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.

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.

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

\[
\text{Sensitivity} = \frac{\text{True Positives}}{\text{All those with pathology}} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} = \frac{24}{24 + 75} = 0.96
\]

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.

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

\[
\text{Sensitivity} = \frac{\text{True Positives}}{\text{All those with pathology}} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} = \frac{24}{24 + 1} = 0.96
\]

Table 1 Test results on 200 subjects

<table>
<thead>
<tr>
<th>Test</th>
<th>Test positive (+)</th>
<th>Test negative (−)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>VFA</td>
<td>42</td>
<td>58</td>
<td>100</td>
</tr>
<tr>
<td>Imaging technology</td>
<td>24</td>
<td>76</td>
<td>100</td>
</tr>
<tr>
<td>VFA, Visual Field Analyser</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

**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 unacceptable 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.


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.

Open Access This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/

REFERENCES