A New Conceptual Approach to Teaching the Interpretation of Clinical Tests.

Shai Linn
School of Public Health, Haifa University
and Rambam Medical Center, Haifa, Israel

Journal of Statistics Education Volume 12, Number 3 (2004),
ww2.amstat.org/publications/jse/v12n3/linn.html

Copyright © 2004 by Shai Linn, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the authors and advance notification of the editor.

Key Words: Bayes' Theorem; Diagnosis; Predictive value

Abstract

Courses in clinical epidemiology usually include acquainting students with a single 2X2 table. All diagnostic test characteristics are explained using this table. This pedagogic approach may be misleading. A new didactic approach is hereby proposed, using two tables, each with specific analogous notations (uppercase and lowercase) and derived equations. This approach makes it easier to discuss the use of Bayes’ Theorem and the two stages of analyses, i.e., using sensitivity to calculate predictive values. Two different types of false negative rates and false positive rates are discussed.

1. Introduction

The standard practice of teaching clinical epidemiology (Baron, 2001) includes acquainting students with the concepts of assessing diagnostic test characteristics. It is often based on a single 2X2 table (Table 1) from which all diagnostic test characteristics are calculated, i.e., the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) Dawson and Trapp 1994; Greenberg, Daniels, Flanders, Eley and Boring 2001). Such a presentation in one table is potentially confusing (Baron 2001, p.243). In fact, test performances in two populations are of interest, and in two stages: First, the evaluation in a selected study population in which the study groups are defined by the disease status (persons with the disease and persons without disease are compared), i.e., a case control study (Pepe 2003). Based on pre-determined numbers of diseased and non-diseased subjects, the relative frequency of diseased subjects is usually much higher in a case control study than in the population from which the cases and controls are drawn. In this population both the diagnostic test and the definitive test (the gold standard) are performed and presented in a 2X2 table (Table 2), and the sensitivity and specificity are calculated. The other population is the general patient (target population) population to which the diagnostic test is applied, in which the study groups are defined by the test status (persons
with positive tests and persons with negative tests are compared) (Table 3).

However, we typically do not have the information on this population because it is often unfeasible and unethical to perform both the diagnostic tests and an additional definitive test to determine the true diagnosis according to the gold standard (Sackett and Haynes, 2002). For example, using angiography as a gold standard for diagnosing cardiac ischemia by electrocardiographic changes is “not a very attractive alternative in terms of discomfort, risk and cost” (Sackett, Haynes, Guyatt, and Tugwell 1991, p. 101).

Therefore, the PPV and NPV are calculated from the sensitivity and specificity and the prevalence of the disease in the target population, using Bayes’ Theorem. Thus, a presentation in one table (Table 1) for analyses in two populations may be pedagogically misleading. A new approach, using two tables (Table 2 and Table 3) instead of one table (Table 1) and specific notations for each table, is hereby proposed.

2. Why Do We Need a New Approach?

All textbooks and papers emphasize that the PPV and NPV should be obtained from the sensitivity and specificity, and information on the prevalence of the disease in the target population, using Bayes’ Theorem. However, in many leading textbooks of biostatistics and epidemiology (Riffenburgh 1993; Sox, Blatt, Higgins, and Marton 1988; Altman 1991; Sackett, Haynes, Guyatt, and Tugwell 1991; Kraemer 1992; Beaglehole, Bonita, and Kjellstrom 1993; Bradley 1993; Essex-Sorlie 1995; Wassertheil 1995; Weiss 1996; Silva 1999; Riegelman 2000; Dawson and Trapp 2001; Greenberg, Daniels, Flanders, Eley, and Boring 2001; Sackett and Haynes 2002; Bhopal 2002) the sensitivity (or specificity) and the PPV (or NPV) are defined using a single table (Table 1).

---

Table 1. Single "generic" table often used to describe both study results and the results of applications of the clinical test to the patients (target) population.

| Clinical Test | Gold Standard | Total |
|---------------|---------------|-------|
|               | S+            | S-    | Total |
| T+            | a=True Positive | b=False Positive | a+b |
| T-            | c=False Negative | d=True Negative | c+d |
| Total         | a+c           | b+d   |

Note: The table demonstrates a misleading presentation in that all test characteristics are calculated in one table.

Sensitivity = a/a+c
Specificity = d/b+d
Positive Predictive Value (PPV) = a/a+b
Negative Predictive Value (NPV) = d/c+d

This may be misleading to many students for the following reasons:
1. A presentation of the definition in one table fails to demonstrate the conceptual distinction of the "study population" in which the test characteristics are determined, and the "patient (target) population" to which the test is applied afterward to obtain the posterior probabilities. Thus, a presentation in a single table implies erroneously a cross-sectional situation, and does not convey the sequence of analyses. Therefore, it is often not clear why one should not calculate the PPV (or NPV) directly from Table 1.

2. The need for using Bayes' Theorem and prevalence to calculate PPV (or NPV) is not obvious from a presentation of calculating sensitivity and PPV in one table. The presentation in Table 1 erroneously implies simultaneous calculations in both axes of the table: the "vertical" disease axis AND the "horizontal" test-results axis. In fact, calculation of the sensitivity and specificity is possible only when the "diseased and non-diseased persons are sampled" (Baron 2001, p. 243), and direct evaluation of the PPV from Table 1 would be "misleading" unless "the proportion of patients with the disease in the study population equals the proportion of patients with the disease in the population in which the test is applied" (Weinstein and Finberg 1980, p. 86-87). In practice, the analyses are performed in two stages: first one uses a selective (study) population in which the sensitivity and specificity are calculated; then, Bayes' Theorem and the prevalence are used, together with the sensitivity and specificity, to calculate PPV or NPV (Hirsch and Riegelman 1996; Greenberg, Daniels, Flanders, Eley and Boring 2001).

3. A presentation of the two stages in one table may be difficult for most students to comprehend. It should be emphasized that Bayes’ Theorem enables students to comprehend the interrelationships between the prevalence, sensitivity and specificity, a fundamental characteristic of clinical epidemiology. When predictive values are calculated from the sensitivity and specificity, their magnitude depends on the prevalence of the disease. Failure to understand the two stages of analyses would not let students appreciate the importance of the prevalence of the disease, in calculations based on Bayes’ Theorem.

Finally, using one table to teach diagnostic test characteristics often makes the definitions of rates unclear (Riffenburgh 1993; Weinstein and Finberg 1980; Hirsch and Riegelman 1996). Is the “true positive rate” referring to the sensitivity (as often defined)? Or is it referring to the predictive value (as often understood by students, physicians or patients)? Analogous considerations are valid for true negative rates, false positive rates and false negative rates. Because of this confusion, Hirsch and Riegelman (1996, pp 11-12) recommended not using these terms at all. We offer a way to overcome these difficulties by presenting the analyses in two tables.

3. A New Approach

We suggest a simple modification of the current teaching practices and notations in order to help students clearly understand the methodology. Two tables should be used, with specific terminologies for each table. One table with lowercase notations (Table 2) is used to calculate the sensitivity and specificity among the selected population in which these measures are examined. Sampling of this population is done “vertically” i.e., the diseased and non-diseased are sampled. Thus, it is appropriate to have a+c and b+d as totals in this table. However, it is inappropriate to have totals of the “horizontal” axis of test results. In this table we have

\[
\text{sensitivity} = P(T^+ | S^+) = \frac{a}{a+c} \quad \text{and} \quad \text{specificity} = P(T^- | S^-) = \frac{d}{b+d}
\]
We use S to denote sickness rather than D which could have been used as an acronym for diseased, because of the use of the letter D in the tables.

Table 2. Table describing study results (in a selected population). Disease-oriented sampling.

| Clinical Test | Gold Standard |
|---------------|---------------|
|               | S+            | S-            |
| T+            | a=True Positive | b=False Positive |
| T-            | c=False Negative | d=True Negative |
| Total         | a+c           | b+d           |

Note: The table demonstrates a more appropriate presentation for the study (selected) population.

Sensitivity = \( \frac{a}{a+c} \)
Specificity = \( \frac{d}{b+d} \)

A second table with uppercase notations (Table 3) is used to explain the predictive values among the target population in which the test would be applied for screening or clinical diagnosis. Sampling of this population is done horizontally, i.e., those with positive and negative tests. Thus, it is appropriate to have \( A+B \) and \( C+D \) as totals in this table if a physician monitors the success of the clinical test by ascertaining the disease status (the gold standard status) of persons with positive and/or negative test results. However, it is inappropriate to have totals of the “vertical” axis of test results. Thus, we define the Positive Predictive Value (PPV) and the Negative Predictive Value (NPV) as follows:

\[
PPV = P(S+ | T+) = \frac{A}{A+B} \quad \text{and} \quad NPV = P(S- | T-) = \frac{D}{C+D}
\]

Table 3. Table describing the results in the patient (target) population to which the clinical test is applied.

| Clinical Test | Gold Standard |
|---------------|---------------|
|               | S+            | S-            | Total |
| T+            | A=True Positive | B=False Positive | A+B   |
| T-            | C=False Negative | D=True Negative | C+D   |
Note: The table demonstrates a more appropriate presentation for the patient target population.

Positive Predictive Value (PPV) = \( \frac{A}{A+B} \)
Negative Predictive Value (NPV) = \( \frac{D}{C+D} \)
False Positive Rate (FPR) = \( \frac{B}{A+B} \)
False Negative Rate (FNR) = \( \frac{C}{C+D} \)

It is now obvious that the translation of information on sensitivity and specificity to PPV or NPV must be done by using Bayes’ Theorem and the prevalence \( P(S+) \).

Positive Predictive Value, PPV

\[
P_{PPV} = P(S+ | T+) = \frac{P(T+ | S+)P(S+)}{P(T+ | S+)P(S+) + P(T+ | S-)P(S-)}
\]

\[
P_{PPV} = \frac{A}{A+B} = \frac{\alpha}{\alpha + c} P(S+) + \frac{b}{b + d} P(S-)
\]

Similarly, Negative Predictive Value, NPV

\[
N_{PV} = P(S- | T-) = \frac{P(T- | S-)P(S-)}{P(T- | S-)P(S-) + P(T- | S+)P(S+)}
\]

\[
N_{PV} = \frac{D}{C+D} = \frac{d}{b + d} P(S+) + \frac{c}{\alpha + c} P(S-)
\]

3.1 Two tables used for two different definition of false positive rates and false negative rates.

The definitions of false positive rates and false negative rates are often unclear in a one-table presentation. Moreover, the definitions of these rates in the literature are inconsistent, often based on errors in the selected study population of Table 2, and sometimes based on errors in the target population, as presented in Table 3.

3.2 Definitions of false positive rates (fpr) and false negative rates (fnr) in a selected population (Table 2).

Most textbooks (Riffenburgh 1993; Sox, Blatt, Higgins, and Marton 1988; Altman 1991; Sackett,
Haynes, Guyatt, and Tugwell 1991; Kraemer 1992; Beaglehole, Bonita, and Kjellstrom 1993; Bradley 1993; Essex-Sorlie 1995; Wassertheil 1995; Weiss 1996; Silva 1999; Riegelman 2000; Dawson and Trapp 2001; Greenberg, Daniels, Flanders, Eley, and Boring 2001; Sackett and Haynes 2002; Bhopal 2002; Pepe 2003) define the false positive rate as the probability that a disease-free patient has a positive test result, and false negative rates as the probability that a diseased patient has a negative test result.

When the diseased and non-diseased are sampled, in a case control study, the definitions are:

False positive rate among persons without the disease is

\[ fpr = \frac{P(T - | S -)}{b + d} \]

i.e., \( fpr = 1 - \text{specificity} \)

and

False negative rate among persons with the disease is

\[ fnr = \frac{P(T + | S +)}{c} \]

i.e., \( fnr = 1 - \text{sensitivity} \).

These definitions of the \( fpr \) and \( fnr \), which are based on Table 2, appear in most of the above-mentioned textbooks.

3.3 Why do we need a different definition of errors, in the target population?

The above-mentioned statistics do not assess the accuracy of the diagnostic test in a clinically useful way, i.e. in a way that can be used by patients and physicians. Clinicians and patients relate to the sequelae of the clinical test results and do not know, at the time of performing the test, who has the disease and who does not (otherwise the test would not have been performed). What is of interest to both physicians and patients is how many of the positively diagnosed patients, in fact, do not have the disease and how often a person with the disease is not diagnosed by the test. The monetary and human cost and consequences of these diagnostic mistakes can be studied and discussed by physicians and patients, based on the consequences of missing a diagnosis or of applying unnecessary additional tests because of a false diagnosis.

3.4 Definitions of false positive rates (FPR) and false negative rates (FNR) in the target population (Table 3).

Fleiss (1981) defined the false positive rate as the proportion of people, among those responding positive to the diagnostic test, who are actually free of the disease. Similarly, Fleiss (1981) defined the false negative rate as the proportion of people, among those responding negative on the test, who nevertheless have the disease. These measures of errors, the FPR and FNR, in the target population are of greater interest to patients and physicians, who are more concerned with wrong diagnoses after applying a diagnostic test than with the errors in a selected case control study, the \( fpr \) and \( fnr \).

Following Fleiss (1981), we can define these measures of interest in the general patient population
False positive rate among persons with a positive test is

$$FPR = P(S^- | T^+) = \frac{B}{A + B}$$

i.e., $FPR = 1 - PPV$.

This statistic indicates the rate of non-diseased persons who would erroneously be classified as having the disease by the clinical diagnostic test.

Clearly, using Bayes’ Theorem:

$$FNR = P(S^+ | T^-) = \frac{C}{C + D}$$

Similarly,

False negative rate among persons with a negative test is

$$FPR = P(S^- | T^+) = \frac{P(T+ | S-)P(S^-)}{P(T+ | S-)P(S^-) + P(T+ | S+)P(S^+)}$$

$$FPR = \frac{B}{A + B} = \frac{b}{b + d} P(S^-) + \frac{a}{a + c} P(S^+)$$

i.e., $FNR = 1 - NPV$.

This statistic indicates the rate of diseased persons who would erroneously be classified as not having the disease by the clinical diagnostic test.

Clearly, using Bayes' Theorem

$$FNR = P(S^+ | T^-) = \frac{P(T- | S+)P(S^+)}{P(T- | S+)P(S^+) + P(T- | S-)P(S^-)}$$

$$FNR = \frac{C}{C + D} = \frac{c}{a + c} P(S^+) + \frac{d}{b + d} P(S^-)$$

Thus, the two-table presentation enables clear pedagogical distinction of the definitions of error rates in (Table 3), using uppercase notations:
the two different populations, the selected case-control study population (Table 2), i.e., $fpr$ and $fnr$, and
the target population (Table 3), i.e., FPR and FNR.

### 3.5 An Example

In a case control study, pathological diagnoses of skin cancer and benign tumors (defining the “disease status”) were recorded as was the pre-resection clinical diagnostic evaluation (the “test”) of a dermatologist (Table 4, analogous to Table 2 above).

#### Table 4. Detection of malignant skin cancer by a physician (analogous to Table 2)

| Clinical Test | Final diagnosis by pathology, the Gold Standard |
|---------------|-----------------------------------------------|
|               | Skin cancer | No skin cancer |
| Diagnosis of skin cancer | T+ | 63 | 6 |
| No diagnosis of skin cancer | T- | 10 | 112 |
| Total | 73 | 118 |

The data for this study indicate a sensitivity of 86.3%, a specificity of 94.9%, a $fpr$ of 5.1% and a $fnr$ of 13.7%. However, PPV, or NPV and the error rates in the general population cannot be calculated from Table 4. Such erroneous estimates would apply to the physician study population alone, and would yield uninformative (and misleading) PPV of 91.3%, NPV of 91.8%, FPR of 8.7% and FNR of 8.2%. Such a single-table presentation would be misleading, because it is incorrect to calculate the PPV and NPV of clinical examinations in the general population from these data. Rather, based on the sensitivity and specificity, a national prevalence of skin cancer of, say, 0.08%, and Bayes’ Theorem, the calculated PPV would be approximately 13.407%, quite different from the PPV for the physician in a dermatology clinic. This discrepancy occurs because of the low prevalence of the disease in the general population. Similar calculations would yield NPV of 99.98845%, FPR of 98.6593% and FNR of 0.001096%. The data for the general population could be reconstructed by first determining the margins according to the prevalence, i.e., 8 patients with melanoma for 10000 persons in the general population. Then, the sensitivity and specificity can be used to yield Table 5, which is the correct presentation for the general population (because of rounding to integers in constructing the table, direct calculations from Table 5 would yield estimates slightly different from the above calculations, based on Bayes’ Theorem).

#### Table 5. Detection of malignant skin cancer in the general population (analogous to Table 3).

| Final diagnosis by pathology, the Gold Standard |
|-----------------------------------------------|
| Skin cancer | No skin cancer |
| 8 of 11 |
The prevalence is 0.08%, thus we expect 8 patients (a rounded number) with melanoma and 9992 healthy persons in 10000 persons.

Using a sensitivity of 86.3%, we calculate $A=7 \times (0.863 \times 8)$.

Using a specificity of 94.9%, we calculate $D=9482 \times (0.949 \times 9992)$.

## 4 Conclusion

This manuscript focuses on an approach to teaching the main characteristics of a diagnostic test, the PPV and NPV, when it is applied in the general population, in a situation such as screening for cancer. The manuscript does not deal with validation of a test through repeated testing.

As has been mentioned above, most textbooks present both the sensitivity and specificity or the PPV or NPV in a single table. Moreover, some would prefer, pedagogically, to begin with a simpler one 2X2 table and then proceed on to a more conceptually correct - but perhaps more complex - two 2X2 table presentation. It is suggested using a two-table presentation for advanced students, or including a transition from a one-table to a two-table presentation even if one begins teaching using a simple one table. Eventually, using two tables to describe diagnostic test characteristics is, in our experience, pedagogically and conceptually more acceptable to students.

Using the two tables and the derived equations demonstrates clearly the use of Bayes’ Theorem, test characteristics (the sensitivity and specificity) and the prevalence to calculate PPV. It is more obvious that the analyses are done in two stages, for two different populations: the selected study population and the target population. This approach makes it easier to discuss and define two different types of false negative rates and false positive rates in the two populations.

## Notation

\[ P(T+) = \text{probability of the diagnostic test being positive} \]

\[ P(T-) = \text{probability of the diagnostic test being negative} \]

\[ P(S+) = \text{probability of the disease, i.e., the prevalence} \]

\[ P(S-) = \text{probability of no disease, i.e., 1-prevalence} \]

vertical line (|) stands for "given that"
Acknowledgements

The author acknowledges the questions and comments made by Mr. Peter Grunau, a medical student, who was one of the first students to see the ideas expressed in this paper.

References

Altman D.G. (1991), Practical statistics for medical research, Chapman & Hall London.

Baron, J.A. (2001), "Clinical epidemiology," in Teaching Epidemiology eds. Olsen J., Saracci R., and Trichopoulos D., Oxford: Oxford University Press, pp. 237-249.

Beaglehole, R., Bonita, R., and Kjellstrom, T. (1993), Basic Epidemiology, Geneva: World Health Organization.

Bhopal, R. (2002), Concepts of Epidemiology, Oxford: Oxford University Press.

Bradley, G.W. (1993), Disease Diagnosis and Decision, New York: John Wiley & Sons.

Dawson, B., and Trapp, R.G. (1994), Basic and Clinical Biostatistics, New York: Lange–McGraw-Hill.

Dawson, B., and Trapp, R.G. (2001), Basic and Clinical Biostatistics, New York: Lange Medical Books-McGraw Hill.

Greenberg, R.S., Daniels, S.R., Flanders, W.D., Eley, J.W., and Boring, J.R. (2001), Medical Epidemiology, London: Lange-McGraw-Hill.

Essex-Sorlie, D. (1995), Medical Biostatistics and Epidemiology, New York: Appleton & Lange/McGraw Hill.

Fleiss, J.L. (1981), Statistical Methods for Rates and Proportions (2nd ed.), New York: John Wiley & Sons.

Hirsch, R.P., and Riegelman R.K. (1996), Statistical Operations, Oxford: Blackwell Science.

Jenicek, M. (1995), The Logic of Modern Medicine, Montreal: EPIDEM International.

Kraemer, H.C. (1992), Evaluation of Medical Tests: Objective and quantitative guidelines, London: Sage Publications.

Pepe, M. S. (2003), The Statistical Evaluation of Medical Tests for Classification and Prediction, Oxford Statistical Science Series 28, Oxford: Oxford University Press.

Riegelman, R.K. (2000), Studying a Study and Testing a Test, Philadelphia: Lippincott Williams & Wilkins.
Riffenburgh, R.H. (1993), *Statistics in Medicine*, San Diego: Academic Press.

Sackett, D.L., Haynes, R.B., Guyatt, G.H., and Tugwell, P. (1991), *Clinical Epidemiology (2nd ed.*), Boston: Little Brown & Company.

Sackett, D., and Haynes, R.B. (2002), "The Architecture of Diagnostic Research," in *The Evidence Base of Clinical Diagnosis*. ed. J.A. Knottnerus, London: BMJ Publishing.

Silva, S.I. (1999), *Cancer Epidemiology: Principles and Methods*. Geneva: International Agency for Research on Cancer, World Health Organization.

Sox, H.C., Blatt, M.A., Higgins, M.C., and Marton K.I. (1988), *Medical Decision Making*, Boston: Butterworth-Heinemann.

Wassertheil, S. (1995), *Biostatistics and Epidemiology*, New York: Springer-Verlag.

Weinstein, M.C., and Finberg, H.V. (1980), *Clinical Decision Analysis*, Philadelphia: W.B. Saunders Co.

Weiss, N.S. (1996), *Clinical Epidemiology*, Oxford: Oxford University Press.

Shai Linn
School of Public Health
Faculty of Welfare and Health Studies
Haifa University
and Unit of Clinical Epidemiology,
Rambam Medical Center
Haifa
Israel
slinn@univ.haifa.ac.il