J Clin Epidemiol Vol. 52, No. 11, pp. 1073–1082, 1999

0895-4356/99/$–see front matter
PII S0895-4356(99)00087-6

The Two by Two Diagram: A Graphical Truth Table

Kevin M. Johnson

Department of Diagnostic Radiology, Yale University School of Medicine, New Haven, Connecticut

ABSTRACT. The two by two table is widely used in statistics, and in particular in the medical literature, to present the results of experimental and clinical studies in which two different operators (such as a reference standard and a new diagnostic test) sort a sample into two groups. The four cells of the table reflect the four possible categories of results. Despite the simplicity of the table, the interplay of its contents is surprisingly complex, and description of this interplay can be confusing if rendered with the conventional descriptive terms alone, such as sensitivity and predictive value. We present a graphical transformation of the table, comprised of a rectangular box in a special coordinate system. This diagram is offered as a flexible and subtle conceptual tool to help physicians, authors, and students to understand, plan, and present clinical research studies. J CLIN EPIDEMIOL 52;11:1073–1082, 1999. © 1999 Elsevier Science Inc.

KEY WORDS. Sensitivity and specificity, truth table, statistics, predictive value of tests, data interpretation, prevalence, contingency table

INTRODUCTION

One can assume that statisticians and epidemiologists have a firm grasp of the myriad statistical concepts invoked in the analysis of diagnostic tests, but many physicians do not [1]. Since the medical literature can improve the lot of patients only if clinicians understand what they read, there is a need for clearer statistical explanations, and a need for these concepts to be more easily taught and retained.

Many researchers employ the two by two table, sometimes called a truth or contingency table, to classify their experimental study results. Consider a discrete sample of any kind in which each individual member of the sample either has or does not have a trait we wish to detect. Let one operator (e.g., a reference test, or "gold standard") separate the sample into two groups, one with trait, and the other without. Let a second operator (e.g., a new diagnostic test) do the same, again starting with the original sample. Each individual then must fall into one of four categories; results are customarily arranged into a 2×2 table of cells (Fig. 1, left).

Despite the seeming simplicity of this table, the number of ways to extract conclusions from it are surprisingly varied [2]. The analysis can be simplified by transforming the table into a simple graphical form. On this "two by two diagram," all of the traditional descriptive terms such as sensitivity and specificity can be readily and clearly defined; the diagram components are easily manipulated to understand the interdependencies of the underlying numbers.

Visual representation can be a powerful conceptual tool; the history of science offers numerous examples including vector diagrams, the periodic table, and Feynman diagrams [3,4]. Although simple notions like sensitivity are easily explained in isolation, the 2×2 diagram permits two or more simple concepts, such as sensitivity and positive predictive value, to be distinguished and the dynamic link between them to be explained clearly. In this way the diagram provides a unifying conceptual framework. However, its chief virtue is that it allows a fluid mental manipulation of more subtle concepts like expected value, verification (referral) bias, and Bayes' theorem among others.

This report introduces the basics of the 2×2 diagram and presents several theoretical and actual clinical examples.

METHODS

For clarity's sake, the discussion will be restricted to diagnostic tests. In this situation, a diagnostic test is pitted against a reference, or "gold," standard. However, the 2×2 diagram applies to any 2×2 table; only the names of the coordinate system hemiaxes are changed to help keep the features of the diagram clear. The diagram can be used just as well to compare two human observers when there is no reference standard, or to analyze the results of a therapeutic trial; these applications cannot be treated here. Its application to diagnostic tests is then a particular instance of a more general use.

Address for correspondence: Kevin M. Johnson, MD, Department of Diagnostic Radiology, Yale University School of Medicine, 333 Cedar Street, New Haven, CT 06510. E-mail: <kevin.johnson@yale.edu>.

Accepted for publication on 15 April 1999.

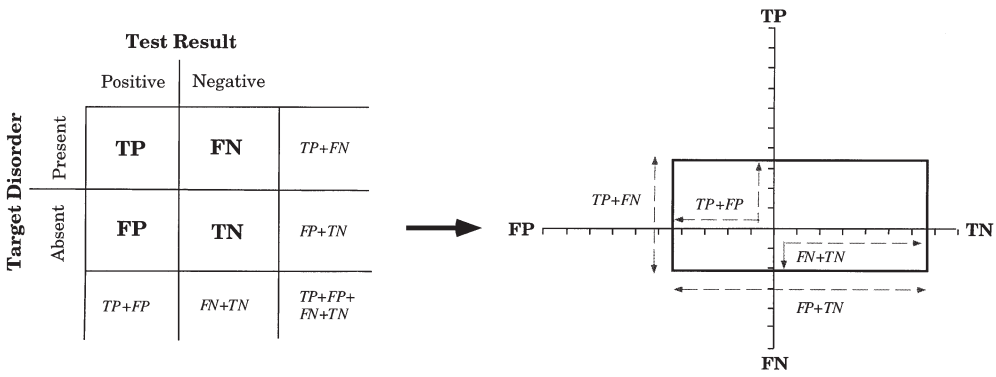

FIGURE 1. The 2 × 2 table and diagram. The 2 × 2 contingency, or "truth," table (left) can be expressed graphically as a diagram (right). A given table defines a unique diagram, and vice versa. By design, each subject's result must fall into one of four categories: TP = true positive, FP = false positive, FN = false negative, and TN = true negative. Each of these categories contributes data to its own cell in the table, or to its own hemiaxis on the diagram. Values are expressed as numbers in the table, but as lengths on the diagram. The sums in italics are the table row and column totals; on the diagram these become sums of lengths along the axes as shown.

The 2 × 2 Table

A test is performed on a patient to determine whether a target disorder is present or absent. Criteria are applied to each test result to categorize it as positive or negative. Each patient is also examined with a reference standard such as surgery or biopsy and categorized as truly normal or truly abnormal. Thus, any given test result falls into one of four possible categories:

True negative (TN): the test result is negative and the disorder is truly absent.
False negative (FN): the test result is negative and the disorder is truly present.
True positive (TP): the test result is positive and the disorder is truly present.
False positive (FP): the test result is positive and the disorder is truly absent.

The results are summarized in a 2 × 2 table (Fig. 1, left). A large body of work extending over many decades has treated such tables in great depth. The data contained therein are described in terms of the contents of individual cells (e.g., number of true positives), proportions (e.g., sensitivity or specificity), and statistical comparison of cell values with the values expected by chance alone (e.g., Pearson's chi-square test). For example, sensitivity of a diagnostic test is defined as the number of true positive results (the upper left cell) divided by the total number of subjects who have the disorders (the top row). Specificity is defined as the number of true negative results (lower right cell) divided by the total number of subjects without the disorder (the lower row). The marginal totals are displayed along the edges of the table as shown. For detailed treatment of the 2 × 2 table the reader is referred to Fleiss [2].

The 2 × 2 Diagram

The 2 × 2 diagram is a method to graphically display all of the data in the 2 × 2 table (Fig. 1, right). It consists of two parts: the coordinate axes and the subject box.

In the coordinate system, each hemiaxis corresponds to one of the four possible result categories (Fig. 2a). The upper vertical hemiaxis of the coordinate system represents the true-positive results (TP), the lower vertical hemiaxis represent false-negative results (FN), the rightward horizontal hemiaxis represents the true-negative results (TN), and the leftward horizontal hemiaxis represents the false-positive results (FP). The dimensions of the hemiaxes are in numbers of patients; values increase positively from the origin in all directions, since there is no such thing as a negative number of patients.

The subject box is constructed by using the reference standard results. The number of truly abnormal patients determines the vertical length of the box, and the number of truly normal patients determines the horizontal length (Fig. 2b). Therefore, the proportions of the box depend on the mixture of normal and abnormal patients in the population sample. The sum of the lengths of any two adjacent sides, termed the half-perimeter, equals the total sample size. A long, low box denotes a low prevalence of the disorder in the population sample. A narrow, tall box denotes a high prevalence.

The position of the box within the coordinate system is determined by the diagnostic test under investigation. A perfect test places the entire box into the right upper quadrant. Note that the box must always include the origin.

Several points need to be emphasized:

1. All abnormal patients lie along the vertical axis: TP + FN.
2. All normal patients lie along the horizontal axis: TN + FP.

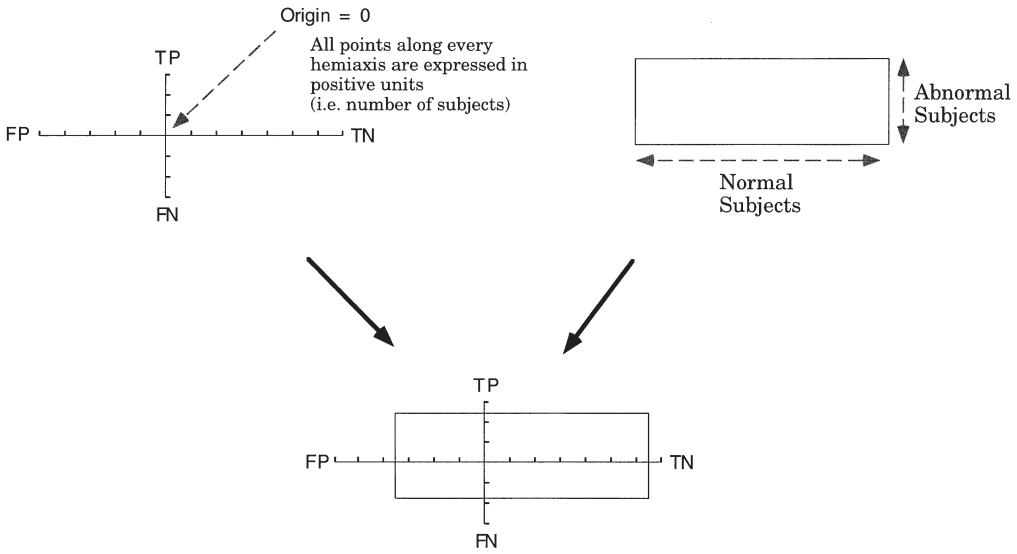

FIGURE 2. The coordinate system and subject box. In the coordinate system (a), the distance from the origin represents the absolute number of patients falling into each of the four result categories. Patients are represented by a rectangular box (b); the number of abnormal patients determines its vertical length and the number of normal patients its horizontal length. The box is superimposed on the coordinate system (c) in a position determined by the diagnostic test under consideration. The origin always falls on or in the box.

3. The length of two adjacent sides of the box, the half perimeter of the box, indicates the total number of patients. TP + FN + TN + FP.
4. The prevalence of the target disorder in the study population is the proportion of the vertical side of the box (abnormal patients) to the half-perimeter of the box (total number of patients). (Fig. 3).
5. Sensitivity is the percentage of the vertical side lying above the horizontal axis: TP / (TP + FN) (Fig. 3).
6. Specificity is the percentage of the horizontal side lying to the right of the vertical axis: TN / (TN + FP) (Fig. 3).

Positive predictive value, negative predictive value, and percent correct answers are represented on the diagram as shown in Figure 3.

Note that the shape of the patient box and its position are independent of one another. The box proportions are determined by the reference standard and the box position

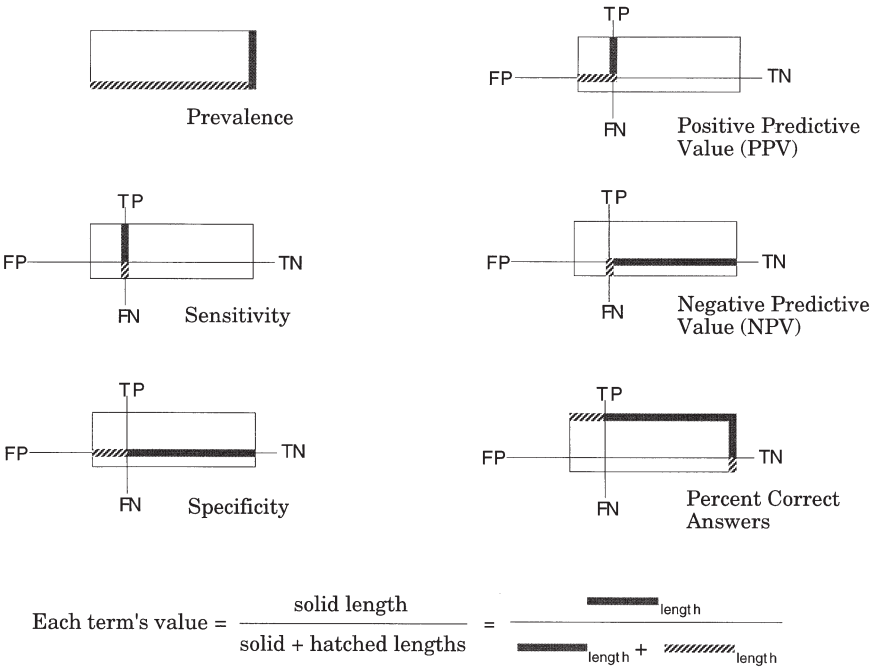

FIGURE 3. Basic descriptors. The 2 × 2 table results are traditionally described by using a number of different terms, each with its corresponding representation on the 2 × 2 diagram. To arrive at a value for each term, the solid length is divided by the sum of the solid and hatched lengths. The particular dimensions used are different for each term, as shown, but the mathematical operation is the same for all terms shown here, as expressed in the equation at the bottom of the figure.

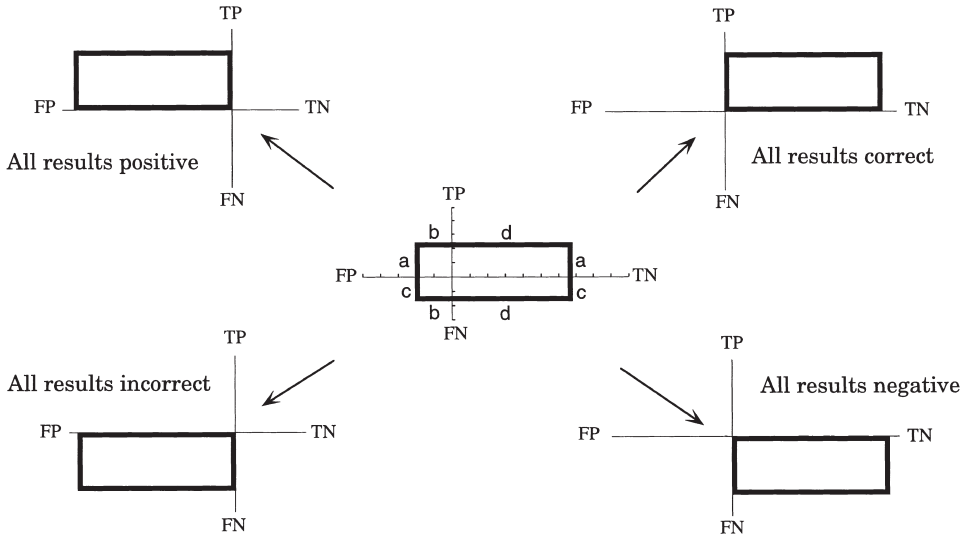

FIGURE 4. Extremes of box position. The position of the subject box is determined by the diagnostic test. The central figure shows the position for a test with mixed results. More extreme behaviors are shown in the other four figures. Each line segment has been labeled. When the entire box is in the right upper quadrant, the test is perfect (top right). The other extremes are all undesirable. The length of the box perimeter in each quadrant has meaning (e.g., the total number of subjects for whom the test result is correct is d + a. The total number of subjects for whom the test result is negative is d + c).

by the diagnostic test. The reader is encouraged to think of the box as potentially freely moveable within the coordinates, but in a particular case restrained by the good or poor behavior of the diagnostic test in question.

Behavior of the 2 × 2 Diagram

Differently sized and proportioned boxes indicate different kinds of patient populations; different positions of the box indicate different test behaviors. Several theoretical examples follow.

THE EXTREMES OF TEST BEHAVIOR. For a perfect test both sensitivity and specificity are 100%, and its box resides entirely in the *right upper* quadrant (Fig. 4). If instead every test result were called positive, the box would reside entirely in the *left upper* quadrant. If every test result was instead called negative, the box would reside in the *right lower* quadrant. In the unlikely event that a test gave 100% wrong answers, sensitivity and specificity would both be 0%, so the box would lie entirely in the *left lower* quadrant. Real tests give a mixture of results so that the box lies somewhere between these extremes.

The right upper quadrant is the only really desirable place for the box. Both costs and benefits are associated with each of the hemiaxes so that when the box resides partly in each quadrant, these costs and benefits sum algebraically. Costs dominate for false-positive and false-negative results; these are opposed by the benefits of true-positive and true-negative results. Therefore, in general the cost–benefit ratio improves as the box moves to the right and/or upwards. Sometimes a diagnostic test criterion, such as a cutoff value, can be selected to find the a position for the box that optimizes this ratio.

THE MEANING OF THE PERCENTAGE OF AREA OF THE BOX IN THE RIGHT UPPER QUADRANT. The percentage of the half-perimeter in the right upper quadrant represents the percentage of patients given correct diagnoses. With a perfect test, all of the patient box will lie in the right upper quadrant. This means the entire half-perimeter will lie there; all of the box's area will lie there too. In real situations, only a portion of the box lies in the right upper quadrant and in these cases the distinction between the behavior of the half-perimeter and of the area becomes important. The percentage of half-perimeter in the RUQ, or "percent correct answers" (PCA), makes no distinction between normal and abnormal patients. With a low prevalence, most of the contribution to this measure is from normal patients, so that PCA is mostly a reflection of the specificity of the test; sensitivity can be quite low without impacting the PCA. With a high prevalence, PCA reflects mostly the sensitivity, and is a poor reflection of the specificity.

In contrast, the percentage area of the box lying in the right upper quadrant seems a better measure of test performance. Percentage area is large when sensitivity and specificity are both large, and small when either one is small. In fact, percent area = sensitivity × specificity; this relationship holds true regardless of the prevalence, or shape of the box. A sensitivity of 90% and a specificity of 80% will place $0.80 \times 0.90 = 0.72$, or 72% of the area of the box into the right upper quadrant. The area percentage is never large

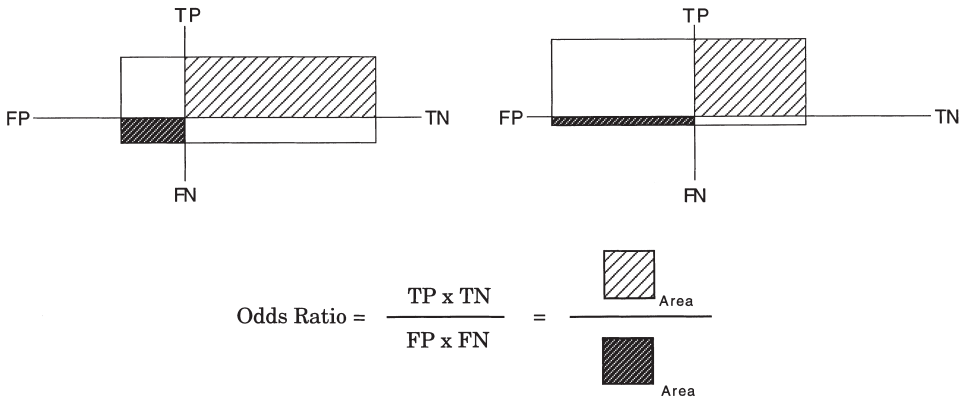

FIGURE 5. Odds ratio. The odds ratio in this context is the odds of a correct test result divided by the odds of an incorrect test result. This can be shown to be the area of the right upper box divided by the area of the left lower box. Both examples shown here have the same odds ratio (7.8), despite disparate test behaviors. As the box moves into the right upper quadrant, the odds ratio increases towards infinity. It is undefined for a box completely above the horizontal axis. Thus, a given odds ratio defines more than one possible position for the box, and so does not describe a unique truth diagram even if the box size and proportions are known.

when either the sensitivity or specificity is poor, regardless of the prevalence.

Another use of the areas is illustrated by the odds ratio, as shown in Figure 5.

THE DIFFERENCE BETWEEN SENSITIVITY AND POSITIVE PREDICTIVE VALUE. Positive predictive value (PPV) should not be confused with sensitivity. Positive predictive value is determined by the proportions of the box in the left upper quadrant (TP/TP + FP); sensitivity is determined by the percentage of the vertical side above the horizontal axis (Fig. 3). Changing the position of the box illustrates that sensitivity and PPV are connected to one another, but are distinct. Sensitivity depends only on the vertical position of the box (i.e., only on the percentage of abnormal patients correctly detected) whereas PPV depends on the vertical position *and* the side-to-side position *and* the box proportions, and therefore on the sensitivity, specificity, and prevalence, respectively.

Analogously, negative predictive value (NPV) should not be confused with specificity. NPV is determined by the proportions of the box in the right lower quadrant (TN/TN + FN), whereas specificity is determined by the percentage of the horizontal side lying to the right of the vertical axis (Fig. 3).

SENSITIVITIES MATCH, BUT NOT SPECIFICITIES. Sensitivities of two tests can be identical even though the 2 × 2 diagrams are very different (Fig. 6). An analogous example could be given for specificity. This underscores that emphasizing only one aspect of test performance can be misleading. To construct a unique diagram, both the box and its position must be known. This requires at least four pieces of information; sensitivity is only one piece.

HIGH PERCENT CORRECT ANSWERS, POOR TEST. A high percent correct answers (percent of agreement) might seem like a straightforward indicator of a good test. However, at extremes of prevalence (which are not unusual in real diagnostic situations) most test results can be correct despite very poor test performance (Fig. 6). Another way of saying this is that a large percentage of the half-perimeter of the box can be in the right upper quadrant (the correct answers quadrant) without much of the *area* of the box being in the right upper quadrant, as illustrated. In this sense the percentage area of the box in the right upper quadrant is a better measure of agreement with the reference standard than is the percentage half-perimeter.

HIGH PREDICTIVE VALUE, POOR TEST. Predictive values are also sensitive to extremes of prevalence. It is possible to have a high positive predictive value but a poor test if the prevalence is high but the specificity is poor. Likewise, it is possible to have a high negative predictive value but a poor test is the prevalence is low and the sensitivity is poor (Fig. 6).

DIFFERENT TESTS COMPARED IN THE SAME POPULATION. Two different tests performed on the same sample can be compared. If costs are ignored, the "better" test gives a higher specificity for a given sensitivity, or in other words places more of the area of the box into the right upper quadrant. In practical situations the costs and benefits of each hemiaxis for each test must be estimated and the absolute net cost of the tests compared. One test may be substantially less specific, but be completely noninvasive and inexpensive, whereas the comparison test might be expensive and invasive, so that the gain in right upper quadrant box area comes at too high a cost.

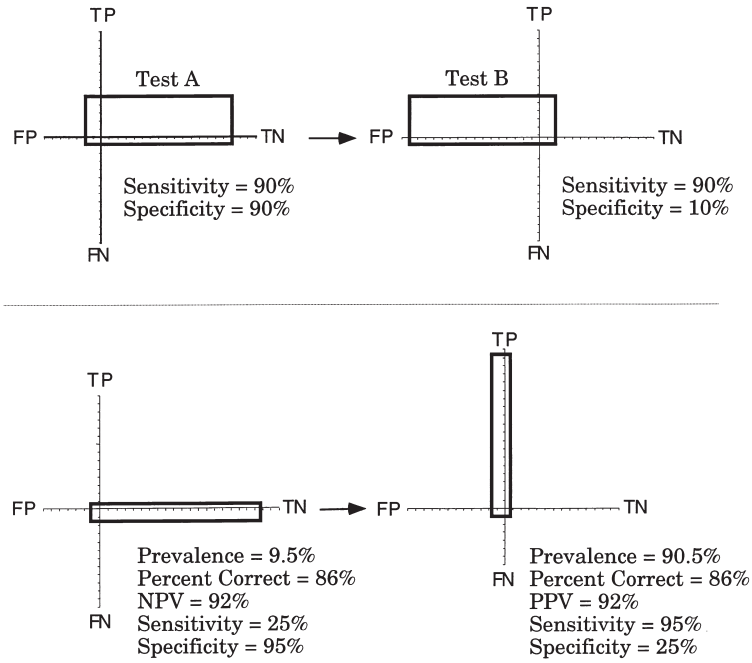

FIGURE 6. Favorable descriptor values, but a poor test. Sensitivity specifies only the percentage of the box lying above the horizontal axis, not its position side to side. In the top pair of diagrams, the sensitivities of tests A and B are equal, but the specificities (also the positive predictive value, negative predictive value, and percent correct answers) are markedly different, with the test on the right showing considerably poorer performance. The bottom diagrams illustrate that predictive value can be high despite poor sensitivity or specificity. The percent correct answers also can be high even though the test is weak in one or more important respects. Care must be taken to avoid the potentially misleading use of isolated descriptive terms.

VERTIFICATION BIAS. After a diagnostic test, usually only some subjects are referred for further verification of their condition by a reference test. Therefore, the spectrum of subjects involved in the verifying test will be different than in the original test. This will distort the sensitivity and specificity calculated for the original test, as illustrated in Figure 7. One remedy is to adequately and randomly subsample the subjects who have negative results and perform the verifying test on them, too. Unfortunately, in medical practice this is rarely feasible, except in the early validation stages of a new test.

Examples From the Literature

Using the 2 × 2 diagram, Figure 8a shows the results of midfield MR imaging for detection of anterior cruciate ligament tears of the knee [5]. Because the reference standard for this study was direct inspection during surgery, the patients shown on the diagram had been stringently preselected from a larger population based on their physician's clinical assessment. This caused the prevalence of a tear in the study population to be quite high, about 80%. If another box were to be drawn to represent all patients presenting to the same physicians with complaints of knee pain (which we cannot actually draw because every patient didn't have a reference test), that box would be bigger and would have significantly different proportions. It would be wide and low, indicating a lower prevalence. This is important because the prevalence affects how well the test result predicts disease. At any given sensitivity and specificity, the predictive value of a positive MRI result will be better in a high prevalence population than in a low prevalence one.

Figure 8b describes the results of a study of electron beam computed tomography (EBCT) for detection of coronary arterial stenosis [6]. Coronary calcification was quantified using EBCT in 491 symptomatic patients and evaluated as a predictor of >50% diameter stenosis somewhere in the coronary arterial tree. The result, using a calcium score of zero (no calcium, *solid box*), is compared to the result using a calcium score of 100 (*dashed box*) in the same set of patients. The patient population was identical in the two instances, but the criterion for calling the test abnormal was different. As the "cutoff" for the calcium score was raised, the box moved towards the right and downwards, following the path indicated by the *curved line*. As the box moved toward the right and downward, specificity increased and sensitivity decreased. The positive predictive value (denoted by the left upper quadrant box) of the selected calcium cutoff score increased and its negative predictive value (right lower quadrant box) decreased (refer to Fig. 3). Each hemiaxis has its own costs; false positives mean more catheterizations, whereas false negatives mean disease will be

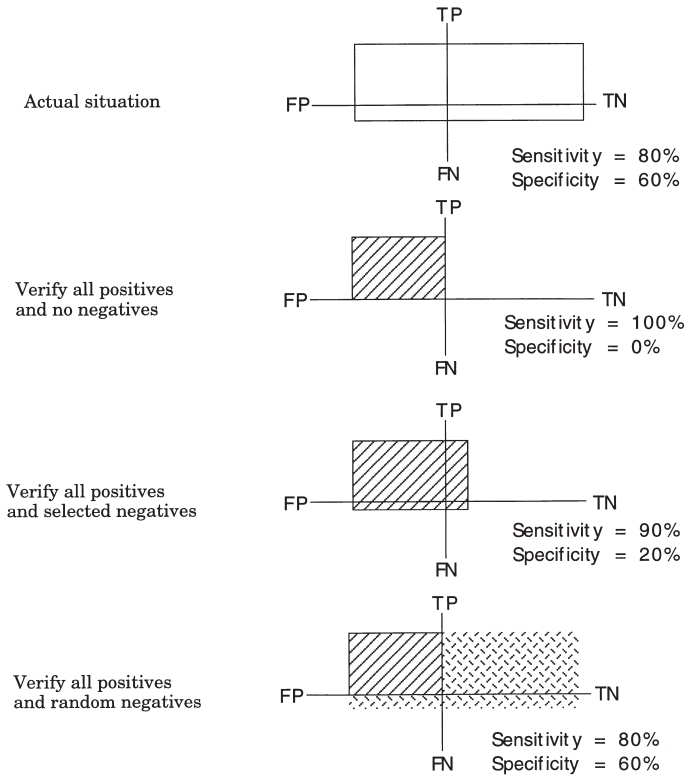

FIGURE 7. Verification bias. The apparent sensitivity and specificity of a test depend on which subjects are referred for verification by the reference standard. In this example, the actual but unknown underlying situation is displayed in the top row. If only patients with positive test results are sent for verification, the diagram we construct would be as in the second row, with sensitivity and specificity grossly misestimated. Also verifying high clinical-suspicion negative cases still gives inaccurate estimates (third row). A better approach is to also send a random subsample of the negative cases for verification. An estimate of the prevalence can then be made, and from that a good estimate of the true box dimensions and position. In ordinary medical practice, random testing of negative cases is usually not a practical option, thus the problem of verification bias persists.

missed. The challenge is to select a cutoff value which gives "adequate" sensitivity without losing "too much" specificity. This is always a problem of balancing the relative costs and benefits of the four hemiaxes.

Figure 8c represents a study of fluoroscopy for the detection of coronary arterial stenosis in asymptomatic aviators at high risk for atherosclerosis [7]. This study tested the ability of fluoroscopically detected calcium to predict the presence of coronary arterial stenosis on subsequent angiography. For a more severe stenosis ($\geq$50%), the patient box was flatter and wider, because fewer patients had this degree of disease. For $\geq$10% stenosis, the prevalence was about double that of the higher grade stenosis, so the box was narrower and taller. In other words, altering the question asked by the test changed the box proportions by changing the assignment of patients between the normal and abnormal groups.

Finally, in a large study of screening mammography [8], the prevalence of breast cancer was about 0.5%, resulting in a very low, very wide box (Fig. 8d). This is typical for a screening test and is in striking contrast to the other examples. Such tests have extreme demands made on their sensitivity and specificity. If the goal is to maximize the area of the box in the right upper quadrant, it is apparent that sensitivity must be high. Every correctly classified abnormal case pushes the box upwards and adds markedly to the percent area in the right upper quadrant. At the same time, even if the specificity is reasonably good, the positive predictive value (reflected by the proportions of the left upper quadrant box) will be relatively low, because normal patients there (false positives) outnumber abnormal patients (true positives) by a large margin. As long as the number of false-negative diagnoses (missed cancers) is kept very low, the cost is dominated by false positives. False positives cost more mammographic views, ultrasound exams, and, in a minority of patients, biopsies. The consensus is that this is an acceptable price to pay, because the alternative would be to move the box to the right and downward, improving the specificity (fewer FPs) but increasing the number of missed cancers (more FNs).

DISCUSSION

The 2 $\times$ 2 table is widely used for the presentation and analysis of experimental results. In this report it is trans-

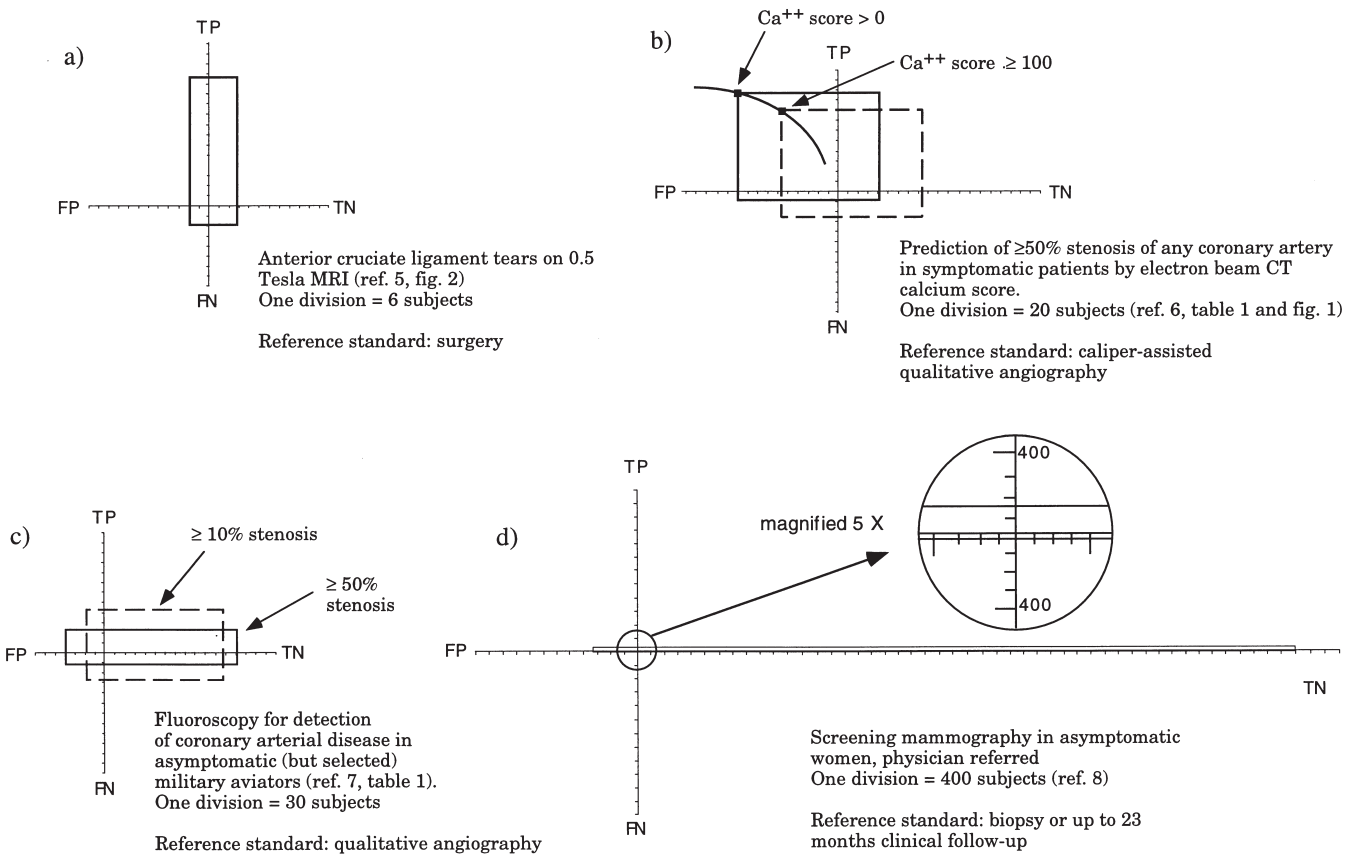

FIGURE 8. Examples from the literature. 2 × 2 diagrams from four individual reports in the medical literature are shown (see text for discussion). The range of prevalence is wide. Note that each diagram is derived from a single publication; other published studies on the same problem may yield different diagrams.

formed into a diagram that has several advantages over the table.

The diagram improves the clarity of presentation because it reduces the number of ways in which the same results can be described, and shows a larger picture rather than emphasizing only one part. It promotes a sense of proportion, and a sense of how the data are interrelated. There is no redundancy: the 2 × 2 table has four cells, and the 2 × 2 diagram has four corresponding dimensions. The temptation to reduce the description of test behavior to fewer dimensions, (i.e., to one or two descriptive numbers) should be resisted, because this cannot be done without sacrificing information. For example, sensitivity and specificity together do not define a unique 2 × 2 diagram, and therefore are not a complete description of the behavior of a test.

ROC Curves

Briefly, the ROC curve data are used to construct a curve in the left upper quadrant, to which the left upper corner of the box is "attached." This is illustrated but not discussed in Figure 8b; we shall call it the test trajectory. The box is constrained to move only where the trajectory allows it to move. Though this trajectory is derived from the ROC curve, *it is not the ROC curve itself*. It is critically important to note this distinction. At a prevalence of 50% the *shape* of the trajectory is precisely that of the ROC curve (only flipped left to right), but the units are completely different. In ROC analysis, the units are dimensionless proportions (i.e., the true-positive rate vs. the false-positive rate). On the 2 × 2 diagram, the dimensions are in absolute number of subjects (i.e., the true-positive subjects vs. the false-positive subjects). There are no rates involved. As the prevalence deviates from 50%, the curve is stretched or compressed along the hemiaxes. How to do this correctly and the implications for interpretation, such as the selection of a cut point, are beyond the scope of the present work.

Why These Axes?

The spatial arrangement of the hemiaxes used here is obviously not the only possible one. Indeed, exactly 24 possible schemes exist and fall into three groups of 8 each. Within each group, the various arrangements are simply flipped and/or rotated versions of one another, so they are topologically identical. Therefore, there are three real choices:

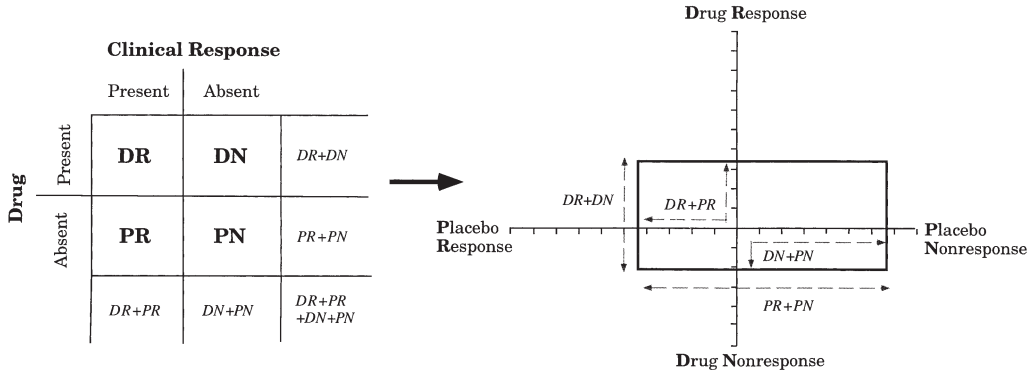

FIGURE 9. The 2 × 2 diagram for a drug trial. This version of the diagram is analogous to the version for diagnostic tests, but the labels have been changed appropriately. A given 2 × 2 table defines a unique diagram; the cell values in the table map to the diagram as shown. The vertical side of the box now represents the subjects who in fact were given the drug, and the horizontal side those given the placebo. Various positions of the box define the ways in which the drug effects differ from placebo effects. Again, the most desirable position for the box is in the right upper quadrant. For example, if the box were entirely in the left upper quadrant, this would mean every subject given the drug responded but so did every patient given the placebo, so that the expense and possible side effects of the drug have not been justified. If instead the box were over the origin, this means half the subjects given the drug responded, but so did half the subjects given the placebo. Again the superiority of the drug has not been established. DR = drug response, PR = placebo response, PN = placebo nonresponse, DN = drug nonresponse.

a. all normal subjects lie along one axis and all the abnormal subjects along the other (the present scheme),
b. all true results lie along one axis and all false results along the other,
c. all positive results lie along one axis and all negative results along the other.

Arrangement "a" has the advantage that it separates the data into two clearly delineated parts: the box size and shape are determined solely by attributes of the subjects (as established by the reference test), and the box position is determined solely by the behavior of diagnostic test under study. The other two arrangements mix these factors together, and lead to a more complicated picture. For example, as test performance changes, instead of a simple translation in position, the shape of the box would change. Incorporating receiver operating characteristic curve information into the diagram would be made more difficult; this is discussed more below.

Applications Beyond Diagnostic Testing

The 2 × 2 diagram can be applied to any 2 × 2 table, not just those describing diagnostic tests. For example, in a therapeutic trial, one operator is the presence of the drug vs. placebo, and the other is clinical improvement vs. no improvement. Since whether the drug was in fact given or not is known, the first operator replaces the "gold standard" in the diagnostic test application, and the effect vs. no effect results replace the diagnostic test results. The axes of the diagram are appropriately renamed, and the diagram then appears as in Figure 9.

Analogous diagrams can be drawn for other situations, for example, to compare the performance of two human observers, or to two test combinations. The derivation and limitations of the concept of expected values is an example of a more subtle application of the diagram that we will develop in a later paper. Cost–benefit analysis also might be more clearly described using the diagram. These alternative aspects cannot be treated in detail here.

CONCLUSION

The 2 × 2 diagram lends itself to mental manipulation in a way that the 2 × 2 table does not. Visual thinking can aid the interpretation of complicated phenomena [3,4], and has been applied at times in statistics [9–11].

It would be interesting to investigate quantitatively whether students in fact find this graphical scheme easier to understand and remember than a list of terms. If so, editors might encourage use of the diagrams as simple, clear displays of experimental results, and physicians might use them to analyze old studies, to plan new studies, and ultimately to make more intelligent decisions for patients.

The author thanks Dr. Laura J. Horvath for her review of the manuscript and helpful discussions.

References

1. Simpson JM. Teaching statistics to non-specialists. **Stat Med** 1995; 14: 199–208.
2. Fleiss JL. **Statistical Methods for Rates and Proportions.** 2nd Ed. New York: Wiley; 1981.
3. Polya G. **How to Solve it: A New Aspect of Mathematical Method.** 2nd Ed. Princeton, NJ: Princeton University Press; 1957.
4. Feynman RP. **The Character of Physical Law.** Cambridge, MA: MIT Press; 1965.

5. Vellet AD, Lee DH, Munk PL, *et al*. Anterior cruciate ligament tear: Prospective evaluation of diagnostic accuracy of middle- and high-field-strength MR imaging at 1.5 and 0.5 T. **Radiology** 1995; 197: 826–830.
6. Detrano R, Hsiai T, Wang S, *et al*. Prognostic value of coronary calcification and angiographic stenoses in patients undergoing coronary angiography. **JACC** 1996; 27: 285–290.
7. Loecker TH, Schwartz RS, Cotta CW, Hickman JR. Fluoroscopic coronary artery calcification and associated coronary disease in asymptomatic young men. **JACC** 1992; 19: 1167–1172.
8. Sickles EA, Ominsky SH, Sollitto RA, Galvin HB, Monticciolo DL. Medical audit of a rapid-throughput mammography screening practice: Methodology and results of 17,114 examinations. **Radiology** 1990; 175: 323–327.
9. Brismar J, Jacobsson B. Definition of terms used to judge the efficacy of diagnostic tests: A graphic approach. **AJR** 1990; 155: 621–623.
10. Singer PA, Feinstein AR. Graphical display of categorical data. **J Clin Epidemiol** 1993; 46: 231–236.
11. Feinstein AR, Kwoh CK. A box-graph method for illustrating relative-size relationships in a 2 × 2 table. **Int J Epidemiol** 1988; 17: 222–224.