REVIEW ARTICLE

Requirements and Assessment of Laboratory Tests

Part 5 of a Series on Evaluation of Scientific Publications

Wilfried Bautsch

The following article does not deal directly with the evaluation of medical publications. It nevertheless appears as part of this series, as it describes a related problem—the statistical evaluation of a situation in which a clinical decision is necessary. The text deals with the theme of the positive predictive value, which is repeatedly found in scientific publications.

SUMMARY

Background: Current laboratory tests exhibit high sensitivity and specificity combined with comparatively low costs thus favoring broad and uncritical ordering habits.

Methods: Introduction of Bayes' theorem and discussion of its implications for laboratory test results in a mostly non-technical form, accompanied by a selective literature review.

Results and conclusions: According to Bayes' theorem the positive predictive value of laboratory test results is directly dependent on the prevalence of the disease in a given patient cohort. Thus, the clinical value of a given test result is critically dependent on a precise indication. Ordering of tests that are not indicated in a given patient is clinically useless and undesirable, where detailed information on disease prevalence is missing. These considerations are valid irrespective of ethical or economic considerations.

Dtsch Arztebl Int 2008; 105(24): 403–6
DOI: 10.3238/arztebl.2008.0403

Key words: laboratory diagnostics, blood analysis, diagnosis, PSA test, borreliosis

Institut für Mikrobiologie, Immunologie und Krankenhaushygiene, Städtisches Klinikum Braunschweig gGmbH: Prof. Dr. med. habil., Dr. rer. nat. Bautsch

In practice, laboratory tests are often ordered in a highly uncritical manner. They are comparatively cheap (for example, in comparison to imaging procedures), but highly sensitive and specific. This implies that if many different laboratory parameters are measured, this will supply clinically relevant information on the disease fast, with little effort and relatively cheaply. This is even taken to be the case if the tested parameters have little or nothing to do with the patient's symptoms. This includes routine profiles (which may be very extensive), as well as screening for diseases such as cancer which should be diagnosed before clinical symptoms develop and infectious diseases, such as borreliosis, which develop in phases.

This overlooks the fact that the reliability of test results depends on a clear indication. Although this aspect is frequently mentioned in public discussions of the value of screening (1), it is also important in daily medical practice. This does not of course apply to recommended screening tests (such as neonatal screening), as these issues are explicitly considered in the recommendations.

The present article sketches the underlying relationships in a largely non-mathematical form and explains the consequences for ordering diagnostic tests in daily medical practice. This problem is related to statistics, an area in which intuitive ideas are often misleading. The underlying problem is displayed in the following multiple choice question:

A laboratory test (for example, for borreliosis) has a diagnostic specificity of 98%. How probable is it that a patient who gives a positive test result does in fact have this disease?

a) You have to know the sensitivity too to be able to answer this question.
b) 98%
c) (1-specificity) × 100 (%) = 2%
d) None of these answers is correct.

Readers who can answer this question correctly can stop here. (The solution is at the end of the article). This article can be very helpful for practical medical work, as the underlying problem appears repeatedly in many different variations.

Confronted with this problem, most people attempt to solve it with the help of specificity alone. The specificity states the proportion of healthy subjects for whom a negative test result is (correctly) given. Conversely, 1-specificity gives the proportion of healthy subjects for

BOX 1

Sensitivity, specificity and positive predictive value

Sensitivity and specificity are statistical parameters which are well known for most tests, as they are easy to determine in principle. This is done by testing a defined number of samples from patients who are either known to be healthy or are known to be suffering from the disease in question. Each patient sample can give either a positive or a negative result. The results can be presented in a 2 × 2 matrix (4-field table). Instead of defined samples, a field study can be performed. A reference procedure must then be used retrospectively to establish whether the tested material comes from a healthy or an ill person. *Table 1* gives the result of such a field experiment in brackets.

The sensitivity gives the proportion of ill persons positively recognized by the test. It can be seen immediately from the 4-field table that the sensitivity must be a/(a+c), where (a+c) is the number of ill persons in the test cohort, of whom (a) patients give a positive test result. Thus the sensitivity of the test in this example is 398/(398+22) = 0.9476, or roughly 94.8%.

TABLE 1

		"True Value"	
		Ill	Healthy
Test result	positive	a (398)	b (12)
	negative	c (22)	d (1012)

The specificity gives the proportion of healthy persons in this test giving a negative test result, corresponding to d/(b+d), where (b+d) is the total number of healthy test persons in the test cohort, of whom (d) persons give a negative test result. Thus the specificity of the test in this example is 1012/(1012+12) = 0.9883, or roughly 98.8%.

How probable is it that a person with a positive test result is in fact ill? This is what mainly interests the responsible physician in a clinical situation. This probability is called the positive predictive value (PPV). It can be seen in the 2x2 matrix that the number of persons testing positive is (a+b) = 410. Of these, 398 (a) are in fact ill. The probability that a person with a positive test result is also ill in our example (the PPV) is then a/(a+b) = 398/410 = 0.9707 or about 97.1%.

This is somewhat different from the specificity. When determining the specificity, the right column (healthy) must be evaluated, but when determining the PPV, the upper line (positive test result) must be used. In the first case lines are evaluated, in the second columns.

Now the difference in this example is not very great (specificity 98.8%, in comparison with PPV 97.1%). The reason for this is that the proportion of ill persons (prevalence of the disease) was very high in this example, corresponding to (a+c)/(a+b+c+d)—about 29.1%. The number of healthy persons in the field experiment could be ten times higher, or 96.1%, corresponding to the prevalence of 3.9% *(Table 2)*. This is often a realistic assumption. This gives rise to the following values:

TABLE 2

		"True Value"	
		Ill	Healthy
Test result	positive	a (398)	b (120)
	negative	c (22)	d (10 120)

The sensitivity is then still 94.8% and the specificity is also unchanged: d/(b+d) = 10 120/(10 120 + 120) = 0.9883 or about 98.8%. However, this has a major influence on the PPV, as this is now only a/(a+b) = 398/(398+120) = 0.768 or 76.8%.

If we set the prevalence even lower, for example, to 0.41%—corresponding to a further 10-fold increase in the number of healthy test persons—the PPV would drop to 24.9%. In this case, a positive test result means that there is a probability of more than 75% than the test person does not have the disease.

Thus, a positive result can come from either an ill or a healthy person. In the latter case, it is a false positive or non-specific result. When the prevalence of the disease is lower, there are fewer ill persons in the test cohort, more healthy persons are tested and the probability increases that a positive test result is false.

In summary, this means that the positive predictive value (PPV) not only depends on the sensitivity and specificity, but also on the prevalence of the disease in the test cohort. The lower the prevalence, the lower the PPV is.

Analogous arguments apply to the negative predictive value (NPV), the probability that a person with a negative test result is indeed not ill. It can be seen from *Table 1* that the NPV = d/(c+d) = 1012/(1012+22) = 97.7%. In contrast to the PPV, the NPV decreases with increasing prevalence. If the number of ill persons in the test cohort is increased by a factor of 100, corresponding to the prevalence of 97.6%, the following values can be calculated *(Table 3)*: The sensitivity, a/(a+c) = 39 800/(39 800+2200) = 94.8%, and specificity, d/(d+c) = 1012/(1012+12) = 98.8% are unchanged. However, the NPV is now only d/(c+d) = 1012/(2200+1012) = 31.5%, i.e. only 31.5% of patients with a negative test result are in fact healthy. Thus, a negative test result can come from either healthy or ill persons. In the latter case, it is a false negative result. With increasing prevalence, more and more ill persons are tested. Corresponding to this, the probability that a negative test result is a false negative also increases. These mathematical relationships can also be presented as calculations of probability *(Box 2)*.

TABLE 3

		"True Value"	
		Ill	Healthy
Test result	positive	a (39 800)	b (12)
	negative	c (2200)	d (1012)

BOX 2

Bayes' theorem

Sensitivity, specificity, and positive predictive value can also be expressed as conditional probabilities. Let p(B/A) be the probability that the event B occurs under the condition A. For example, the conditional probability p (test result positive/test person ill) is exactly the same as sensitivity as defined in *Box 1* and is given by a/(a+c). However, the inverse probability p (test person ill/test result positive), the positive predictive value, is usually of more interest. In some sense, effect and cause are swapped. The probability is usually known that the cause (the disease) leads to a positive test result. We are often interested in another aspect, namely how a positive test result can lead to the conclusion of the cause (disease). The correct mathematical relationship is given by Bayes' theorem, as originally presented by Thomas Bayes: $p(A/B) = p(B/A) \times p(A)/p(B)$, or the transformation of this, as p(B) is not known directly: $p(A/B) = p(B/A) \times p(A) : [p(B/Ac) \times p(Ac) + p(B/A) \times p(A)]$, where Ac is the complement of A, i.e., A is not present.

whom a positive result is wrongly given (false positive rate). The intuitive tendency is to think that we now have all the necessary information and that the probability is 98%. However, this is wrong. The correct solution of the problem requires two additional pieces of information, the test sensitivity and the prevalence of the disease in the test cohort. The latter is the proportion of persons with the illness relative to all persons for whom the doctor has ordered this test. The reason for this is explained in *Box 1*.

Use additional parameters

What are the consequences for the example of borreliosis testing discussed in the introduction? The prevalence of active borreliosis in the population is not precisely known. Estimates range from 10 to 237 cases per 100 000 inhabitants (2), with major differences between the regions (3). The Robert Koch Institute published a value of 25 per 100 000 for Germany in 2003 (4). This will be used in the following, to simplify the calculations. Modern serological immune tests for borreliosis coupled to the recommended immunoblot are assumed to be at least 98% specific (5), although this figure is not known exactly and probably depends on the test system. We will assume that the specificity is 98%. This means that 25 genuine positive results are actually obtained for 100 000 tests in the population. It will be neglected that the sensitivity of the test is less than 100%. However, there are two additional fundamental problems in the interpretation of serological test results for borreliosis, which also exist when tests are only ordered for strict indications.

- A negative test result does not reliably exclude active borreliosis—particularly in the early stages—as the tests are less than 100% sensitive.
- The available serological tests cannot reliably distinguish between active borreliosis and a titer after recovery from borreliosis, so that even unambiguously positive serological findings per se are not an indication for treatment.

Aside from the 25 genuine positive results, there will also be 2000 false positive test results, as 1-specificity = 2%. There will therefore be a total of 2025 positive test results, of which 25 are caused by active borreliosis. This corresponds to a probability of about 1.25% that a test person with a positive test result really is suffering from active borreliosis. It follows that this test is clearly unsuitable for population screening, as it is almost 99% certain that a positive result is wrong.

The physician can influence the prevalence of a disease, meaning the prevalence of a disease in the test cohort for whom he orders the test. Thus, if he orders borreliosis testing for every patient—whatever the symptoms—, the reliability of the individual results is close to that for population screening, as everyone goes to the doctor at one time or another. The reliability of the positive result is then close to zero.

The situation is quite different if the test is ordered for a specific indication, for example, if the patient comes with acute peripheral facial palsy. The prevalence of borreliosis in patients with acute facial palsy has not been very well studied. A recent Norwegian article gives the value of about 10% (6); the value for children is certainly greater. The results are quite different for this patient cohort. Of 1000 tests, 18 will be false positive (1-specificity = 2% of 900 negative patients), but there will be 100 genuine positive findings. The probability that a patient with a positive test result genuinely has borreliosis is then (100/100 + 18) x 100 ~ 85%. This figure will certainly be greater for children.

Conclusion

Sensitivity and specificity are test-specific properties which the physician cannot actively influence. This assumes that the test is properly performed and evaluated, including the steps before and after the analysis. On the other hand, the reliability of a positive test result—the positive predictive value—is critically dependent on the prevalence of the disease in the test cohort and this is something the physician can influence. As a matter of principle, tests should only be ordered when they are indicated, as it is only then that the test result can be clinically evaluated. Results from non-indicated orders are clinically useless without a well founded database on the prevalence of the disease and should therefore not be ordered. This is unrelated to economic or ethical considerations.

Although the borreliosis test was used as an example, this applies to all laboratory tests. The arguments apply equally well to laboratory tests or to other investigations,

including X-ray, endoscopic, sonographic, electrocardiographic or clinical procedures. If the test or investigation is not indicated, this reduces its positive predictive value and increases the number of false positive test results.

The correct answer to the initial multiple choice question was—d.

Conflict of interest statement
The authors declare that no conflict of interest exists according to the guidelines of the International Committee of Medical Journal Editors.

Manuscript received on 6 February 2007, revised version accepted on 19 October 2007.

Translated from the original German by Rodney A. Yeates, M.A., Ph.D.

REFERENCES

1. Bögermann C, Rübben H: Früherkennung des Prostatakarzinoms. Dtsch Arztebl 2007; 104(8): A 503-04.
2. O´Connell S, Granström M, Gray JS, Stanek G: Epidemiology of European Lyme borreliosis. Zentralbl Bakteriol 1998; 287: 229–40.
3. Talaska T: Borreliose-Epidemiologie. Brandenburgisches Ärzteblatt 2002; 11: 338–40. www.laekb.de/15/15Beitraege/95021TH0211.pdf
4. Mehnert WH, Krause G: Surveillance of lyme borreliosis in Germany, 2002 and 2003. Euro Surveill 2005; 10: 83–5.
5. Goettner G, Schulte-Spechtel U, Hillermann R, Liegl G, Wilske B, Fingerle V: Improvement of Lyme borreliosis serodiagnosis by a newly developed recombinant immunoglobulin G (IgG) and IgM line immunoblot assay and addition of VlsE and DbpA homologues. J Microbiol 2005; 43: 3602–9.
6. Ljostad U, Okstad S, Topstadt T, Mygland A, Monstad P: Acute peripheral facial palsy in adults. J Neurol 2005; 252: 672–6.

Corresponding author
Prof. Dr. med. habil., Dr. rer. nat. Wilfried Bautsch
Institut für Mikrobiologie,
Immunologie und Krankenhaushygiene
Städtisches Klinikum Braunschweig gGmbH
Celler Str. 38
38814 Braunschweig, Germany
w.bautsch@klinikum-braunschweig.de