PERSPECTIVE

Appraising evaluations of screening/diagnostic tests: the importance of the study populations

Robert Harper, David Henson, Barnaby C Reeves

Sensitivity and specificity are the indices most commonly reported when describing the performance of a screening or diagnostic test. These indices, and their corresponding predictive values or likelihood ratios, are fundamental test properties since they allow the user to determine the consequences of selecting a particular cut off criterion for referral or further investigative tests. (Sensitivity is the proportion of diseased individuals correctly identified as diseased and specificity is the proportion of non-diseased individuals correctly identified as non-diseased. The positive predictive value (PPV) is the proportion of patients with positive screening test results who are found to have disease and the negative predictive value (NPV) is the proportion of patients with negative screening test results who are found not to have disease, based on the gold standard. Given a test result, a likelihood ratio describes how many times more likely a patient with disease is to have that test result, compared with a patient without the disease.)

A recent article[1] has highlighted the importance of complying with methodological standards[2–4] when evaluating diagnostic or screening tests, in order that the findings of a study can be applied with confidence to clinical practice. These standards need to be considered at both the study design stage and the reporting stage (see Table 1). Standards 1 and 2 are closely related, since they are both concerned with the way in which the sensitivity and specificity of a test may vary depending on the clinical and demographic characteristics of a population (for example, disease stage, age, sex). These standards allow clinicians wishing to use a test to judge whether the sensitivity/specificity reported by the evaluation can be applied to their own population of patients. Standard 3 also relates to a study's population, being concerned with the bias which can arise if only a proportion of the total number of subjects included in an evaluation are referred to receive "gold standard" verification. The extent of compliance with these standards in evaluations of ophthalmic tests has been reported elsewhere.[1]

The purpose of this perspective article is to discuss in more detail the importance of the study population when evaluating a screening test. The results from three studies, which have independently evaluated tonometry as a screening test, are discussed in order to illustrate how the selection of the population in a study can make a considerable difference to the findings. The aim is to highlight the need to consider carefully the choice of the populations in an evaluation of diagnostic/screening accuracy and the implications of the chosen populations on the applicability of the findings to clinical practice.

Evaluations of tonometry as a screening test

The performance of a screening test can be represented graphically by a receiver operator characteristic (ROC) curve, which illustrates the trade off in test sensitivity and specificity as the cut off criterion for classifying patients as diseased or non-diseased (that is, in this example, glaucomatous or non-glaucomatous) is systematically varied. For convenience, specificity is usually plotted on a reversed scale; curves that approach the top left hand corner of the ROC diagram indicate tests that achieve both high sensitivity and high specificity.

Figure 1 shows three different ROC curves for Goldmann tonometry, representing the sensitivity/specificity estimates derived from three independent studies.[5–7] All three curves illustrate that tonometry is not an ideal screening test, since no *single* IOP cut off criterion has both high sensitivity and high specificity for detecting glaucoma. However, it is striking that the screening accuracy of tonometry varies considerably between these studies. For example, although the sensitivity estimates for an IOP cut off criterion of >21 mm Hg (indicated by "20" in Fig 1) are comparable for all studies (sensitivity is ~50%), the specificity varies from 97% (95% CI=93–99%) in the study by Harper and Reeves[7] to 49% (95% CI=43–55%) in the study reported by Daubs and Crick.[5] This large discrepancy in specificity has a considerable effect on

Table 1 Methodological standards for the evaluation of diagnostic tests. Standards 1, 2, and 7 relate to the relevance of the results to particular settings, whereas standards 3, 4, and 6 are primarily about the validity of the results. Standard 5 relates to the need for careful reporting of the results*

(1) **Specification of spectrum composition** This standard requires at least three of the following four descriptors to be reported for the study population: the age and sex distribution, the presenting clinical symptoms and/or disease stage of the populations studied, and the eligibility criteria for the subjects included.
(2) **Analysis of pertinent subgroups** This standard requires the evaluation to cite the indices of accuracy for any pertinent demographic or clinical subgroup of the population.
(3) **Avoidance of work-up (verification) bias** This standard requires an evaluation to ensure all subjects receive both diagnostic testing and gold standard verification.
(4) **Avoidance of review bias** This standard requires an evaluation to make a clear statement about the independence in interpreting both the test and the gold standard procedure.
(5) **Presentation of precision of results for test accuracy** This standard requires an evaluation to report the 95% CI or standard error associated with the indices of diagnostic accuracy
(6) **Presentation of indeterminate test results** This standard requires an evaluation to state the number of indeterminate results and whether or not these results had been included or excluded when the indexes of accuracy were calculated.
(7) **Presentation of test reproducibility** This standard requires that the reproducibility of a test should be reported, or that the report should cite other sources of this information.

*Jaeschke *et al*[2 3] and Reid *et al*.[4]

Table 2 Sensitivity and specificity estimates for the criterion of an IOP ⩾21 mm Hg and the associated positive likelihood ratios (that is, sensitivity/1 − specificity) and PPV (%) for glaucoma, given an IOP ⩾21 mm Hg. The PPV estimates have assumed a prevalence of undetected glaucoma of 1% and no additional patient risk factors (that is, pretest probability of glaucoma of 1%)*

IOP criterion ⩾21 mm Hg	*Sensitivity (95% CIs)*	*Specificity (95% CIs)*	*Likelihood ratio (+ve)*	*Positive predictive value (%)*
Daubs and Crick[5]	52% (48–56)	49% (43–55)	1.0	1.0
Tielsch *et al*[6]	51% (44–58)	84% (83–85)	3.2	3.1
Harper and Reeves[7]	51% (39–63)	97% (94–100)	17.0	14.7

PPV = (prevalence × sensitivity)/((prevalence × sensitivity) + (1 − prevalence × 1 − specificity)).
95% CIs are not given for these PPV estimates, since, for the purposes of this comparison, we have used Bayes's theorem and the sensitivity/specificity at an IOP cut off criterion ⩾21 mm Hg (assuming the same glaucoma prevalence rate in each case), rather than working back from the frequency of study cases.

the positive likelihood ratio and the associated post-test probability of glaucoma (Table 2), with consequent implications for the management of individual patients. For example, at an IOP level of 21 mm Hg the data of Daubs and Crick[5] gives a likelihood ratio of 1, suggesting that the post-test probability of glaucoma is exactly the same as the pretest probability (that is, no information gain). In contrast, the data of Harper and Reeves[7] give a likelihood ratio of 17, representing a "large" change from pretest to post-test probability.[3]

A criterion free measure of the diagnostic/screening accuracy of a test is provided by the area under the ROC curve.[8] This statistic allows the overall performance of tonometry estimated by the three studies to be compared (see Table 3). We have calculated non-parametric areas, and tests of significance of differences between areas using the method of De Long *et al*.[9] The area under the ROC curve is significantly less for the study of Daubs and Crick[5] than the other two studies. Similarly, the area under the ROC curve is significantly less for the study of Tielsch *et al*[6] than the study of Harper and Reeves.[7] How might one account for these discrepancies?

Table 3 Area under the ROC curves (A_z) and the associated 95% confidence intervals

Study	*A_z*	*A_z 95% CI*
Daubs and Crick[5]	0.54	0.50–0.58
Tielsch *et al*[6]	0.73	0.69–0.77
Harper and Reeves[7]	0.87	0.81–0.93

Comment

There are differences in the methods used in the three studies that should be recognised, including differences in the gold standard (see Table 4). However, we believe that the most likely explanation for the observed discrepancies in test performance is the differences in the study populations and the ways in which patients were selected for inclusion.

Firstly, let us consider what might influence test specificity, since the major discrepancy in test effectiveness would appear to be due to differences in this index. Since specificity estimates are derived from the "normal" or "non-glaucomatous" populations, we must consider the possible differences in these populations in order to explain the discrepancies. Consider the cut off criterion of ⩾21 mm Hg (indicated by "20" in Fig 1). Although the sensitivity estimates are very comparable for this criterion, the specificity estimates vary from 97% to 49%, implying a considerable difference in the distribution of IOP in these populations. The data from Daubs and Crick[5] suggest that 51% of their non-glaucomatous population had an IOP >20 mm Hg in contrast with the 3% with an IOP >20 mm Hg suggested by the data from Harper and Reeves.[7] Figure 2 illustrates the distributions of IOP in the non-glaucomatous sample of each study, derived from the reported specificities at each level of IOP (see legend to Fig 2). The data of Daubs and Crick[5] have a highly atypical distribution of IOP in their non-glaucomatous population. There are far more non-glaucomatous patients with high IOP than found with the data from the other studies. This atypical distribution might have arisen if the sample comprised people who had been referred to the hospital eye service as glaucoma suspects but who were subsequently confirmed as non-glaucomatous by an ophthalmological examination. Such a population would include a much higher proportion of patients with ocular hypertension than one drawn from the general population, since ocular hypertension is a common reason for referrals from primary care. The effect of having a high proportion of ocular hypertensives in the non-glaucomatous population is to shift the entire ROC curve towards the diagonal "indecision" line. Evaluation of tonometry in this population is clearly of little relevance to the performance of tonometry as a screening test.

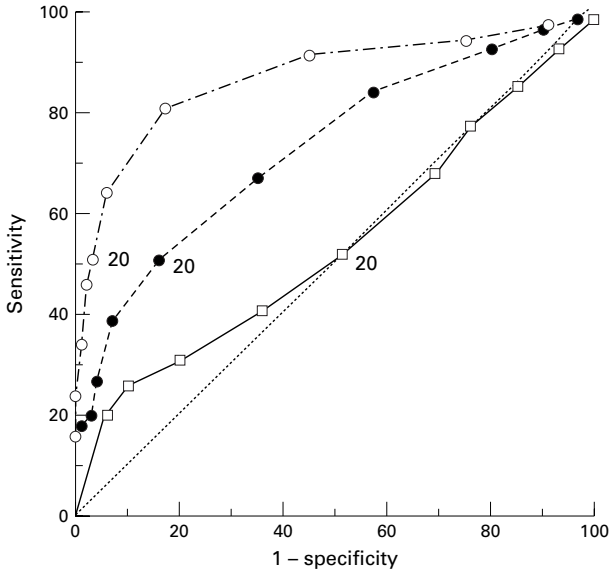

Figure 1 ROC curves for tonometry, drawn from the data of Daubs and Crick (open squares), Tielsch et al (solid circles), and Harper and Reeves (open circles). In each case the data points represent the sensitivity and specificity at different levels of IOP from 10 mm Hg to 28 mm Hg in 2 mm Hg steps.

The studies by Tielsch *et al*[6] and Harper and Reeves[7] both estimate the specificity of tonometry to be higher than the study by Daubs and Crick.[5] However, the specificities observed in these two studies also differ, although for most cut off criteria this difference is less extreme than the differences between each study and that of Daubs and Crick. The relatively small sample of non-glaucomatous individuals studied by Harper and Reeves[7] illustrates how the vagaries of sampling can cause apparent differences. None of the 145 non-glaucomatous subjects had an IOP in either eye >26 mm Hg (Fig 2), leading to estimates of specificity of 100% (one sided 97.5% CI=97–100%) for this IOP cut off criterion, compared with an estimate of specificity of 97% (95% CI=96–97%) in the study by Tielsch *et al*.[6] While the difference in specificity is small, it has a considerable impact on the post-test probability estimates of "being glaucomatous given an IOP >26 mm Hg", which are 6.3% based on the data of Tielsch *et al*[6] and

Table 4 Summary of study populations

Study	*Setting*	*Sample size*	*Diagnostic criteria for glaucoma**
Daubs and Crick[5]	Existing glaucoma patients and referrals to King's College Hospital, London	566 glaucoma patients 273 non-glaucoma	Glaucomatous visual field loss using modification of Friedmann VFA† ("the character of the loss was appropriate when taken in conjunction with the fundus and optic disc appearances")
Tielsch *et al*[6]	Population based survey in Baltimore, Maryland	5308 subjects screened (196 cases with glaucoma)	Glaucomatous optic nerve head damage based upon visual fields (Humphrey and Goldmann), optic disc and nerve fibre layer findings
Harper and Reeves[7]	Glaucoma cases identified from referrals to the HES. Normals recruited by systematic sampling from list of a general practice	67 glaucoma patients 145 non-glaucoma	Glaucomatous visual field loss on a 132 point static suprathreshold field test (Henson CFS 2000, field score >25 and "the distribution of missed stimuli consistent with glaucomatous defects")

*Patients included in King's College study all had "open angle glaucoma". The Baltimore evaluation reported on "any type" of glaucoma (although 161 of the 196 cases had POAG) and the Oxford Study included cases of POAG alone. All studies excluded IOP as a diagnostic criterion for glaucoma.
†Noted in separate publication.[10]

100% based on the data of Harper and Reeves,[7] assuming a prevalence of undetected glaucoma of 1% in both instances (see PPV equation, Table 2).

Although the sensitivity estimates derived from the three studies are comparable, let us consider what might influence this index in order to appraise the relevance of the reported estimates to clinical practice. Sensitivity is derived entirely from the glaucoma population used in each evaluation. The glaucoma populations (Table 3) included in the evaluations reported by Daubs and Crick[5] and Harper and Reeves[7] comprised patients referred to the hospital eye service. Although all subjects included in these evaluations appear to have received the gold standard visual field assessment (on which the diagnosis of glaucoma was based), this method of sampling is still likely to have resulted in verification bias. The bias arises here because raised IOP will almost certainly have been one of the major reasons for referral of suspects to the hospital eye service, thereby increasing the apparent prevalence of raised IOP among the glaucoma cases relative to those with normal IOPs. In contrast, Tielsch *et al*[6] studied a large population based sample, an approach which avoids verification bias. However, in this study, verification bias *is* introduced subsequently, since only patients who failed one or more of the screening tests under evaluation (including tonometry) were referred for the "definitive ophthalmologic examination" upon which the glaucoma cases were validated. Thus, there are likely to have been a small number of truly glaucomatous individuals who "passed" all screening tests and who were not detected because they were not referred for the gold standard—that is, who were presumptively classified as true negatives. These individuals would have been classified as true negatives when in fact they should have been classified as false negatives. This differential misclassification means that the sensitivity estimate reported may be biased upwards. A comparison of the *relative* performance of two or more tests against the same gold standard can be unbiased (despite verification bias), if those positive on any of the tests under evaluation are followed up by the gold standard.[11]

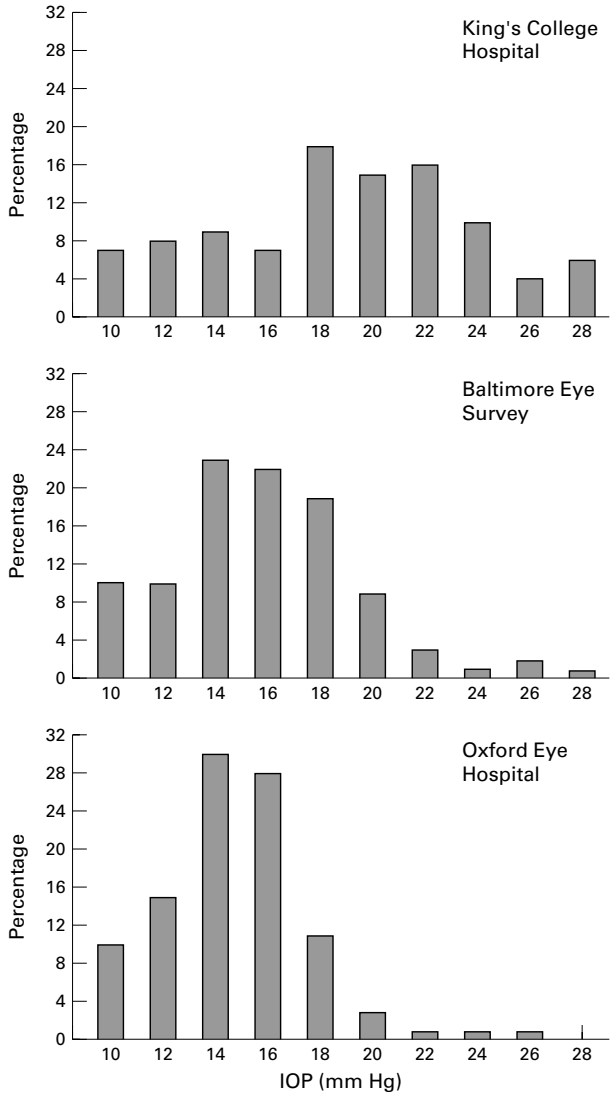

Figure 2 Distributions of IOP (mm Hg) for the "non-glaucomatous" samples used by Daubs and Crick (top), Tielch et al (centre), and Harper and Reeves (bottom). (Since the raw data were unavailable for two of these studies, distributions have been derived from the specificity estimates at each level of IOP, either as reported directly in the paper, or as read from figures illustrating the variation in sensitivity/specificity at specific cut off criteria. Percentages at 10 mm Hg and 28 mm Hg include all cases with IOP below and above these criteria respectively.)

Verification bias is difficult to avoid when evaluating a screening test where the prior probability of a disease is often low, as is the case with glaucoma—that is, the prevalence in people aged 40–89 years is 1.5% to 2%,[12] with about half of the cases being *undetected*. A practicable evaluation must either make assumptions about normality in those who pass the screening tests,[6] or alternatively select a population with a higher proportion of diseased people than would be generated through population screening (for example, established hospital eye service cases, new referrals, etc). As discussed above, selecting cases in this way will almost inevitably result in verification bias, since the reasons for referral will often be associated with the results of the screening test.[7] Alternative (yet still practicable) options for reducing verification bias in the evaluation of screening tests might be to (a) carry out the gold standard on a random sample of those who pass the screening tests (that is, are test negative), adjusting for the sampling fraction,[13] and/or (b) repeat the screening tests on those who initially pass screening in order to demonstrate the continuing absence of disease. While being inherent in many evaluations of screening tests, verification bias also arises in studies of diagnostic tests. Indeed, a recent commentary in the *BJO* has discussed sources of bias in evaluations of optic disc and retinal nerve fibre layer

instruments.[14] While verification bias may be difficult to avoid, it is important to re-emphasise the recommendation of Garway-Heath and Hitchings,[14]—that is, that authors should point out this limitation in the report of an evaluation. We believe that reviewers and journal editors should enforce this requirement.

A second factor which may have influenced the sensitivity estimates for tonometry in these evaluations is whether or not patients included in the glaucoma sample were receiving treatment or not. The glaucomatous population in the evaluation reported by Harper and Reeves[7] comprised newly referred suspects who did not have a previous history of glaucoma or glaucoma treatment. The King's College Hospital sample[5] is not described in sufficient detail to ascertain whether or not the glaucoma sample comprised exclusively pretreatment IOP readings. The proportion of treated cases in the sample of 196 "definite" or "probable" glaucoma patients (any type) is not reported by Tielsch *et al*,[6] but it would appear from associated publications[15 16] that 58 of the 132 cases of "definite" POAG were "aware of it and receiving treatment for it". The inclusion of a proportion of treated cases in the glaucoma sample is likely to lower the sensitivity and specificity estimates across the range of IOP cut off criteria, because one would expect these patients to have had a lower "screening" IOP than might have been recorded before treatment. It should be pointed out, however, that the Baltimore Eye Survey was a population based prevalence survey of glaucoma that had considerably wider epidemiological goals than the other evaluations and the inclusion of all cases of glaucoma was essential in order to derive estimates of glaucoma prevalence, etc. Nevertheless, it is, arguably, the sensitivity and specificity of tonometry for the detection of undiagnosed glaucoma that is most relevant to screening or case finding.

It is, of course, important to consider other aspects of the study population when appraising an evaluation of diagnostic accuracy. For example, although it is the *reporting* of the inclusion/exclusion criteria that forms part of the requirement for compliance with standard 1 (Table 1), it is very important for the reader to appraise whether the test has been evaluated on a population that is representative of the one to which they wish to apply the test. Selection is unlikely to present a problem in the context of tonometry, since the test was applied successfully to almost 99% of the unselected population reported by Tielsch *et al*.[6] However, evaluations of more subjective procedures (for example, visual field screening) or procedures influenced by restricted acuity or co-morbidity such as cataract (for example, imaging systems) sometimes use more selected populations. When applied to unselected populations, up to ~20% of patients are unable to complete some tests,[17] a factor of significance when one considers that co-morbidity, for example, is not uncommon in patients with age dependent conditions such as glaucoma.[18] While compliance with standard 6 (Table 1) merely requires a report to specify the numbers of indeterminate test results and whether or not these were included or excluded in the calculations for sensitivity/specificity, it is important for both the researcher and the reader of the evaluation to assess the generalisability of the findings. Where a prototype test is under evaluation, it may be appropriate to use a selected population, although authors should point out this limitation to readers. Any comment about the wider application of the test requires an evaluation on a more representative population.

In addition to the characteristics of the study populations discussed above, the criteria used to define diseased cases are also crucially important. The evaluations of tonometry discussed in this article all used a gold standard definition of glaucoma that was independent of IOP. In contrast, an evaluation of non-contact tonometry carried out by Vernon *et al*[19] *included* raised IOP >22 mm Hg in the gold standard definition of cases. This group reported a sensitivity and specificity >90% for an IOP cut off criterion of >22 mm Hg, a finding which differs considerably from the three evaluations described above (where sensitivities for this criterion are ~40%). The difficulty with this definition of cases is the very close association of the test under evaluation with the gold standard. A similar problem arises when visual field screening tests are evaluated against a gold standard with similar properties. In these circumstances, the evaluations are not necessarily invalid, but it is important to be aware of the "tautology bias",[1] especially when tests with properties which are similar to the gold standard are being compared with tests which are less well correlated with the gold standard.[7]

Conclusion

This perspective has highlighted how the characteristics of the population studied can influence the results of a screening test, by relating the considerable differences between the diagnostic/screening accuracy of tonometry reported in three independent evaluations to the populations studied. Despite the limited applicability of the King's College Hospital data[5] to a primary care setting, these sensitivity/specificity estimates and the associated PPV estimates have been used to make recommendations about the use of tonometry in optometric practice,[20] whereas the data from the Baltimore Eye Survey[6] probably provide the best estimates available at present. When appraising evidence about the accuracy of screening/diagnostic tests, readers of research findings should consider carefully the extent to which a study has complied with the relevant methodological standards and, in particular, the extent to which the population included in an evaluation is representative of the population in which they wish to use the test.

ROBERT HARPER
DAVID HENSON

Academic Department of Ophthalmology, Manchester Royal Eye Hospital, Manchester M13 9WH, UK

BARNABY C REEVES

Health Services Research Unit, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK

Correspondence to: Dr Harper
robert.harper@man.ac.uk

1 Harper, R, Reeves B. Compliance with methodological standards when evaluating ophthalmic diagnostic tests. *Invest Ophthalmol Vis Sci* 1999;**40**:1650–7.
2 Jaeschke A, Guyatt GH, Sackett DL, for the Evidence-based Medicine Working Group. Users' guides to the medical literature III. How to use an article about a diagnostic test A. Are the results of the study valid? *JAMA* 1994;**271**:389–91.
3 Jaeschke A, Guyatt GH, Sackett DL, for the Evidence-based Medicine Working Group. Users' guides to the medical literature III. How to use an article about a diagnostic test B. What are the results and will they help me in caring for patients? *JAMA* 1994;**271**:703–7.
4 Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research: getting better but still not good. *JAMA*. 1995;**274**: 645–51.
5 Daubs J, Crick RP. Epidemiological analysis of the King's College Hospital glaucoma data. *Res Clin Forums* 1980;**2**:41–59.
6 Tielsch JM, Katz J, Singh K, *et al.* A population-based evaluation of glaucoma screening: the Baltimore Eye Survey. *Am J Epidemiol* 1991;**134**: 1102–10.
7 Harper, R, Reeves B. Glaucoma screening: the importance of combining test data. *Optom Vis Sci* 1999;**75**:537–43.
8 Massof RW, Emmel TC. Criterion-free parameter-free distribution-independent index of diagnostic test performance. *Appl Opt* 1987;**26**:1395–408.
9 De Long ER, De Long DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating curves: a non-parametric approach. *Biometrics* 1988;**44**:837–45.
10 Crick RP. Prevention of blindness from glaucoma using the King's College Hospital computerized problem orientated medical record. *Br J Ophthalmol* 1975;**59**:236–48.
11 Chock C, Irwig L, Berry G, *et al.* Comparing dichotomous screening tests when individuals negative on both tests are not verified. *J Clin Epidemiol* 1997;**50**:1211–7.

12 Tuck MW, Crick RP. The age distribution of primary open angle glaucoma. *Ophthalmic Epidemiol* 1998;**5**:173–83.
13 Irwig L, Glasziou PP, Berry G, *et al.* Efficient study designs to assess the accuracy of screening tests. *Am J Epidemiol* 1994;**140**:759–69.
14 Garway-Heath DF, Hitchings RA. Sources of bias in studies of optic disc and retinal nerve fibre layer morphology. *Br J Ophthalmol* 1998;**82**:986.
15 Sommer A, Tielsch JM, Katz J, *et al.* Relationship between intraocular pressure and primary open angle glaucoma among white and black Americans. *Arch Ophthalmol* 1991;**109**:1090–5.
16 Tielsch JM, Sommer A, Katz J, *et al.* Racial variations in the prevalence of primary open angle glaucoma. *JAMA* 1991;**266**:369–74.
17 McLeod D, Joseph AJ, Henson DB, *et al.* Applicability of scanning laser polarimetry, retinal tomography and threshold perimetry to an unselected population of patients attending a glaucoma clinic. *Invest Ophthalmol Vis Sci*, 1999;**40**:S73.
18 Blomdahl S, Calissendorff B, Tengroth B, *et al.* Blindness in glaucoma patients. *Acta Ophthalmol* 1997;**75**:589–91.
19 Vernon SA, Jones SJ, Henry DJ. Maximising the sensitivity and specificity of non-contact tonometry in glaucoma screening. *Eye* 1991;**5**:491–3.
20 Directorate for Optometric Continuing Education and Training. *Primary open angle glaucoma and its detection.* London: College of Optometrists, 1998.

Video Report (www.bjophthalmol.com)
Capsule staining and mature cataracts: a comparison of indocyanine green and trypan blue dyes. *D F Chang*