td>
<td>Glaucoma: 62 (SD 20.5); Normal: 54 (SD 16.9); OHT: 57 (SD 12.7)</td>
<td>N/S</td>
<td>Canada + Finland</td>
<td>N/S</td>
</tr>
<tr>
<td>Patient source:</td>
<td>not stated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anton 1997[^49]</td>
<td>SAP threshold</td>
<td>Uncertain</td>
<td>Ophthalmic examination</td>
<td>180</td>
<td>180 eyes of 180 people</td>
<td>Glaucoma: 61 (SD 8); Normal: 59 (SD 9)</td>
<td>N/S</td>
<td>Spain</td>
<td>N/S</td>
</tr>
<tr>
<td>Patient source:</td>
<td>Cases and controls: glaucoma unit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Damato 1989[^50]</td>
<td>OKP</td>
<td>Staff experienced in perimetry</td>
<td>Ophthalmic examination</td>
<td>102</td>
<td>102 eyes of 102 people</td>
<td>Glaucoma: 57.3;</td>
<td>N/S</td>
<td>UK</td>
<td>N/S</td>
</tr>
<tr>
<td>Patient source:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Already suspect population (case-control studies)

<table>
<thead>
<tr>
<th>Study id</th>
<th>Index test(s)</th>
<th>Test(s) carried out and interpreted by</th>
<th>Reference Standard</th>
<th>Enrolled (people)</th>
<th>Analysed</th>
<th>Mean age (range)</th>
<th>Gender</th>
<th>Country</th>
<th>Time period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airaksinen 1984</td>
<td>RNFL photography</td>
<td>N/S</td>
<td>Follow-up confirmation</td>
<td>142</td>
<td>132 eyes of 132 people</td>
<td>Glaucoma: 62 (SD 20.5); Normal: 54 (SD 16.9); OHT: 57 (SD 12.7)</td>
<td>N/S</td>
<td>Canada + Finland</td>
<td>N/S</td>
</tr>
<tr>
<td>Patient source:</td>
<td>not stated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anton 1997</td>
<td>SAP threshold</td>
<td>Uncertain</td>
<td>Ophthalmic examination</td>
<td>180</td>
<td>180 eyes of 180 people</td>
<td>Glaucoma: 61 (SD 8); Normal: 59 (SD 9)</td>
<td>N/S</td>
<td>Spain</td>
<td>N/S</td>
</tr>
<tr>
<td>Patient source:</td>
<td>Cases and controls: glaucoma unit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Damato 1989</td>
<td>OKP</td>
<td>Staff experienced in perimetry</td>
<td>Ophthalmic examination</td>
<td>102</td>
<td>102 eyes of 102 people</td>
<td>Glaucoma: 57.3;</td>
<td>N/S</td>
<td>UK</td>
<td>N/S</td>
</tr>
<tr>
<td>Patient source:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study id</td>
<td>Index test(s)</td>
<td>Test(s) carried out and interpreted by</td>
<td>Reference Standard</td>
<td>Enrolled (people)</td>
<td>Analysed</td>
<td>Mean age (range)</td>
<td>Gender</td>
<td>Country</td>
<td>Time period</td>
</tr>
<tr>
<td>----------</td>
<td>---------------</td>
<td>---------------------------------------</td>
<td>--------------------</td>
<td>-------------------</td>
<td>----------</td>
<td>-----------------</td>
<td>--------</td>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>Cases: not stated Controls: dermatology ward, hospital staff, relatives/friends of patients, patients with unilateral non-glaucomatous disease affecting the fellow eye</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enger 1987</td>
<td>SAP threshold</td>
<td>N/S</td>
<td>Ophthalmic examination</td>
<td>112</td>
<td>170 eyes of 112 people</td>
<td>Glaucoma: 61 (28 to 80); Normal: 51 (26 to 75)</td>
<td>N/S</td>
<td>USA</td>
<td>N/S</td>
</tr>
<tr>
<td>Harper 1994</td>
<td>OKP; SAP suprathreshold</td>
<td>Uncertain</td>
<td>Ophthalmic examination</td>
<td>212</td>
<td>193 (OKP); 212 (SAP)</td>
<td>Glaucoma: 67.8 (43 to 85); Normal: 61.5 (41 to 85)</td>
<td>N/S</td>
<td>UK</td>
<td>N/S</td>
</tr>
<tr>
<td>Heeg 2005</td>
<td>FDT C-20-1; FDT C-20 full threshold</td>
<td>N/S</td>
<td>Ophthalmic examination</td>
<td>1112</td>
<td>208 (FDT C-20-1); 1112 (FDT C-20 full threshold)</td>
<td>Glaucoma: 65 (13 to 91); Normal: 63 (33 to 94)</td>
<td>Eligible; Glaucoma: M: 509; F: 542 Normal:</td>
<td>Netherlands (Groningen Longitudinal Glaucoma Study)</td>
<td>Jul 2000 – Jun 2001</td>
</tr>
<tr>
<td>Study id</td>
<td>Index test(s)</td>
<td>Test(s) carried out and interpreted by</td>
<td>Reference Standard</td>
<td>Enrolled (people)</td>
<td>Analysed</td>
<td>Mean age (range)</td>
<td>Gender</td>
<td>Country</td>
<td>Time period</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------------------</td>
<td>----------------------------------------</td>
<td>---------------------</td>
<td>-------------------</td>
<td>-----------</td>
<td>-----------------</td>
<td>--------</td>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>department</td>
<td>Controls: old people’s homes, blood bank, other public places</td>
<td>HRT II; SAP suprathreshold</td>
<td>Optometrists</td>
<td>66</td>
<td>66 eyes of 66 people (both tests)</td>
<td>Glaucoma: 69; Normal: 60</td>
<td>M: 118; F: 119</td>
<td>UK</td>
<td>N/S</td>
</tr>
<tr>
<td>Ieong 2003</td>
<td>Patient source: Cases: glaucoma subjects Controls: partners of cases, optometrist practice</td>
<td>FDT C-20-1</td>
<td>N/S</td>
<td>108</td>
<td>160 eyes of 108 people</td>
<td>Glaucoma: 64 (35 to 85); Normal: 46 (18 to 81)</td>
<td>USA</td>
<td>N/S</td>
<td></td>
</tr>
<tr>
<td>Johnson 1999</td>
<td>Patient source: not stated</td>
<td>Optic disc photography; RNFL photography</td>
<td>Ophthalmologists</td>
<td>175</td>
<td>294 eyes of ? people (both tests)</td>
<td>Readable photos: Glaucoma: 52.7 (SD 2.78); Glaucoma suspect: 45.2 (SD 1.56); Normal:</td>
<td>M: 86; F: 89</td>
<td>USA</td>
<td>Jan 1978 – Apr 1979</td>
</tr>
<tr>
<td>Study id</td>
<td>Index test(s)</td>
<td>Test(s) carried out and interpreted by</td>
<td>Reference Standard</td>
<td>Enrolled (people)</td>
<td>Analysed</td>
<td>Mean age (range)</td>
<td>Gender</td>
<td>Country</td>
<td>Time period</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------------------------------------------------------------------------</td>
<td>----------------------------------------</td>
<td>--------------------</td>
<td>-------------------</td>
<td>----------</td>
<td>----------------------------------------</td>
<td>--------</td>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>Sommer 1979</td>
<td>Optic disc photography; RNFL photography</td>
<td>N/S</td>
<td>Follow-up confirmation</td>
<td>Unclear</td>
<td>223 eyes of ? people (both tests)</td>
<td>N/S</td>
<td>N/S</td>
<td>USA</td>
<td>N/S</td>
</tr>
<tr>
<td>Wollstein 2000</td>
<td>Optic disc photography</td>
<td>Photos taken by trained technicians; assessed by glaucoma consultants, glaucoma fellow, clinical glaucoma technician</td>
<td>Ophthalmic examination</td>
<td>123</td>
<td>123 eyes of 123 people</td>
<td>Glaucoma: 65.1 (SD 10.06); Normal: 57.1 (SD 12.52)</td>
<td>N/S</td>
<td>UK</td>
<td>N/S</td>
</tr>
<tr>
<td>Study id</td>
<td>Index test(s)</td>
<td>Test(s) carried out and interpreted by</td>
<td>Reference Standard</td>
<td>Enrolled (people)</td>
<td>Analysed</td>
<td>Mean age (range)</td>
<td>Gender</td>
<td>Country</td>
<td>Time period</td>
</tr>
<tr>
<td>---------------</td>
<td>-----------------------</td>
<td>------------------------------------------------</td>
<td>--------------------</td>
<td>-------------------</td>
<td>-------------------</td>
<td>------------------</td>
<td>--------</td>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>Wood 198761</td>
<td>Ophthalmoscopy</td>
<td>Ophthalmologists; junior doctors</td>
<td>Ophthalmic examination</td>
<td>22</td>
<td>43 eyes of 22 people</td>
<td>(32 to 75)</td>
<td>N/S</td>
<td>UK</td>
<td>N/S</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Patient source: not stated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes:
1. N/S, not stated.
2. Numbers analysed are people unless otherwise stated.
3. Study ids in brackets eg [Vernon 1991] are secondary reports that also contribute outcome data.
Table 2  HSROC analysis: all studies compared with higher quality studies.

<table>
<thead>
<tr>
<th></th>
<th>Optic disc photography</th>
<th>HRT II</th>
<th>FDT C-20-5</th>
<th>SAP threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sensitivity % (95% CrI)</td>
<td>Specificity % (95% CrI)</td>
<td>Sensitivity % (95% CrI)</td>
<td>Specificity % (95% CrI)</td>
</tr>
<tr>
<td>All studies</td>
<td>73 (61 to 83)</td>
<td>89 (50 to 99)</td>
<td>86 (55 to 97)</td>
<td>89 (66 to 98)</td>
</tr>
<tr>
<td>Higher quality</td>
<td>74 (30 to 95)</td>
<td>82 (45 to 97)</td>
<td>93 (58 to 99)</td>
<td>85 (47 to 97)</td>
</tr>
</tbody>
</table>

Notes:
1. Optic disc photography (all studies n = 6, higher quality studies n = 3); HRT II (all studies n = 3, higher quality studies n = 2); FDT C-20-5 (all studies n = 5, higher quality studies n = 2); SAP threshold (all studies n = 5, higher quality studies n = 2).
Table 3  Sensitivity, specificity, DOR and relative DOR at the common cutoff for studies directly comparing tests.

<table>
<thead>
<tr>
<th>Study id</th>
<th>Test</th>
<th>Common cutoff</th>
<th>Sens % (95% CI)</th>
<th>Spec % (95% CI)</th>
<th>DOR (95% CI)</th>
<th>RDOR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vitale 2000&lt;sup&gt;34&lt;/sup&gt;</td>
<td>SAP supra 3 adjacent points missed</td>
<td>50 (37 to 63)</td>
<td>83 (76 to 88)</td>
<td>5 (3 to 9)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Optic disc photo VCDR &gt; 0.6</td>
<td>77 (62 to 89)</td>
<td>59 (50 to 67)</td>
<td>5 (2 to 11)</td>
<td>0.99 (0.36 to 2.75)</td>
<td></td>
</tr>
<tr>
<td>Ieong 2003&lt;sup&gt;55&lt;/sup&gt;</td>
<td>SAP supra Optometrist judgement</td>
<td>72 (53 to 87)</td>
<td>95 (82 to 99)</td>
<td>46 (9 to 237)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HRT II Global/one of six segments abnormal</td>
<td>69 (49 to 85)</td>
<td>95 (82 to 99)</td>
<td>39 (8 to 198)</td>
<td>0.85 (0.08 to 8.54)</td>
<td></td>
</tr>
<tr>
<td>Robin 2005&lt;sup&gt;30&lt;/sup&gt;</td>
<td>SAP threshold AGIS score ≥ 3 [common cutoff]</td>
<td>63 (38 to 84)</td>
<td>74 (68 to 80)</td>
<td>5 (2 to 13)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HRT II ≥ 1 borderline or 1 severe abnormality</td>
<td>95 (74 to 100)</td>
<td>81 (75 to 85)</td>
<td>75 (10 to 574)</td>
<td>15.01 (1.57 to 143.82)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FDT C-20-5 One abnormal point</td>
<td>84 (60 to 97)</td>
<td>55 (49 to 61)</td>
<td>7 (2 to 23)</td>
<td>1.31 (0.27 to 6.43)</td>
<td></td>
</tr>
<tr>
<td>Spry 2005&lt;sup&gt;46&lt;/sup&gt;</td>
<td>SAP threshold GHT outside normal limit and/or p &lt; 0.05 with the PSD global index in one/ both eyes</td>
<td>80 (52 to 96)</td>
<td>52 (34 to 69)</td>
<td>4 (1 to 18)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FDT C-20 matrix</td>
<td>100 (78 to 100)</td>
<td>27 (13 to 46)</td>
<td>12 (1 to 222)</td>
<td>2.83 (0.11 to 72.91)</td>
<td></td>
</tr>
<tr>
<td>Ivers 2001&lt;sup&gt;23&lt;/sup&gt;</td>
<td>SAP supra 3 or more points missing</td>
<td>89 (80 to 94)</td>
<td>73 (71 to 74)</td>
<td>20 (10 to 39)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAT IOP &gt; 22 mmHg</td>
<td>14 (7 to 23)</td>
<td>98 (97 to 98)</td>
<td>6 (3 to 12)</td>
<td>0.31 (0.12 to 0.78)</td>
<td></td>
</tr>
<tr>
<td>Wang 1998&lt;sup&gt;36&lt;/sup&gt;</td>
<td>SAP supra Absolute or relative defects ≥ 17</td>
<td>70 (57 to 80)</td>
<td>67 (59 to 74)</td>
<td>5 (2 to 9)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAT IOP &gt; 21 mmHg</td>
<td>28 (17 to 40)</td>
<td>96 (93 to 98)</td>
<td>9 (4 to 19)</td>
<td>1.89 (0.70 to 5.13)</td>
<td></td>
</tr>
</tbody>
</table>

Notes:
1. RDOR = Relative DOR= index test DOR/ SAP DOR.
2. RDOR calculated as all direct studies had SAP as one of the tests. Values of RDOR > 1 indicate that the test performed better than SAP in the study and values < 1 indicate that the test performed worse than SAP.
3. AGIS, Advanced Glaucoma Intervention Study; GHT, Glaucoma Hemifield Test; PSD, Pattern Standard Deviation.
Table 4  Pair-wise indirect comparisons of tests in a single HSROC model.

<table>
<thead>
<tr>
<th></th>
<th>Ophthalmoscopy (60%, 94%) versus</th>
<th>Optic disc photography (73%, 89%) versus</th>
<th>RNFL photography (75%, 88%) versus</th>
<th>HRT II (86%, 89%) versus</th>
<th>OKP (86%, 90%) versus</th>
<th>SAP supra (71%, 85%) versus</th>
<th>SAP threshold (88%, 80%) versus</th>
<th>FDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optic disc photo</td>
<td>-12 (-46 to 20) 6 (-7 to 21)</td>
<td>-14 (-50 to 26) 6 (-7 to 30)</td>
<td>-24 (-57 to 14) 5 (-9 to 30)</td>
<td>-20 (-54 to 19) 4 (-9 to 26)</td>
<td>-10 (-43 to 20) 9 (-4 to 22)²</td>
<td>-26 (-58 to 2) 14 (-2 to 37)⁴</td>
<td>-26 (-58 to 2) 14 (-2 to 37)⁴</td>
<td></td>
</tr>
<tr>
<td>RNFL photo</td>
<td>-14 (-50 to 26) 6 (-7 to 30)</td>
<td>-2 (-31 to 34)</td>
<td>-12 (-38 to 22) -1 (-18 to 24)</td>
<td>-6 (-43 to 30) -1 (-26 to 22)</td>
<td>-4 (-29 to 38)</td>
<td>-16 (-37 to 5) 5 (-12 to 28) ²</td>
<td>-26 (-52 to -0) 3 (-11 to 18)</td>
<td></td>
</tr>
<tr>
<td>HRT II</td>
<td>-24 (-57 to 14) 5 (-9 to 30)</td>
<td>-10 (-45 to 25)</td>
<td>-1 (-25 to 24)</td>
<td>-4 (-39 to 16)</td>
<td>-16 (-37 to 5) 5 (-12 to 28) ²</td>
<td>-16 (-37 to 5) 5 (-12 to 28) ²</td>
<td>-30 (-62 to -0) 3 (-11 to 18)</td>
<td></td>
</tr>
<tr>
<td>OKP</td>
<td>-20 (-54 to 19) 4 (-9 to 26)</td>
<td>-8 (-35 to 27)</td>
<td>-6 (-43 to 30) -1 (-26 to 22)</td>
<td>-4 (-29 to 38)</td>
<td>-16 (-37 to 5) 5 (-12 to 28) ²</td>
<td>-16 (-37 to 5) 5 (-12 to 28) ²</td>
<td>-30 (-62 to -0) 3 (-11 to 18)</td>
<td></td>
</tr>
<tr>
<td>SAP supra</td>
<td>-10 (-43 to 20) 9 (-4 to 22)²</td>
<td>2 (-23 to 25)</td>
<td>4 (-31 to 29) 3 (-21 to 18)</td>
<td>10 (-24 to 34)</td>
<td>-16 (-37 to 5) 5 (-12 to 28) ²</td>
<td>-16 (-37 to 5) 5 (-12 to 28) ²</td>
<td>-30 (-62 to -0) 3 (-11 to 18)</td>
<td></td>
</tr>
<tr>
<td>SAP threshold</td>
<td>-26 (-58 to 2) 14 (-2 to 37)⁴</td>
<td>-14 (-38 to 7) ³ 8 (-11 to 31)</td>
<td>-12 (-46 to 12) 8 (-17 to 32)</td>
<td>-16 (-37 to 5) 5 (-12 to 28) ²</td>
<td>-16 (-37 to 5) 5 (-12 to 28) ²</td>
<td>-16 (-37 to 5) 5 (-12 to 28) ²</td>
<td>-30 (-62 to -0) 3 (-11 to 18)</td>
<td></td>
</tr>
<tr>
<td>FDT C-20-1</td>
<td>-30 (-62 to -0) 3 (-11 to 18)</td>
<td>-18 (-42 to 6) ³ -6 (-21 to 12)</td>
<td>-16 (-50 to 10) -5 (-29 to 13)</td>
<td>-10 (-42 to 14) -8 (-22 to 9)</td>
<td>-4 (-23 to 18)</td>
<td>-13 (-36 to 6) 5 (-12 to 28) ²</td>
<td>-30 (-62 to -0) 3 (-11 to 18)</td>
<td></td>
</tr>
<tr>
<td>FDT C-20-5</td>
<td>-11 (-49 to 32) 19 (-2 to 53)⁴</td>
<td>1 (-30 to 40) 12 (-10 to 47)</td>
<td>3 (-36 to 44) 12 (-16 to 47)</td>
<td>9 (-29 to 49) 14 (-13 to 49)</td>
<td>15 (-11 to 53)</td>
<td>19 (-10 to 57) 5 (-23 to 41) ²</td>
<td>19 (-10 to 57) 5 (-23 to 41) ²</td>
<td></td>
</tr>
<tr>
<td>GAT</td>
<td>15 (-22 to 47) -0 (-12 to 7)</td>
<td>27 (-4 to 53) ⁴ -6 (-21 to 3) ³</td>
<td>29 (-10 to 57) -6 (-30 to 4)</td>
<td>35 (-2 to 62) ³ -9 (-22 to 0) ³</td>
<td>41 (14 to 64) ³</td>
<td>45 (17 to 68) ³ -19 (-53 to -0) ³</td>
<td>45 (17 to 68) ³ -19 (-53 to -0) ³</td>
<td></td>
</tr>
</tbody>
</table>

Notes:
1. In the column headings the summary sensitivity and specificity values from the HSROC meta-analysis models are shown after the name of the test.
2. Test A (column) versus test B (row) = A – B. For each comparison, within each cell, the top row is the median difference in sensitivity (95% CrI) and the bottom row is the median difference in specificity (95% CrI).
3. Statistically significant difference at 5% significance level.
4. Statistically significant difference at 10% significance level.
5. Statistically significant difference at 20% significance level.
### Appendix 1  Summary of sensitivity, specificity and DOR for tests included in the HSROC meta-analysis models.

<table>
<thead>
<tr>
<th>Test</th>
<th>Number of studies</th>
<th>Common cutoff</th>
<th>Sensitivity % (95% CrI)</th>
<th>Specificity % (95% CrI)</th>
<th>DOR (95% CrI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ophthalmoscopy</td>
<td>5</td>
<td>VCDR ≥ 0.7</td>
<td>60 (34 to 82)</td>
<td>94 (76 to 99)</td>
<td>26 (6 to 110)</td>
</tr>
<tr>
<td>Optic disc photography</td>
<td>6</td>
<td>VCDR ≥ 0.6</td>
<td>73 (61 to 83)</td>
<td>89 (50 to 99)</td>
<td>22 (3 to 148)</td>
</tr>
<tr>
<td>RNFL photography</td>
<td>4</td>
<td>Diffuse and/or localised defect</td>
<td>75 (46 to 92)</td>
<td>88 (53 to 98)</td>
<td>23 (4 to 124)</td>
</tr>
<tr>
<td>HRT II</td>
<td>3</td>
<td>≥ 1 borderline or outside normal limits</td>
<td>86 (55 to 97)</td>
<td>89 (66 to 98)</td>
<td>51 (11 to 246)</td>
</tr>
<tr>
<td>FDT C-20-1</td>
<td>3</td>
<td>1 abnormal point</td>
<td>92 (65 to 99)</td>
<td>94 (73 to 99)</td>
<td>181 (25 to 2139)</td>
</tr>
<tr>
<td>OKP C-20-5</td>
<td>5</td>
<td>1 abnormal point</td>
<td>78 (19 to 99)</td>
<td>75 (57 to 87)</td>
<td>10 (0.7 to 249)</td>
</tr>
<tr>
<td>SAP suprathreshold</td>
<td>9</td>
<td>≥ 3 points missing</td>
<td>71 (51 to 86)</td>
<td>85 (73 to 93)</td>
<td>14 (6 to 34)</td>
</tr>
<tr>
<td>SAP threshold</td>
<td>5</td>
<td>AGIS score ≥ 3</td>
<td>88 (65 to 97)</td>
<td>80 (55 to 93)</td>
<td>30 (6 to 159)</td>
</tr>
<tr>
<td>GAT</td>
<td>9</td>
<td>IOP &gt; 21 mmHg</td>
<td>46 (22 to 71)</td>
<td>95 (89 to 97)</td>
<td>15 (4 to 49)</td>
</tr>
</tbody>
</table>

**Notes:**

1. The common cutoff was considered to also include the following cutoffs: Ophthalmoscopy (discs graded as normal or suspicious, subjective criteria); Optic disc photography (VCDR ≥ 0.7, normal/glaucomatous disc based on majority opinion of observers); RNFL photography (NFL lost); HRT II (global or 1 of the 6 segments flagged abnormal); OKP (1 or more points missing, if ≥ 1 chart numbers consistently made the black stimulus disappear); SAP suprathreshold (≥ 17 relative or absolute defects and/or cluster of 8 in any one quadrant, ≥ 4 abnormal points in any single quadrant, sufficient points to drop the indicator into the suspicious zone or below, 3 abnormal adjacent points, ≥ 1 missed point, optometrist judgement, at least 1 absolute defect associated with 1 relative defect or 3 adjacent relative defects or 4 non-adjacent relative defects or sure nasal step); SAP threshold (cross meridional, GHT abnormal/borderline, LDA 59 points, mirror image method, GHT outside normal limit and/or PSD p<0.05 in on or both eyes); GAT (IOP ≥ 21 mmHg, IOP 21-22 mmHG, IOP > 22 mmHG).

2. NFL, nerve fibre layer, AGIS, Advanced Glaucoma Intervention Study, GHT, Glaucoma Hemifield Test, LDA, Logistic Discriminant Analysis.