Ophthalmic statistics note 5: diagnostic tests—sensitivity and specificity

Luke J Saunders,[1] Haogang Zhu,[1,2] Catey Bunce,[3] Caroline J Doré,[4] Nick Freemantle,[5] David P Crabb,[1] on behalf of the Ophthalmic Statistics Group

This fifth note from the Ophthalmic Statistics Group illustrates the utility of measurements of sensitivity and specificity in assessing the usefulness of a test for predicting the presence of pathology

[1]Department of Optometry and Visual Science, City University London, London, UK
[2]Institute of Ophthalmology, University College London, London, UK
[3]NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK
[4]UCL Comprehensive Clinical Trials Unit, University College London, London, UK
[5]Department of Primary Care and Population Health, University College London, London, UK

Correspondence to
Dr David P Crabb, Department of Optometry and Visual Science, School of Health Sciences, City University London, Northampton Square, London EC1V 0HB, UK; d.crabb@city.ac.uk

Received 26 August 2014
Revised 11 November 2014
Accepted 16 November 2014
Published Online First 8 December 2014

To cite: Saunders LJ, Zhu H, Bunce C, *et al. Br J Ophthalmol* 2015;**99**:1168–1170.

ABSTRACT

This is the fifth statistics note produced by the Ophthalmic Statistics Group (OSG) which is designed to be a simple guide to ophthalmic researchers on a statistical issue with an applied ophthalmic example. The OSG is a collaborative group of statisticians who have come together with a desire to raise the statistical standards of ophthalmic researcher by increasing statistical awareness of common issues.

A typical question facing an ophthalmologist in clinic is: 'How good is a particular test at diagnosing a particular pathology'? For example, how useful are the results from a Visual Field Analyser (VFA) in diagnosing glaucoma? Although a positive test result can indicate the presence of a symptom, such as visual field loss, full clinical diagnosis of a disease or condition is considerably more complicated. For instance, in glaucoma, a full clinical examination involving a battery of tests evaluating visual field loss, intraocular pressure, the optic nerve and functional tests such as standard automated perimetry are necessary to make a diagnosis.[1] Performing each test is costly and time consuming, and in some countries only a single test may be available for a particular condition. Where multiple tests are available, it is desirable, therefore, to compare their performance in a quantitative way through assessing their diagnostic utility and by examining the sensitivity and specificity with their respective CIs.

Suppose, for example, I have 200 individuals referred for assessment in my glaucoma clinic, 100 of the subjects referred are assessed using a VFA, while the other 100 are assessed using a new imaging technology. Individuals are classed as positive or negative according to whether their test result is positive or negative. The positive and negative refer to presence of an abnormality in visual field or optic disc, rather than the outlook for the patient, though it may seem perverse that a positive test result actually translates to something which might be viewed as negative news for the patient.

It may be tempting to assume from table 1 that the VFA is more useful in the diagnosis of glaucoma because it is identifying more subjects as positive. However, this may not reflect the full picture as a diagnostic test may not always give the correct classification. There are two types of error possible: the first type is when a healthy person is told that they have the pathology, a *false positive*; while the second type is when an individual with pathology is not identified by the test, a *false negative*. A false positive is essentially a false alarm, which could lead to subjects being incorrectly referred, causing them unnecessary anxiety and wasting clinical time and resources, or worse still, in other pathologies, potentially having to undergo unnecessary treatment if the correct diagnosis is not made in the clinic. A false negative can have equally dire consequences; if a patient has the pathology but is not diagnosed, then this would lead to them being falsely reassured that all is well and not receiving appropriate treatment at the earliest stage of their disease.

One approach to assessing the usefulness of a test is to perform it in a number of patients with known pathology, comparing the test results with this diagnosis. In this example, we consider the number of patients with known glaucoma who test positively using the VFA test (table 2) or the new imaging technology (table 3).

There are 25 subjects with known glaucoma and 75 subjects without the condition. The proportions of these two groups correctly diagnosed by the VFA were 24/25=0.96 and 57/75=0.76. These two proportions are given similar sounding names and are clearly described in a statistics note without application to ophthalmology.[2]

Sensitivity is the proportion of subjects *with* a diagnosis who are correctly identified by the test.

$$\begin{aligned} \mathit{Sensitivity} &= \frac{\mathit{True\ Positives}}{\mathit{All\ those\ with\ pathology}} \\ &= \frac{\mathit{True\ Positives}}{\mathit{True\ Positives} + \mathit{False\ Negatives}} \\ &= \frac{24}{24+1} = 0.96 \end{aligned}$$

Table 1 Test results on 200 subjects

Test	Test positive (+)	Test negative (−)	Total
VFA	42	58	100
Imaging technology	24	76	100

VFA, Visual Field Analyser.

Table 2 VFA test results for subjects with known glaucoma status

	Glaucoma (true status)		
Test—VFA	Present	Absent	Total
Test positive	24 True positive	18 False positive	42
Test negative	1 False negative	57 True negative	58
Total	25	75	100

VFA, Visual Field Analyser.

Specificity is the proportion of subjects *without* a diagnosis who are correctly identified by the test.

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{All those without pathology}}$$
$$= \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} = \frac{57}{57+18} = 0.76$$

We can say based on our test results that we expect 96% of patients with glaucoma to have abnormal VFA results (test positive), while 76% of subjects without glaucoma would have normal VFA results (test negative).

If we now consider the imaging technology, we have the following test results: of the 25 glaucoma patients, the imaging technique detects 15 as positive. The sensitivity is therefore 15/25=0.60 (or 60%). Of the 75 subjects without glaucoma, the imaging technique correctly identifies 66. The imaging technique, therefore, has a specificity of 66/75=0.88 (or 88%).

The VFA has higher sensitivity (0.96 vs 0.60) and a lower specificity (0.76 vs 0.88) than the imaging technique.

Sensitivity and specificity are proportions, so CIs can be calculated for them using standard methods.[3] It is important to ensure that the method adopted for calculating the CI is correct for the type of sample that you have. One indication that this is not the case would be when a CI exceeds 100%, which clearly it can never do.[4 5]

If this is done for this scenario we obtain the results shown in table 4.

When considering sensitivities and specificities it is also essential to consider the merits of the study used to provide these estimates. Initiatives such as the STARD statement (STAndards for the Reporting of Diagnostic accuracy studies)[6] and QUADAS[7] (QUality Assessment of Diagnostic Accuracy Studies) have been developed in order to aid experimenters and readers to assess the internal and external validity of diagnostic studies, and we would urge readers to look through these checklists to optimise what can be established from these studies. For further information on this topic, we would refer the interested reader to other relevant articles.[8 9]

Table 3 Imaging technology test results for subjects with known glaucoma status

	Glaucoma (true status)		
Test—Imaging technology	Present	Absent	Total
Test positive	15 True positive	9 False positive	24
Test negative	10 False negative	66 True negative	76
	25	75	100

Table 4 95% CIs for estimates of sensitivity% and specificity%

	VFA	Imaging technology
Sensitivity	96% (88.3% to 100%)	60% (40.8% to 79.2%)
Specificity	76% (66.3% to 85.7%)	88% (80.6% to 95.4%)

VFA, Visual Field Analyser.

CHOOSING A TEST

So, which is the best test here? On the one hand, using the more sensitive but less specific method (VFA in this fictitious example) will lead to more false positives and thus greater cost and time burdens to clinics and more patients with undue concern, while, on the other hand, using the more specific but less sensitive method (imaging technology) will lead to more false negatives, that is, more patients not being diagnosed, or being diagnosed late on in the disease if the test is repeated. Unsatisfactory as it may be, the answer is: it depends. Consideration needs to be given to the consequences of being a false positive or a false negative. If a false negative error is serious, as would be the case if diagnosing ocular tumours, a high sensitivity would take priority. If there is a need to avoid false positives, as might be the case in the glaucoma example and for other diseases with a low prevalence in the population, then a high specificity would be desirable.

A positive test that has high sensitivity is not a guarantee that the individual has the condition, particularly when specificity is not very high and the condition in question is rare. For instance, in our example scenario, which has a high prevalence of patients with glaucoma (25%), the sensitivity of VFA is high (96%), meaning that a high proportion of individuals with glaucoma would be correctly referred (only 4% would be missed). However, almost half of the positive tests overall are, in fact, false positives (18/42=42.8%), which represents an unacceptably large drain on resources in the context of glaucoma and treatment of many other conditions. It is, therefore, important to bear in mind that, in this context, a positive test requires further confirmation, while a negative test is more likely to be correct because, with a highly sensitive test, there are few false negative results, as seen in table 2. Indeed, if a test is 100% sensitive, there will be no false negatives, so a negative test will always be a true negative.

Obviously, it would be ideal for both sensitivity and specificity to be high, a perfect test having 100% sensitivity and 100% specificity, but, in practice, selecting the cut-off is always a compromise between sensitivity and specificity because there is generally a trade-off between the two; as one measure increases, the other decreases. The cut-off selected to define a positive test for a continuous testing method should be selected to optimise the balance between sensitivity and specificity. For example, using more stringent criteria for abnormal VFA results would decrease the proportion of individuals wrongly classified as having glaucoma, although it would also result in some patients who had correctly been classified with glaucoma now being misdiagnosed as without the condition. In other words, for a given cut-off, the sensitivity can be increased at the expense of specificity and vice versa. When comparing two tests, it is a good practice to select cut-offs that fix specificity at a particular level for both tests, which allows test sensitivities to be compared at equivalent specificities.

LESSONS LEARNED

- Sensitivity and specificity are useful summary measures for describing the diagnostic utility of a testing method. Test utility measures should be summarised using CIs.

- As a condition will usually have a low prevalence in the population, a positive test result is likely to require independent confirmation even when using a highly sensitive test, while a negative finding is more likely to be truly negative.
- The cut-off used to define a positive test can be selected to obtain an optimal compromise between sensitivity and specificity. A useful strategy for comparing diagnostic test performance is to compare sensitivities at equivalent specificities.

Collaborators Valentina Cipriani, Jonathan Cook, Phillippa Cumberland, Gabriela Czanner, Paul Donachie, Andrew Elders, Marta Garcia Finana, Rachel Nash, Neil O'Leary, Toby Prevost, Ana Qaurtilho, Chris Rogers, Selvaraj Sivasubramaniam, Simon Skene, Irene Stratton, Joana Vasconcelos, Wen Xing.

Contributors LJS, HZ and DPC designed and drafted the paper. LJS, HZ, CB, CJD, NF and DPC reviewed and revised the paper.

Funding CB is partly funded by the National Institute for Health Research (NIHR) Biomedical Research Centre (BRC) at Moorfields Eye Hospital NHS foundation Trust and UCL Institute of Ophthalmology.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 National Institute for Health and Clinical Excellence. *NICE Clinical Guidelines—Glaucoma: Diagnosis and management of chronic open angle glaucoma and ocular hypertension*. National Institute for Health and Clinical Excellence, 2009. http://publications.nice.org.uk/glaucoma-cg85 (accessed Jun 2014).

2 Altman DG, Bland JM. Diagnostic tests 1: Sensitivity and specificity. *Br Med J* 1994;308:1552.

3 Gardner MJ, Altman DG. Calculating confidence intervals for proportions and their differences. In: Gardner MJ, Altman DG. eds. *Statistics with confidence*. London: BMJ Publishing Group, 1989:28–33.

4 Deeks JJ, Altman DG. Sensitivity and specificity and their confidence intervals cannot exceed 100%. *Br Med J* 1999;318:193.

5 Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Stat Sci* 2001;16:101–33.

6 Bossuyt PM, Reitsma JB, Bruns DE, *et al*. Standards for Reporting of Diagnostic Accuracy. Toward complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. *Ann Intern Med* 2003;138:40–4.

7 Whiting PF, Rutjes AW, Westwood ME, *et al*. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155:529–36.

8 Mallett S, Halligan S, Thompson M, *et al*. Interpreting diagnostic accuracy studies for clinical care. *Br Med J* 2012;345:e3999.

9 Linnet K, Bossuyt PMM, Moons KGM, *et al*. Quantifying the accuracy of a diagnostic test or marker. *Clin Chem* 2012;58:1292–301.