Ann. Ist. Super. Sanità
Vol. 27, N. 3 (1991), pp. 395-400

THE MISUSE OF PREDICTIVE VALUE - OR WHY YOU MUST CONSIDER THE ODDS

R. GAMBINO

Columbia University, New York, NY, USA

Summary. - *The predictive value of a test is often misinterpreted because it is presented as a percent. It is intuitive to assume that low percentages (70 % or less) are "bad" and high percentages are "good". A positive predictive value of 20 %, for example, was cited as proof that a test should not be used even though the positive likelihood ratio for that same test was 50. A likelihood ratio of 50 means that the post test odds of disease for a positive test result will be 50 times higher than the pretest odds of disease. Now, that is a large increase in the odds. Critics of laboratory medicine fail to recognize that sensitivity and specificity vary with the strength of the signal. Thus, a value well above the cutoff is far more likely to indicate disease than does a value just above the cutoff - even though both are reported as "positive". Tables of likelihood ratios for a wide range of specific test results, or for multiple test results, provide more information than a simple four-by-four predictive value table. Likelihood ratios are also more informative than predictive values or ROC curves. Finally, critics of laboratory medicine fail to take into account the information to be derived from a confirmatory test, a repeat test at a later time, and from other tests.*

KEY WORDS: (clinical) sensitivity and specificity, predictive value, odds (ratio), likelihood ratio.

Riassunto (L'uso improprio del valore predittivo, ovvero perché bisogna considerare il rapporto di previsione). - *Il valore predittivo di un test viene spesso male interpretato perché si indica come una percentuale. Intuitivamente si assume che le percentuali basse (70% o meno) sono "cattive", mentre quelle alte sono "buone": così, ad esempio, si è riportato come prova che un test non dovrebbe essere usato un valore predittivo per un risultato positivo del 20%, benché fosse 50 il rapporto di verosimiglianza del risultato positivo per lo stesso test. Un rapporto di verosimiglianza di 50 vuol dire che la prevedibilità di malattia "post-test" di un risultato positivo sarà 50 volte più grande di quella "pre-test": e questo è un bell'aumento di prevedibilità. Gli esperti di medicina di laboratorio hanno difficoltà a riconoscere che sensibilità e specificità variano con l'intensità del segnale: un risultato ben al di sopra del valore di discrimine ha una probabilità di gran lunga maggiore di indicare uno stato di malattia che non un risultato appena oltre il discrimine - pur essendo considerati "positivi" entrambi i risultati. Tabelle di rapporti di verosimiglianza per un vasto intervallo di valori risultanti per test specifici, o per test multipli, sono assai più informative di una semplice tabella di contingenza dei valori predittivi. Inoltre, i rapporti di verosimiglianza danno molte più informazioni delle curve ROC. Infine, gli esperti di medicina di laboratorio tendono a non prendere in considerazione l'informazione che deriva da test di conferma, da test ripetuti nel tempo e da test aggiuntivi.*

PAROLE CHIAVE: sensibilità (clinica), specificità (clinica), valore predittivo, rapporto di previsione, rapporto di verosimiglianza.

Introduction

In 1975, Dr. R. Galen and I wrote a book entitled "Beyond normality". Our book helped to popularize the concepts of sensitivity, specificity and prevalence - especially prevalence [1]. We provided many examples of how important it was to consider the prevalence of disease when speaking of the "accuracy" or usefulness of a test. We did not invent these concepts - they were first formulated by the Reverend T. Bayes in 1763 - but we did show how they could be used to decide which tests were better and which tests might be abandoned. Unfortunately, as with any concept, it can mislead if misapplied or is taken to extremes [2]. Moreover, we did not discuss the usefulness of likelihood ratios, the importance of the strength of the signal, or the value of a negative result.

Reprint from *Lab. Report*, September 1989, Vol. 11, no. 9, pp. 65-72 by permission of R. Gambino, author of the article and editor of *Lab. Report*.

This article has been reprinted because of its great relevance in the framework of this monographic issue and its interest for all readers.

The misuse of Bayes' theorem to damn tests for "AIDS"

A typical recent example of the misuse of Bayes' theorem can be found in a new book entitled "Innumeracy: mathematical illiteracy and its consequences", by Prof. J.A. Paulos (which by the way is a book that I recommend very highly) [3]. Prof. Paulos, in order to illustrate the importance of conditional probability, talks about a theoretical cancer test that has a sensitivity of 98% and a specificity of 98%. He also assumes that the prevalence of cancer is 0.5%, or 1 per 200. And he states - right up front - that he is using this example because it has important implications for testing for AIDS.

Prof. Paulos asks the reader to assume that he (the reader) has taken this test for cancer and that the result is reported as "positive". The question Paulos now asks is: "How depressed should you be?". His conclusion (incorrect) is that you should be cautiously optimistic. Why? Because, as Paulos stresses, the predictive value of a positive result is *only* 19.8 % (Table 1). The error made by Professor Paulos is to assume that a "low" percentage for predictive value is bad. On an intuitive basis, of course, a low percentage for the predictive value is "obviously" bad. But intuition, as Prof. Paulos points out time and again, can be misleading.

A positive predictive value of 19.8 % is not bad if you consider the test's ability to increase the probability of disease

We are so used to thinking that 100 % is best - and zero percent the worst - that intuitively we find it hard to accept that a percentage figure of around 20 could be "good". What Prof. Paulos fails to consider is the ability of this test to increase the probability of disease.

To make my point I will structure Prof. Paulos' question in a different way. Let us say that we had a test to detect potential winners of a horse race, and that this test had a sensitivity of 98% and a specificity of 98 percent. Also assume that the prevalence of "winners" among all horses that are racing is only 1 per 200, or 0.5 %. Now the prior probability of picking a "winner" at random from among 10,000 horses is only 0.005. A probability of 0.005 is equal to odds of 0.00502 to 1 - not at all favorable for a bettor who wants to win.

Table 1. - *Positive predictive value of cancer test "X"*

Sensitivity 0.98
Specificity 0.98
Prevalence of cancer 0.005
10,000 subjects tested
Number of true cancers: 0.005 x 10,000 = 50
Number of **true positives**: 50 x 0.98 = **49**
Number without cancer: 0.995 x 10,000 = 9,950
Number of **false positives**: 0.02 x 9,950 = **199**
Total number of positives: 49 + 199 = 248
% of positives that are true positives: 49/248 = 19.8 %

So, in order to improve our chances of picking a winner, we apply our test to the entire field of 10,000 horses and come up with a new field of 248 potential winners to chose from. But Paulos says this is not very good because only 49 of these 248 horses are going to true winners! He has a negative opinion of the test because the "success" rate is less than 20 %. Now who among you wouldn't pay a tout for this secret test - a test that increases the prevalence of winners from a prior figure of 1/200 to a posterior (after the test) figure of 1/5! *What the test has done is to increase the prevalence of "winners" by a factor of 40.* Moreover, as you will see below, the test increases the *ODDS* of having a winner by a factor of 50 - from only 0.00502 to 1 before the test, to 0.25 to 1 after the test. Now that isn't bad at all if you are a betting person.

Calculate the likelihood ratio - not the predictive value

Calculation of the likelihood ratio (LR) provides better insight into what a test can and cannot do for you. The LR for a positive result is obtained by dividing the sensitivity of the test by 1 minus the specificity (the true positive fraction by the false positive fraction):

$$+\,LR = \frac{\text{Sensitivity}}{(1 - \text{specificity})}$$

or

$$+\,LR = \frac{\text{True positive fraction}}{\text{False positive fraction}}$$

In the case of Prof. Paulos' cancer test - or our horse race test - the LR for a positive result is obtained by dividing 0.98 by 0.02 yielding a ratio of 50. Once you know the positive likelihood ratio you can easily convert pre-test odds into post-test odds by simple multiplication. In our example, the pre-test odds of winning are only 0.005 to 1, but the post-test odds are 0.25 to 1 - a 50 fold increase.

Odds and *probability* are related, *but the numbers are not identical.* Odds are expressed as fractions, with the number one as the denominator. Probability is usually expressed as a decimal fraction ranging from zero to one, but probability also can be expressed as a percentage. Odds are useful because only odds - not probability - can be multiplied by the likelihood ratio to obtain the new odds of disease after a test is performed [4-6] (Tables 2 and 3).

Finally, the natural logarithm of the likelihood ratio is a measure of the "weight of evidence" [7].

The likelihood of disease varies with the strength of the signal

The next error Prof. Paulos makes is to fail to consider the strength of the signal. Most problems arise when results are in the "overlap" zone - where the distribution of

results for subjects without the condition overlaps the distribution of those with the condition. He assumes that all positive results are equal. That simply is not correct. Sensitivity and specificity vary with the strength of the signal. A result near the cutoff between "positive" and "negative" has poorer sensitivity or specificity than a result that is far above or below the cutoff. In fact, once a result is outside the overlap zone it has either 100 % sensitivity or 100 % specificity (Fig. 1).

To improve the diagnostic utility of test results that are in the overlap zone, it is possible and desiderable to develop a whole family of likelihood ratios for different reported values. Such tables have been published for creatine kinase isoenzyme MB (CK-2) by Dr. Bernstein and associates at the Bridgeport Hospital in Bridgeport, Connecticut; and earlier by Drs. van der Helm and Hische for glucose [8, 9].

Dr. Bernstein and his associated calculated what they called "Bayes factors" for a large number of specific CK-2 values obtained on admission and at 12 h after admission. For any combination of CK-2 values at zero time and 12 h they calculated a "Bayes factor". Their "Bayes factor" is in actuality the likelihood ratio. It would have been better if they had simply provided the reader with the likelihood ratio.

Drs. van der Helm and Hische calculated likelihood ratios for diabetes based on 2 h post prandial glucose values reported in a study by Reiman and Wilkerson. Drs. van der Helm and Hische point out that it is too simplistic to simply derive likelihood ratios from raw sensitivity and specificity data [9]. You must also consider the shape of the distribution curves for the diseased and non-diseased populations - specifically in the overlap zone - *and recalculate likelihood ratios based on the ratio of the heights of the overlapping distributions at the specific test values in question* (Table 4).

Table 2. - *Conversion formulae for odds and probability*

$$\text{Odds} = \frac{\text{Probability (as decimal fraction)}}{\text{1 - probability}}$$

$$\text{Probability} = \frac{\text{Odds}}{\text{1 + Odds}}$$

Table 3. - *Relationship of odds to probability*

Odds	Probability
0.25/1	0.20
0.50/1	0.33
1.00/1	0.50
2.00/1	0.66
4.00/1	0.80

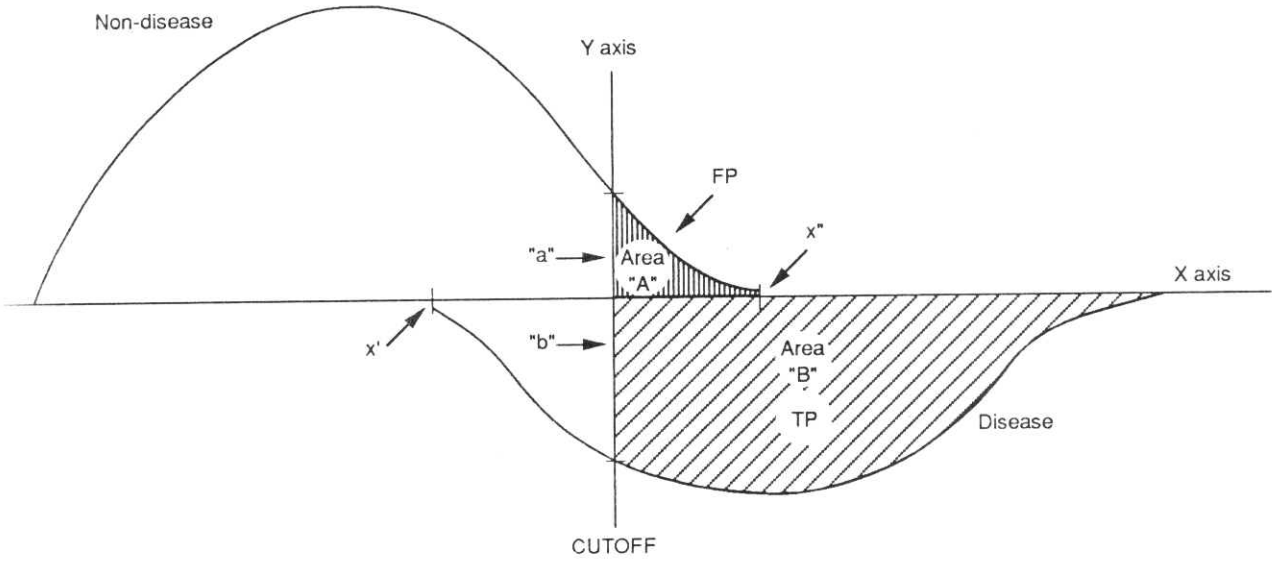

Fig. 1. - Likelihood ratios for overlapping distributions. The figure shows two overlapping distributions placed one on top of the other in a flip-flop fashion in order to make it easier to analyze. In the left upper quadrant is a theoretical distribution for a non-diseased population. In the right lower quadrant is a theoretical distribution for a diseased population. The X axis represents increasing values for a test result. Test results that are less than the lowest value (marked x') for the diseased population have 100% sensitivity because disease is always present. Test results that are greater than the highest value (marked x") for the non-diseased population are 100% specific because non-disease is always absent.
The figure shows the two ways to calculate a likelihood ratio in the overlap zone for a specific cutoff - in this case indicated by the Y axis. The less correct way is to define the likelihood ratio for cutoff Y as the ratio of area B (the total number of true positive above the cutoff, including positive results above the overlap zone) to that of area A (the total number of false positive above the cutoff). In the example shown, the likelihood ratio for disease - given a result at Y - is approximately 15 since area B is approximately 15 times greater than area A.
The more correct way to define the likelihood ratio for a specific cutoff value in the overlap zone is to compare the heights of the frequency distributions at that cutoff. In the example shown, this would be the ratio of "b" (the number of true positive cases at that value) to "a" (the number of false positive cases at that value). This ratio is approximately 1.5, a value that is far less than that obtained when areas are used. For this test and study population the likelihood of disease for a value at Y is not too great - as you can see by looking at the distributions.

Table 4. - *Example of the change in the likelihood ratio (simplistic and correct) with a change in the reported result for glucose* [8]

Glucose level 2 h post prandial mmol/l (mg/dl)	Simplistic likelihood based on ratio of area of frequency curves	Correct likelihood based on ratio of heights of frequency curves at specific test values
3.89 (70)	1.08	0.09
4.44 (80)	1.30	0.13
5.00 (90)	1.80	0.26
5.55 (100)	2.93	0.20
6.11 (110)	5.39	1.70
6.66 (120)	9.52	1.61
7.22 (130)	20.74	2.88
7.77 (140)	**95.17**	**35.50**
8.33 (150)	125.00	14.50
8.88 (160)	235.00	21.00
9.44 (170)	> 1000	> 1000
9.99 (180)	> 1000	> 1000
10.55 (190)	> 1000	> 1000
11.10 (200)	> 1000	> 1000

It is interesting to note that either method of calculating the likelihood ratio yields a striking increase in the ratio at the usual discriminant value (7.77 mmol/l) for a post prandial glucose test. But the likelihood ratio obtained by the more accurate method - which is derived from the exact ratio of the *heights* of the frequency curves at the specific test value - is one-third of that of the less accurate method - which is derived from the gross ratio of the *areas* of the frequency distribution curves above and below the cut point. It is also interesting to note that the more accurate method correctly provides negative odds for diabetes when post prandial glucose levels are less than 5.55 mmol/l (100 mg/dl).

And what about ROC curves?

As defined by J.A. Swets, who has written the most on the subject, ROC can stand for "Receiver Operating Characteristic" when used in the field of signal detection, or for "Relative Operating Characteristic" when used in generalized applications [10]. The ROC curve is a plot of sensitivity on the Y axis, and 1 minus the specificity on the X axis (or in other words, the true positive fraction on the Y axis and the false positive fraction on the X axis) for selected reported values of a test. The shape of the curve provides a gestalt as to the usefulness of a test. Tests that plot in the upper left hand corner are better than tests that plot closer to the center. That's because the upper left corner defines high sensitivity and high specificity. The ROC curve can also be plotted as sensitivity versus specificity, and in that case the best tests plot in the upper right hand corner (Fig. 2) [11].

The area under the ROC curve is a measure of accuracy. The preferred measure of overall accuracy is the proportion of the area of the entire graph that lies beneath the

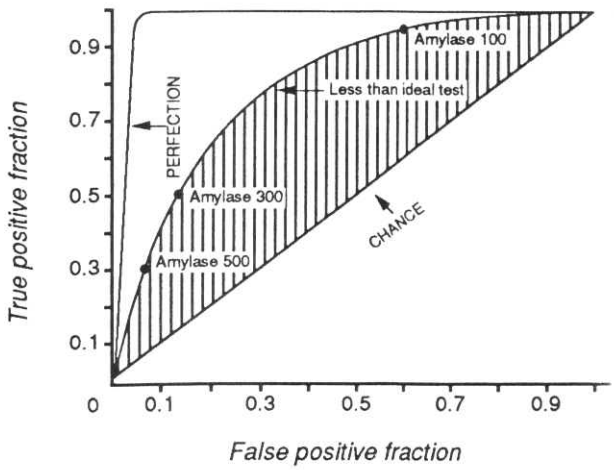

Fig. 2. - Theoretical ROC curves. The figure shows three theoretical ROC curves. The Y axis is the true positive fraction (sensitivity) of the test, ranging from zero at the bottom to 1.0 at the top. The X axis is the false positive fraction (1 minus specificity), ranging from zero on the left to 1.0 on the right. ROC curves are generated by calculating the sensitivity and specificity of a test for varying cutoff values in the overlap zone and then plotting the true positive and false positive fractions obtained. The information content of the ROC curve is increased if the specific test value - used to obtain a particular TP/FP point - is written beside the plotted point.

The diagonal line running from the lower left hand corner to the right upper corner at a 45 degree angle represents the results you would obtain by chance alone, *i.e.* by a flip of an unbiased coin. The curve that runs straight up the Y axis on the left and then along the very top to the right is the curve for a perfect test - one with 100 % sensitivity and specificity. The final curve (the one in the middle) represents a less than ideal but typical laboratory test. In this case serum amylase for pancreatitis. Note that for an amylase value of 100, the true positive fraction is nearly 1.0, but the false positive fraction is high at 0.6 (high sensitivity but low specificity). For a value of 500, on the other hand, the true positive fraction is now only 0.3, but the false positive fraction is less than 0.1 (low sensitivity but high specificity).

Table 5. - *Some typical test combinations with higher information value than that provided by a single test*

- Thyroxine or free thyroxine with TSH
- BUN with creatinine
- Glucose with fructosamine or glycohemoglobin
- Iron and iron-binding capacity followed by ferritin
- Cholesterol with HDL cholesterol
- Apolipoprotein B with apolipoprotein A-I

curve. Specifically, if the area beneath the curve is 0.5 or less of the area of the entire graph, then no discrimination exists because such a result can be achieved by chance alone; but if the proportion of the area under the curve is 1.0 of the entire graph, then perfect discrimination exists. In the case of perfect discrimination, the curve lies along the upper Y axis and along the top of the graph (Fig. 2).

I find ROC curves most useful for comparing one or more tests with another. ROC curves are also helpful in making an estimate of optimum cutoff values for "positive" and "negative" results. But ROC curves are not as useful as likelihood ratios for actually defining what a test is going to do for you - given a specific result - independent of the reference range for the test. Nor do ROC curves provide a direct measure of likelihood. Finally, a ROC curve is not the best way to demonstrate how well a test discriminates between two diseases [12]. A plot or table of likelihood ratios provides more quantitative information than a simple plot of sensitivity and specificity.

A test should never be interpreted in isolation

Still another error made by Prof. Paulos is to fail to discuss the utility of a confirmatory test and a repeat test at a later date. A test for antibody to HIV, for example, never stands by itself. Any result considered positive is confirmed by an independent test. Confirmation is a general testing principle that is used widely in science and in medicine - but the value and need for confirmation is too frequently forgotten when a writer, such as E.R. Pinckney, a former associated editor of JAMA, wants to damn the routine use of laboratory tests [13]. Confirmation is also obtained by performing one or more different tests (Table 5). An analysis of multiple test results almost always provides more information than does an analysis of a single test. For example, the combination of thyroxine and TSH is far more informative than either test alone. *Finally, and most important, confirmation is obtained by what actually happens to patients.*

REFERENCES

1. GALEN, R.S. & GAMBINO, S.R. 1975. *Beyond normality: the predictive value and efficiency of medical diagnoses*. John Wiley, New York.
2. FEINSTEIN, A.R. 1977. Clinical biostatistics XXXIX. The haze of Bayes, the aerial palaces of decision analysis, and the computerized Quija board. *Clin. Pharmacol. Ther.* **21**: 482-496.
3. PAULOS, J.A. 1988. *Innumeracy: mathematical illiteracy and its consequences*. Hill and Wang, New York.
4. GAMBINO, S.R. 1986. Odds, probability and likelihood ratios. *Lab Report* **8**: 69-71.
5. ALBERT, A. 1982. On the use and computation of likelihood ratios in clinical chemistry. *Clin. Chem.* **28**: 1113-1119.
6. SIMEL, D.L. 1985. Playing the odds. *Lancet* **i**: 329-330.
7. GOOD, I.J. 1960. Weight of evidence, corroboration, explanatory power, information and the utility of experiments. *J. R. Statis. Soc.* **22**(B): 319-331.
8. BERNSTEIN, L.H., GOOD, I.J., HOLTZMAN, G.I., DEATON, M.L. & BABB, J. 1989. Diagnosis of acute myocardial infarction from two measurements of creatine kinase isoenzyme MB with use of nonparametric probability estimation. *Clin. Chem.* **35**: 444-447.
9. van der HELM, H.J. & HISCHE, E.A.H. 1979. Application of Bayes' theorem to results of quantitative clinical chemical determinations. *Clin. Chem.* **25**: 985-988.
10. SWETS, J.A. 1988. Measuring the accuracy of diagnostic systems. *The Sciences* **240**: 1285-1293.
11. WERNER, M., STEINBERG, W.M. & PAULEY, C. 1989. Strategic use of individual and combined enzyme indicators for acute pancreatitis analyzed by receiver-operator characteristics. *Clin. Chem.* **35**: 967-971.
12. HARTZ, A.J. 1984. Validity and bias in laboratory tests. *Arch. Pathol. Lab. Med.* **108**: 769-771.
13. PINCKNEY, E.R. & PINCKNEY, C. 1989. Unnecessary measures: physicians are relying too heavily on medical tests. *The Science* **29**: 20-27.