Diagnostic skills

Evidence based diagnosis: does the language reflect the theory?

Matt T Bianchi, Brian M Alexander

Much effort is directed towards optimising doctor-patient communication and avoiding misunderstandings. The language of everyday diagnostic reasoning as it routinely occurs among doctors in teaching hospitals could benefit from similar attention

Partners Neurology, Massachusetts General Hospital and Brigham and Women's Hospital, Boston, MA 02114
Matt T Bianchi
resident

Partners Radiation-Oncology, Massachusetts General Hospital and Brigham and Women's Hospital
Brian M Alexander
resident

Correspondence to: M Bianchi
mtbianchi@partners.org

BMJ 2006;333:442–5

Although interest in evidence based medicine has increased in recent years, and it is taught in most medical schools, evidence based strategies have been adopted inconsistently into routine care.[1 2] One aspect of evidence based medicine involves understanding the limitations of inherently imperfect diagnostic tests. Many trainees appreciate the concepts of sensitivity and specificity and learn how to combine the "art" of the history and physical exam (pre-test probability of disease) with the "science" of diagnostic testing (post-test probability of disease) without explicit use of quantitative probability theory. Nevertheless, it seems that quantitative reasoning is neither intuitive nor well understood. As diagnostic testing is a common and critical component of evaluating patients, it is worth considering whether the manner in which we verbally communicate these ideas may represent a fundamental (yet reparable) hindrance to diagnostic reasoning. We discuss common examples of diagnostic language that do not accurately reflect the underlying theory, and review the evidence for inadequate clinical application of bayesian strategies.

Innocent generalisations?

As trainees, we can all recall hearing pearls of wisdom conveyed in the form of: "Any patient presenting with this sign/symptom is assumed to have disease *X* until proved otherwise." The common mnemonic "SPin/SNout" is used to indicate that positive results from specific tests rule in disease, while negative results from sensitive tests rule out disease. One may hear sensitivity or specificity discussed in isolation ("that test is so sensitive that a negative result rules out disease") or, more commonly, of a test having good positive or negative predictive value. Certain findings are called "non-specific" because they manifest in multiple diseases. Although this language seems to capture simple diagnostic generalisations, does it actually reflect the bayesian logic that underlies diagnostic reasoning? The accuracy of such language is easily overlooked because in common practice test results agree with clinical suspicion and the details of sensitivity, specificity, and predictive value become arguably less important.

The basics of bayesian logic

To interpret any diagnostic test, one must have information not only about the test's characteristics but also about the patient (or a population with similar characteristics). Few tests are inherently accurate enough to "rule in" or "rule out" disease effectively in all cases. We should look at results as altering disease probability. This requires estimation of a pre-test probability that will be adjusted up or down by the test results. This is bayesian logic, which uses an adjustment factor called the likelihood ratio (LR) to convert a pre-test probability into a post-test probability (fig 1).[3 4] The upward adjustment of the probability after a positive result is called the LR(+) and is a number >1, while the downward adjustment after a negative result is the LR(−) and is a fraction <1. The key feature of the likelihood ratio is that it incorporates both the sensitivity and the specificity. Ruling disease in or out (or considering subsequent decisions on management) depends on a comparison of post-test probability with thresholds for further action based on factors such as severity of disease, risks of further testing, or side effects of treatment.[5]

English theologian and mathematician Thomas Bayes, 1702-61

Simply remembering that the likelihood ratio incorporates both sensitivity and specificity protects against the common misconception that sensitivity and specificity can be considered in isolation.[6] Although it is true in general that sensitivity impacts LR(−) more than specificity (and specificity impacts LR(+) more than sensitivity), the likelihood ratio is derived from both measurements. In fact, for every sensitivity (or specificity) less than 100%, there is a specificity (or sensitivity) that renders the $LR = 1$ (that is, no change in probability of disease). The fact that most tests are imperfect and therefore do nothing more than adjust

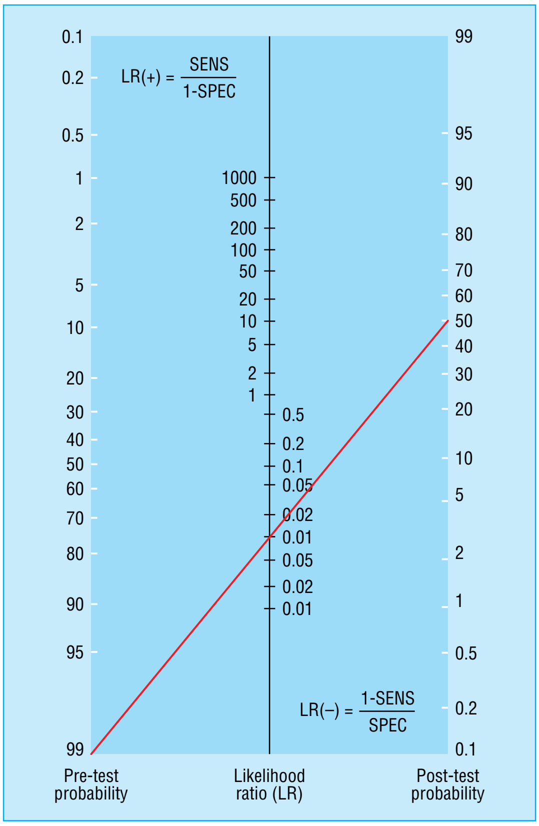

Fig 1 Nomogram (adapted from www.CEBM.net with permission) to convert pre-test probability to post-test probability using the likelihood ratio. The line refers to a text example

probability (which may or may not “rule in” or “rule out” the disease depending on the situation) protects against the misconception that a result can be interpreted without considering pre-test probability. Several studies have shown deficiencies in using pre-test probability when interpreting test results.[7–16]

“Affected until proved otherwise . . .”

This language is commonly used to emphasise that certain symptoms can represent the first presentation of a serious disease. For example, a positive result on faecal occult blood testing in an adult could indicate “colon cancer until proved otherwise.” This seemingly innocent statement translates into bayesian language: “colon cancer has a pre-test probability of >99%, and further investigation is needed to reduce its probability to <1%” (an arbitrary certainty of not having cancer). Although few clinicians use this strict interpretation of high pre-test probability, the bayesian consequences warrant discussion. Consider a test with exceptional sensitivity and specificity: 99% each. Colonoscopy may approach such numbers for detection of neoplasm, yielding an LR(−) of ~0.01. A negative result on colonoscopy would reduce the chances from 99% to ~50% (fig 1)—hardly ruling it out. Yet most physicians would stop investigating stools positive for blood after a negative result on colonoscopy. The gap between the language and the practice is that the actual pre-test probability of colon cancer in the example is far less than 100%, so the negative colonoscopy is informative. The intended message of “affected until proved otherwise” is actually that the threshold for further evaluation is low, not that the pre-test probability is high.

It is worth considering more realistic numbers. A negative result from what might be called a “good” test, with 90% sensitivity and 90% specificity, would reduce the disease probability only slightly, from 99% to ~90%. For a single test with such characteristics (LR(−) of ~0.1) to render disease probability <1%, the nomogram shows that pre-test probability would have to be no greater than 10% (fig 1). Negative results from two independent tests with exceptional sensitivity and specificity (99% each) would be needed to reduce disease probability from 99% to 1%, or four consecutive negative results from independent tests with sensitivity and specificity of 90% each. Test independence means that the result from one test cannot bias the outcome of the next, such that the post-test probability after one test becomes the pre-test probability of the subsequent test.

“This test has good predictive value . . .”

The language of predictive value is more problematic, yet understanding predictive value is critical for moving beyond the simplicity of sensitivity and specificity for interpretation of test results. Referring generally to the “predictive value of a test” gives the false impression that a test’s predictive power stands alone (in the same way, theoretically, as its sensitivity or specificity) and therefore can be applied to any patient. In fact, the predictive value is a reflection of the pre-test probability as well as the discriminative power (sensitivity and specificity) of the test (fig 2). Therefore, the predictive value is a characteristic of a test result in a specific patient (or representative population) not of the test result in general, nor of the test itself. It is inappropriate, for example, to describe a negative d-dimer test result as having good negative predictive value for pulmonary embolism. Doing so ignores the impact of pre-test probability—that is, it ignores the information provided by clinical judgment. If the pre-test probability of pulmonary embolism were high, then the negative d-dimer result would not rule out pulmonary embolism, and thus the d-dimer test is most useful in the setting of lower pre-test probability.[17]

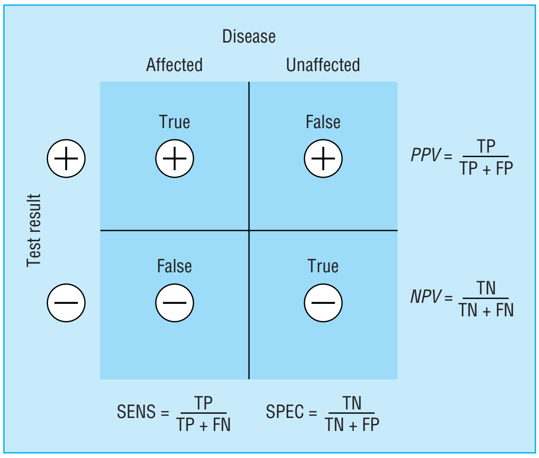
Disease
Affected
Unaffected
Test result
+
−
True +
False +
False −
True −
PPV = $\frac{TP}{TP + FP}$
NPV = $\frac{TN}{TN + FN}$
SENS = $\frac{TP}{TP + FN}$
SPEC = $\frac{TN}{TN + FP}$

Fig 2 Standard 2×2 box illustrating the determination of sensitivity, specificity, and predictive value for a dichotomous test

One of the potentially confusing aspects of predictive value is that it seems to be determined by simple calculations with the 2×2 box, similar to sensitivity and specificity, and therefore it may be misconstrued as a characteristic of the test itself. Whereas the calculations of sensitivity and specificity are unaffected by the proportion of affected versus unaffected patients ("vertical" calculations in fig 2), this is not the case for predictive value ("horizontal" calculations), which depends heavily on the disease prevalence. Thus, test results cannot be said to have predictive value; only a test result in a given patient (or population) has predictive value. Rather than a mere semantic distinction, this fundamental issue in test interpretation has been reported to be deficient at all levels of training.[7-16]

Specificity refers to the control population from which it was derived

The concept of specificity itself presents hidden challenges. One may refer to a test as being either "specific for a disease," to indicate that few other diseases could produce a positive test result, or as "non-specific," to indicate that it may yield positive results in multiple diseases (or in health). Specificity, like sensitivity, is often considered an intrinsic property of a test and therefore independent of the population under study. As specificity is determined by unaffected individuals who have positive results (fig 2), however, it is in fact dependent on the characteristics of this comparison population.

Consider the finding of fever: it is called "non-specific" for obvious reasons, but if a study of pharyngitis investigated a population of 10 year old children with sore throat it is unlikely that the unaffected control children would have fever. Therefore, fever might be considered a highly specific finding in such a study. A more practical challenge involves the mechanism of a false positive: stochastic assay variation (no biological meaning) versus a "real" false positive, arising from a different disease present in some members of the control population.

Consider next the finding of oligoclonal bands in the cerebrospinal fluid of a patient suspected of having multiple sclerosis. While several texts and reviews report 92-98% specificity (comparing patients with multiple sclerosis with "normal" controls), is that value relevant if the clinician is also considering alternative diagnoses such as lupus or Sjogren's, which can also manifest with oligoclonal bands in cerebrospinal fluid?[18 19] In this situation, it cannot be said that oligoclonal bands are "specific for multiple sclerosis," regardless of the reported specificity as previous control populations might not have contained patients with lupus or Sjogren's. Conversely, it has been suggested that the 14-3-3 protein assay in cerebrospinal fluid is not specific for Creutzfeldt-Jakob disease because the protein can also occur with other diseases, including central nervous system malignancy, infection, or stroke.[20] If imaging and evaluation of the cerebrospinal fluid can reduce the likelihood of such confounding conditions, however, a positive 14-3-3 in that setting might then be considered more "specific." Interpretation of specificity requires careful attention not only to the control population but also to the test's performance in other diseases that are being considered. Specificity should not be considered an intrinsic property of a test because it depends in part on the characteristics (even subclinical) of the control population from which it was derived. It is therefore critical to evaluate the study design from which the specificity of a test has been determined and to consider whether the test can be used more appropriately to distinguish one disease from another or to distinguish the presence or absence of disease.

Summary points

Most tests are imperfect and thus can only adjust disease probability, which requires estimation of the pre-test probability of disease

Likelihood ratios adjust disease probability by using both sensitivity and specificity

Clinical sayings of the type "affected until proved otherwise" indicate that the threshold for further evaluation is low, not that the pre-test probability is high

Predictive value is a characteristic of a test result in a specific patient, not of the test result in general, nor of the test itself

Specificity is not an intrinsic property of a test because it depends in part on the characteristics (even subclinical) of the control population

Conclusion

As with any non-intuitive skill, understanding statistical reasoning depends on the frequency of use in practice. Despite general awareness of the other concepts of evidence based medicine, the estimation pre-test probability and adjustment of disease probability in the setting of thresholds for testing and treating is not commonplace. Incomplete epidemiological information that facilitates estimation of pre-test probability certainly contributes to the challenge. Are easily "digestible" pearls of wisdom compromising the importance of pre-test probability and the concepts of bayesian logic? Can we afford to dismiss these concerns as mere semantics at a stage in training when bayesian concepts are not well understood? Perhaps these details go unnoticed or uncontested because most of the time test results agree with our expectations and the details of probability theory become less relevant. However, one could argue that the art of medicine is most reflected in the approach to the unexpected finding, a situation where generalisations carry more risk, and where knowledge of pre-test probability and bayesian logic is indispensable.

Contributors and sources: The authors are both residents in their third postgraduate year, with educational backgrounds in ion channel biophysics (MTB) and mathematics (BMA). The commentary was motivated by their personal challenges involving the complexities of interpreting diagnostic tests and the contrast between these complexities and the seemingly superficial manner in which test interpretation is often discussed clinically. The medical literature was searched with keywords

such as sensitivity, specificity, review, predictive value, medical decision making, pre-test probability; relevant papers were also isolated by "similar article" searches via www.PubMed.com. Both authors contributed equally to this work.

Funding: None.
Competing interests: None declared.

1 Green ML. Evidence-based medicine training in graduate medical education: past, present and future. *J Eval Clin Pract* 2000;6:121-38.
2 Ghosh AK, Ghosh K, Erwin PJ. Do medical students and physicians understand probability? *QJM* 2004;97:53-5.
3 Gallagher EJ. Clinical utility of likelihood ratios. *Ann Emerg Med* 1998;31:391-7.
4 Halkin A, Reichman J, Schwaber M, Paltiel O, Brezis M. Likelihood ratios: getting diagnostic testing into perspective. *QJM* 1998;91:247-58.
5 Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med* 1980;302:1109-17.
6 Boyko EJ. Ruling out or ruling in disease with the most sensitive or specific diagnostic test: short cut or wrong turn? *Med Decis Making* 1994;14:175-9.
7 Dolan JG, Bordley DR, Mushlin AI. An evaluation of clinicians' subjective prior probability estimates. *Med Decis Making* 1986;6:216-23.
8 Kassirer JP, Kopelman RI. Cognitive errors in diagnosis: instantiation, classification, and consequences. *Am J Med* 1989;86:433-41.
9 Lyman GH, Balducci L. Overestimation of test effects in clinical judgment. *J Cancer Educ* 1993;8:297-307.
10 Lyman GH, Balducci L. The effect of changing disease risk on clinical reasoning. *J Gen Intern Med* 1994;9:488-95.
11 Richardson WS, Wilson MC, Guyatt GH, Cook DJ, Nishikawa J. Users' guides to the medical literature: XV. How to use an article about disease probability for differential diagnosis. Evidence-Based Medicine Working Group. *JAMA* 1999;281:1214-9.
12 Noguchi Y, Matsui K, Imura H, Kiyota M, Fukui T. Quantitative evaluation of the diagnostic thinking process in medical students. *J Gen Intern Med* 2002;17:839-44.
13 Cahan A, Gilon D, Manor O, Paltiel O. Probabilistic reasoning and clinical decision-making: do doctors overestimate diagnostic probabilities? *QJM* 2003;96:763-9.
14 Attia JR, Nair BR, Sibbritt DW, Ewald BD, Paget NS, Wellard RF, et al. Generating pre-test probabilities: a neglected area in clinical decision making. *Med J Aust* 2004;180:449-54.
15 Heller RF, Sandars JE, Patterson L, McElduff P. GPs' and physicians' interpretation of risks, benefits and diagnostic test results. *Fam Pract* 2004;21:155-9.
16 Phelps MA, Levitt MA. Pretest probability estimates: a pitfall to the clinical utility of evidence-based medicine? *Acad Emerg Med* 2004;11:692-4.
17 Kelly J, Hunt BJ. The utility of pretest probability assessment in patients with clinically suspected venous thromboembolism. *J Thromb Haemost* 2003;1:1888-96.
18 West SG, Emlen W, Wener MH, Kotzin BL. Neuropsychiatric lupus erythematosus: a 10-year prospective study on the value of diagnostic tests. *Am J Med* 1995;99:153-63.
19 Delalande S, de Seze J, Fauchais AL, Hachutta E, Stojkovic T, Ferriby D, et al. Neurologic manifestations in primary Sjogren syndrome: a study of 82 patients. *Medicine (Baltimore)* 2004;83:280-91.
20 Green AJ. Use of 14-3-3 in the diagnosis of Creutzfeldt-Jakob disease. *Biochem Soc Trans* 2002;30:382-6.

(Accepted 17 June 2006)

doi 10.1136/bmj.38915.558738.55

When I use a word

Sometimes, never

If something always happened, what percentage frequency would you assign to that event? Presumably 100%. And if something never happened? Presumably 0%. Well, not everyone shares that opinion. By "always" some mean as infrequently as 91% of the time, and "never" can mean as often as 2% of the time. The combined results of seven studies of what people mean when they use words such as always, commonly, often, frequently, occasionally, sometimes, seldom, rarely, and never are summarised in the table (for references see *Drug Safety* 2005;28:851-70). For comparison, I have also included definitions from the *Oxford English Dictionary*.

Look, for example, at "occasionally," "infrequently," and "seldom"; according to the dictionary they all mean roughly the same thing, but the frequencies that people think these words represent do not overlap at all. Perhaps the lexicographers should reconsider some of their definitions—although surely not "never"—nohow. And perhaps when we use words like this we should remember what the German conductor Hans Richter supposedly once said: "Up with your damned nonsense will I put twice, or perhaps once, but sometimes always, by God, never."

Jeff Aronson *clinical pharmacologist, Oxford (jeffrey.aronson@clinpharm.ox.ac.uk)*

Interpretations of words used to indicate frequencies

Word	Interpretation (range of mean percentages)	Definition in the *Oxford English Dictionary*
Invariably/always	91-100	At every time, on every occasion, at all times, on all occasions. Opposed to sometimes, occasionally
Almost always	85-94	—
Normally	71-81	Under normal or ordinary conditions; as a rule, ordinarily
Usually	70-84	In a usual or wonted manner; according to customary, established, or frequent usage; commonly, customarily, ordinarily; as a rule
More often than not	64	—
Common(ly)	56-69	As a usual circumstance; as a general thing; in ordinary cases; usually, ordinarily, generally
Often	42-71	Many times; at many times; on numerous occasions; frequently; for a significant amount or proportion of the time
Frequent(ly)	36-72	At frequent or short intervals, often, repeatedly
Not infrequently	24-35	Rather frequently
Occasionally	17-21	Now and then, at times, sometimes; irregularly and infrequently
On occasion	12	As need or opportunity arises; now and then, occasionally
Infrequently	12-14	Not frequently; somewhat rarely, seldom
Sometimes	11-33	On some occasions; at times; now and then
Seldom	7-8	On few occasions, in few cases or instances, not often; rarely, infrequently
Almost never	2	Scarcely ever
Very rare(ly)	0.8-3	—
Rare(ly)	0.5-9	Seldom, infrequently, in few instances
Exceptionally	0.4-1	Uncommonly, unusually
Never	0-2	At no time or moment; on no occasion; not ever