Chapter 3
DIAGNOSTIC AND PROGNOSTIC CLASSIFICATION

© 2001 Dennis A. Noe

DIAGNOSTIC STUDY PERFORMANCE

A diagnostic laboratory study is one that is designed, or has been discovered, to improve the clinician's ability to discriminate between persons suffering from a disorder or condition of interest and persons free from the disorder or condition. The degree to which a study accomplishes this discrimination is referred to as its diagnostic performance.

Sensitivity, specificity, and ROC curves

The fundamental measures of the diagnostic performance of a laboratory study are its sensitivity and specificity. Sensitivity is the frequency with which a study indicates the correct diagnosis in persons with the disease. Specificity is the frequency with which a study indicates the correct diagnosis in individuals who are disease-free.

As an example, the data obtained in a clinical investigation concerned with the laboratory diagnosis of iron deficiency in infants (Dallman *et al.* 1981) can be used to quantify the performance of the study, transferrin saturation (the ratio of plasma iron concentration to plasma iron-binding capacity). Transferrin saturation was determined in capillary blood specimens from 165 1-year-olds who were suspected of having iron-deficiency anemia. Infants were classified as iron deficient if the transferrin saturation was less than 10% and as iron replete if the transferrin saturation was greater than 10%. The study classifications, categorized according to the final diagnostic classification, are presented in Table 3.1. The numerical entries indicate the number of study subjects in each category. The sensitivity of the study is calculated as the frequency of correct diagnosis in the iron-deficient infants. In this case, the frequency is 29 divided by 55 which equals 0.53. A little better than one-half of the iron-deficient infants are properly identified. The specificity is the frequency of correct diagnosis in the iron-replete subjects; here, it is 82 divided by 110 which equals 0.75. Three quarters of the iron-replete infants are correctly identified. A table of classification categories, as used in the example, can be constructed for any diagnostic study (Table 3.2). The designations true or false and positive and negative are usually assigned to the categories as shown. From the table,

$$sensitivity = \frac{true\ positives}{true\ disease}$$

and

$$specificity = \frac{true\ negatives}{true\ disease\ free}$$

Because a study's specificity is determined by the frequency distribution of results in a stable reference population, it will remain constant. There will be some variability in its measurement because the composition of the sample of subjects will vary by chance. However, as long as the subjects are chosen at random from the same reference population, estimates of the study's specificity will cluster around the value that would be found were the entire reference population to be studied. Constancy of estimation is not true for sensitivity. The sensitivity of a diagnostic study is usually greater in individuals with more advanced or severe forms of a disease. In the case of iron deficiency, as the condition persists, and the iron deficit deepens, all of the diagnostic studies used to identify iron deficiency, transferrin saturation included, show increased sensitivity.

Table 3.1
Classification Categories for Transferrin Saturation

Classification Using Transferrin Saturation	Final Diagnostic Classification	
	Iron-replete	Iron-deficient
Iron-replete	82	26
Iron-deficient	28	29
Total	110	55

Table 3.2
Classification Categories for a Diagnostic Study

Classification Using Study Result	Final Diagnostic Classification	
	Disease-free	Disease
Disease free	true negative	false negative
Disease	false positive	true positive
Total	true disease free	true disease

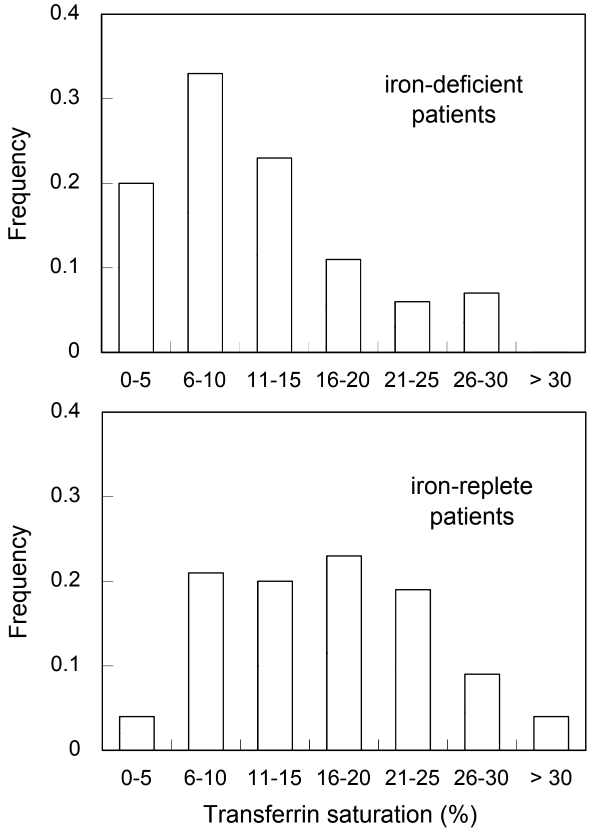

Figure 3.1 Reference frequency histograms for transferrin saturation.

ROC curves. The diagnostic performance of a study depends upon the choice of the critical value. This is the study result used to separate the diagnostic classes. In the example from Dallman *et al.* (1981), the critical value of transferrin saturation that was used was 10%. Selection of a different critical value would have resulted in different values for sensitivity and specificity. The set of sensitivity and specificity pairs that are generated by considering every possible critical value for a laboratory study constitute the performance characteristic function. This function completely defines the performance of the study when applied to a given pair of reference frequency distributions. Consequently, it is the most informative way to record the findings from an investigation of the study's performance (Henderson 1993, Zweig and Campbell 1993, Beck and Shultz 1986). Using it, one can identify the critical value that generates a desired pairing of sensitivity and specificity.

A performance characteristic function for transferrin saturation can be obtained by again referring to the data reported by Dallman *et al.* (1981). The authors include in their article histograms indicating the distribution of study values in the two reference populations. Their histograms are recast as frequency distributions in Figure 3.1. To construct the performance characteristic function, first select an extreme study value (here 0% is a likely choice) and calculate the sensitivity and specificity that would result were this the critical value. No iron-deficient subject has a transferrin saturation less than 0% so the sensitivity is 0. All the iron-replete subjects have saturations greater than 0% so the specificity is 1.0. Then, repeat the calculations using the next permissible value of the study, 5%, as the critical value. Since 20 percent of the iron-deficient infants have transferrin saturations below 5%, the sensitivity is 0.2. Of the iron-replete subjects, 4 percent have saturations less than 5% so only 96 percent of these subjects are correctly identified. Thus, the specificity is 0.96. This procedure is repeated until all the possible critical values have been considered. The results for these data are shown in Table 3.3. Performance characteristic functions are often presented in their graphic form which, for historical reasons, are called receiver operating characteristic or ROC curves. Figure 3.2 (squares) shows the ROC curve for transferrin saturation.

Table 3.3
Performance Characteristic Function for Transferrin Saturation

Critical Value	Performance Characteristic	
	Sensitivity	Specificity
0%	0.00	1.00
5%	0.20	0.96
10%	0.53	0.75
15%	0.76	0.55
20%	0.87	0.32
25%	0.93	0.13
30%	1.00	0.04

As discussed in Chapter 1, the distribution of study results in reference populations can be represented by frequency distribution models. Such modeling yields two significant benefits in the construction of ROC curves. First, irregularities in the empirical data attributable to measurement variability are smoothed out and, in turn, so are the derived values of sensitivity and specificity. Second, gaps in the data corresponding to study values that were not recorded among the reference subjects can be filled in. Indeed, the use of continuous distribution models allows for the construction

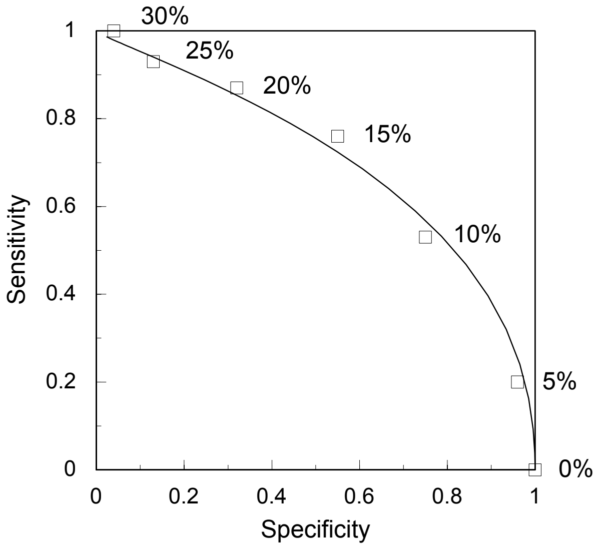

Figure 3.2 ROC curves for transferrin saturation. The squares represent the points constructed from the observed frequency data (Figure 3.1). The continuous line is the curve constructed from the lognormal frequency distribution models of the data.

of continuous ROC curves. Figure 3.2 (line) shows the ROC curve that results from modeling the data of Dallman *et al.* (1981) with lognormal frequency distributions. Another example of ROC curve construction using lognormal modeling can be found in Krieg *et al.* (1989). ROC curve construction based on kernel density smoothing is discussed by Zou *et al.* (1997).

Variability in study performance. It is often mistakenly assumed that one of the reasons that ROC curves are considered such a useful way to describe study performance is because the sensitivity and specificity values that make up the curves are invariant features of a laboratory study. In fact, sensitivity and specificity can vary and even vary widely. For instance, study performance can differ from laboratory to laboratory because of differences in analytical methodology and in staff and equipment quality. Also, study performance can vary from clinical population to clinical population because of differences in the spectrum of disease in the different populations. Additionally, study performance can vary over time either as a result of changes in the methods used to perform the study and or consequent to alterations in the spectrum of disease over time.

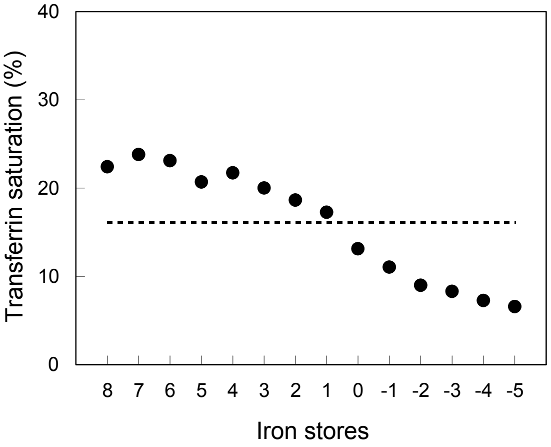

Figure 3.3 Changes in transferrin saturation during the development of iron deficiency caused by serial phlebotomy as studied in normal volunteers. The dashed line indicates the lower limit of the reference range.

Clearly, spectrum of disease, which represents the range in clinical expression and severity of disease in a clinical population, is an especially important determinant of study performance. This is so because it is rare for a study to yield the same result regardless of the severity of a disease or the pathobiologic stage of a disorder. Examples of laboratory studies that are essentially invariant are the chromosome and DNA studies used to diagnose certain genetic diseases. However, for the vast majority of laboratory studies, study results are affected by the level of activity of or severity of the disease and the degree to which the disease has compromised normal body function. Typically, the more severe the disease or the greater the level of dysfunction, the greater the displacement of the study value. For example, in iron deficiency, the larger the iron deficit, the smaller the value of transferrin saturation. This is illustrated in Figure 3.3 for iron deficiency resulting from serial phlebotomy. Transferrin saturation begins to decline once the body iron stores fall below about 4 mg/kg and decreases progressively as the iron deficit increases (Skikne *et al.* 1990).

Because the location and width of the distribution of study results in diseased individuals varies with disease severity and activity, the location and width of the aggregate distribution of study results in a clinical population will depend upon the spectrum of disease in the population. For instance, in a clinical population consisting mostly of individuals who have early or mild forms of a disease, individual study results will tend to be at most modestly abnormal; thus, the aggregate distribution of study results will usually not be far removed from the distribution of results in the disease-free members of the population. This is the typical situation when screening for a disease among asymptomatic individuals and explains the considerable challenge in finding sensitive screening tests that are also specific. On the other extreme, if a clinical population consists primarily of patients with advanced or severe disease, almost all individual study results will be very abnormal causing the aggregate distribution of study results to be widely separated from that of the disease-free members of the population. At any stipulated critical value, the sensitivity of the study will be much greater in the second population than in the first. In addition, the specificity will usually be less because the disease-free individuals in the second population almost always have other medical conditions that explain their presence in this clinical population—conditions that will tend to broaden the distribution of study results and thereby lower the specificity.

Other measures of diagnostic performance

Two alternative measures of diagnostic performance need to be mentioned. Both incorporate the effect that the prevalence of a condition, i.e. the proportion of persons in the clinical population who have the condition, will have upon the classification accuracy of a diagnostic study. The first measure is diagnostic efficiency, defined as the overall frequency of correct diagnostic classifications when a study is applied in a clinical setting. Thus, from Table 3.2,

$$\textit{efficiency} = \frac{\textit{true negatives} + \textit{true positives}}{\textit{true disease free} + \textit{true disease}}$$

The dependence of efficiency upon disease prevalence is indicated by redefining it in terms of sensitivity and specificity. The number of true positives equals the sensitivity of the study times the number of tested individuals who have the disease, the number of tested individuals who have the disease equals the prevalence times the number of individuals individuals, and the number of true negatives equals the specificity of the study times the number of tested individuals times one minus the prevalence. Thus,

$$\textit{efficiency} = \textit{prevalence} \cdot \textit{sens} + (1\text{-}\textit{prevalence})\ \textit{spec}$$

where *sens* stands for study sensitivity and *spec* stands for study specificity. This formula reveals the validity of a number of intuitive insights regarding the behavior of diagnostic efficiency. First, when the disease prevalence is low, the efficiency of a study is determined largely by its specificity and second, when the disease prevalence is high, the efficiency of a study depends mostly upon its sensitivity.

The other alternative measure of diagnostic performance is the predictive value of a study result. Predictive value is the frequency with which a classification study is correct in a given clinical setting,

$$\textit{predictive value of a positive result} = \frac{\textit{true positives}}{\textit{true positives} + \textit{false positives}} = \frac{\textit{prevalence} \cdot \textit{sens}}{\textit{prevalence} \cdot \textit{sens} + (1 - \textit{prevalence})(1 - \textit{spec})}$$

$$\textit{predictive value of a negative result} = \frac{\textit{true negatives}}{\textit{true negatives} + \textit{false negatives}} = \frac{(1 - \textit{prevalence})\,\textit{spec}}{(1 - \textit{prevalence})\,\textit{spec} + \textit{prevalence}\,(1 - \textit{sens})}$$

These definitions, as well as good sense, demonstrate that the predictive value of a positive study result increases with increasing prevalence and with increasing study sensitivity and specificity. When the prevalence is low, the probability that a positive study result is correct is small, unless the study specificity is nearly one. This is an extremely important point when a study is being used to identify individuals with rare disorders. The predictive value of a negative study result increases with decreasing prevalence and with increasing study sensitivity and specificity. When the prevalence is low, the frequency of correct negative study results is high even when the diagnostic performance of the study is poor.

Repeating and combining studies

The performance of a diagnostic study can be altered by repeating the study or by using the study in combination with one or more other diagnostic studies. The performance that results from such multiple testing depends largely upon two new considerations: the positivity rule used to make the ultimate diagnostic classifications and the classification correlation between repeated tests or among combinations of tests.

Repeat testing. The two most frequently used positivity rules for repeat testing are illustrated in the

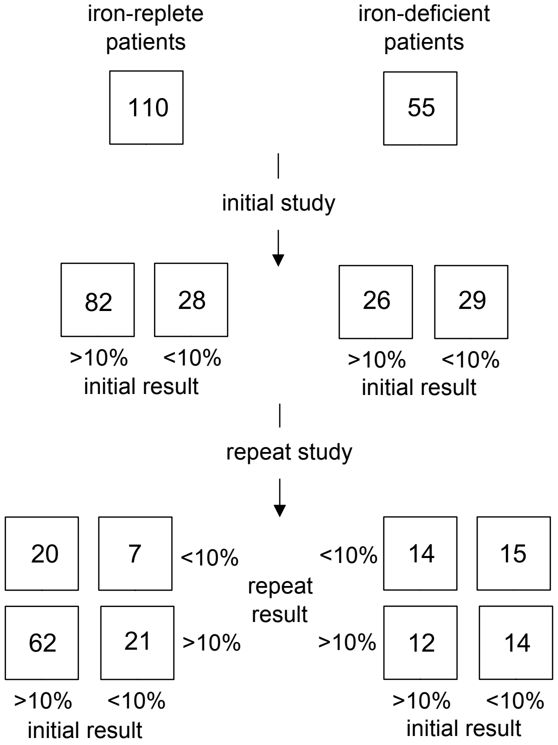

Figure 3.4 Diagnostic performance of repeat testing using transferrin saturation, assuming a repeat classification correlation of zero.

following example. The single test diagnostic performance of transferrin saturation at the critical value of 10% saturation recommended by Dallman *et al.* (1981) consists of a sensitivity of 0.53 and a specificity of 0.75. What happens if the study is repeated in the same patients? If there is no classification correlation between the initial and repeat study—that is, if the performance characteristics of the study in the diagnostic subgroups formed by the initial application of the study are the same as they are in the population as a whole—the repeat study results will be as shown in Figure 3.4. For the 82 iron-replete patients initially classified as negative, 62 will have negative results with the repeat study but 20 will have positive results. For the 29 iron-deficient patients with positive results from the first test, 15 will have a positive repeat study result and 14 a negative result. And so on for the other categories.

One way to categorize these patients clinically is to decide that the test series is positive if either the initial or repeat study result is positive. This positivity rule is designated "believe-the-positive." The diagnostic performance resulting from this rule is indicated at the top of Figure 3.5. Forty-three iron-deficient patients will have at least one positive test result, so the sensitivity of the series is 0.78 (43 divided by 55). Of the patients who are iron replete, 62 will have both results negative and will therefore be appropriately categorized. The specificity is therefore 0.56 (62 divided by 110). Another way to categorize the patients is to consider the test series positive only if both study results are positive. This positivity rule is referred to as "believe-the-negative." Its performance, as demonstrated at the bottom of Figure 3.5 is a sensitivity of 0.27 (15 divided by 55) and a specificity of 0.94 (103 divided by 110).

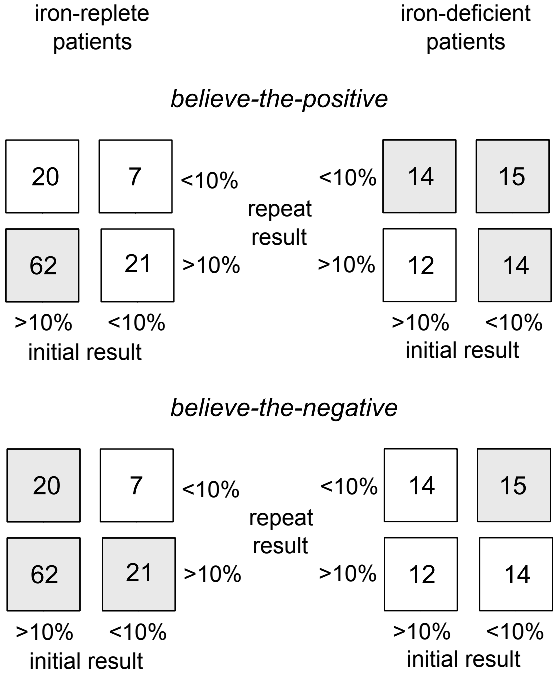

Figure 3.5 Application of different positivity rules to repeat testing using transferrin saturation, assuming a repeat classification correlation of zero.

As shown in the example, the "believe-the-positive" positivity rule leads to an increased sensitivity and a decreased specificity compared to a single application of the study. This is because individuals who have the disorder have two opportunities to be detected while those who do not have the disorder have two chances to be misclassified. In contrast, use of the "believe-the-negative" positivity rule results in decreased sensitivity but increased specificity. With this rule patients who have the disorder have two opportunities to be misclassified while those who do not have the disorder have two chances to be correctly identified. With additional repetitions of the study, the diagnostic performance of the series is removed further still from that of the single study.

All of the foregoing calculations have been based upon the condition that there is no classification correlation between repeat studies. In reality, classification correlation usually exists (Politser 1982). Because intraindividual variability in study results is usually fairly small, a repeat study in an individual is likely to yield a result close to a previous result and, therefore, to give a similar diagnostic classification, even if it is a misclassification. When classification correlation is present, the actual diagnostic performance of a test series will differ from that computed under the assumption that the classification correlation is zero. For the "believe-the-positive" rule, the sensitivity will be greater than predicated and the specificity will be less; for the "believe-the-negative" rule, the sensitivity will be greater than predicted and the specificity will be less.

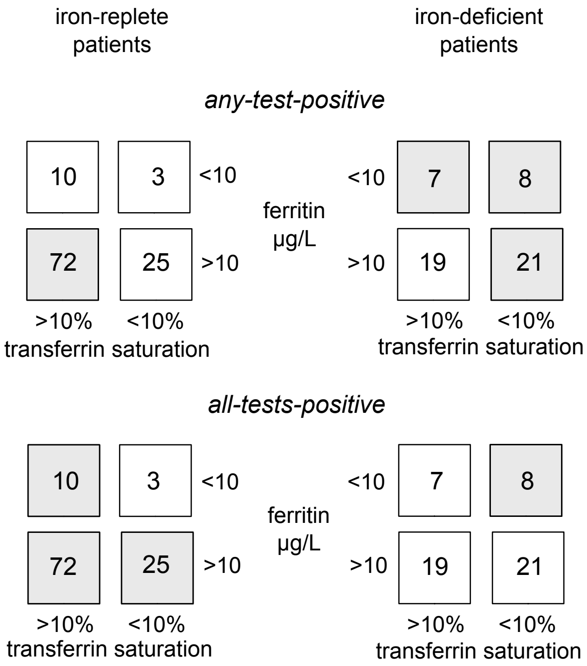

Figure 3.6 Application of different positivity rules to combination testing using transferrin saturation (critical value, 10%) and ferritin concentration (critical value, 10 µg/L). A classification correlation of zero is assumed.

Combination testing. Two popular positivity rules for combination testing are analogous to those used for repeat testing. The "any-test-positive" rule, for which the test combination is considered positive if any of the constituent study results are positive, is the same as the "believe-the-positive" rule. The "all-tests-positive" positivity rule is equivalent to the "believe-the-negative" rule for repeat testing. And, just as for repeat testing, the first rule leads to an increased sensitivity and decreased specificity compared to the individual studies and the second rule results in decreased sensitivity but increased specificity (Cebul *et al.* 1982). This is shown in Figure 3.6 for the combination of ferritin and transferrin saturation as comarkers of iron deficiency. Using the data of Dallman *et al.* (1981), at a critical value of 10%, transferrin saturation has a sensitivity of 0.53 and a specificity of 0.75 and, at a critical value of 10 μg/L, plasma ferritin concentration has a sensitivity of 0.28 and a specificity of 0.87. If there is no classification correlation between the studies, using the any-test-positive rule gives the test combination a sensitivity of 0.65 and a specificity of 0.65; using the all-tests-positive rule yields a sensitivity of 0.15 and a specificity of 0.97. The any-test-positive rule is frequently used in multiphasic health screening. A multiphasic health screen is a combination of 12, 18, 24 and sometimes more laboratory studies performed upon a single blood specimen for the purpose of detecting clinically silent disease in asymptomatic individuals. The presence of any study result outside of its reference interval is supposed to identify persons who should be evaluated further for subclinical disease. The diagnostic specificity of such testing is low, however. Very low! For one laboratory study with a reference interval based upon a specificity of 0.95, the probability that a healthy person will have a test result outside of the reference interval is 0.05. For a combination of *j* uncorrelated laboratory studies, each of which has a reference interval chosen to give a specificity of 0.95, the chance of one or more positive results in a healthy individual is

$$1 - (0.95)^j$$

This means that for multiphasic screens of 12, 18, and 24 tests the probability of at least one positive result is 0.46, 0.60, and 0.71, respectively, for an individual who is, in fact, absolutely healthy. In clinical practice most physicians deal with this problem by ignoring positive results that represent only small deviations outside of the reference range. They do respond to larger deviations. This is appropriate because larger deviations are associated with greater specificities among the individual tests and, therefore, with a reasonable level of overall specificity for the multiphasic screen.

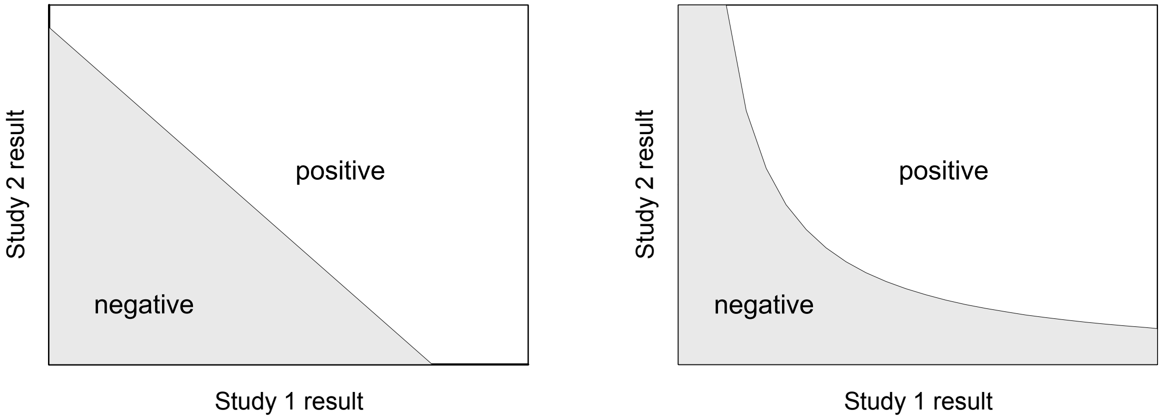

Figure 3.7 Discriminant function positivity rules for the interpretation of a two-test combination. Left graph, diagnostic spaces defined by a linear discriminant function; right graph, diagnostic spaces defined by a quadratic discriminant function.

Multivariate positivity rules. The positivity rules for combination testing just discussed rely upon critical values derived from the univariate (one test) result frequency distributions for the reference populations. In the setting of combination testing it is also possible, and often desirable, to define positivity rules which arise from a consideration of the multivariate (multiple test) result frequency distributions that arise from the application of the test combination to the respective reference populations. These are called multivariate positivity rules.

Discriminant functions. Positivity rules based on discriminant functions separate diagnostically positive test result combinations from diagnostically negative combinations by defining a curve (two tests) or surface (multiple tests) that divides the space of test result combinations into the two diagnostic regions (Figure 3.7). When a linear discriminant function is used, the diagnostic regions are separated by a straight line or a plane. The slope of the line, or the orientation of the plane, is selected by statistical rules to yield maximum separation of the result frequency distributions of the diagnostic classes and, therefore, maximum diagnostic discrimination (Solberg 1978, Strike 1996). The location of the line or plane, which is specified by the value of an axis intercept, establishes the separation of the diagnostic categories and thereby serves as the critical value determining the performance characteristics of the study combination. For combinations of two tests, the diagnostic classification of individuals can be accomplished graphically, by plotting their test results, or algebraically, by calculating a discriminant score,

$$\textit{discriminant score} = \textit{b1 result 1} + \textit{b2 result 2}$$

and comparing the score to the stipulated critical score value. For more than two test results, the algebraic approach is used.

Although linear discriminant function positivity rules are common in the medical literature, the valid application of this technique is limited by its statistical constraints. In particular, it is necessary that, in the clinical population of interest, the variances of the individual test results as well as the covariances of all test pairs must be the same for individuals with the disease and those who are disease-free. This is a criterion that is rarely satisfied. Fortunately, quadratic discriminant analysis can be used when the individual test variances and the test pair covariances are unequal; however, the test result distributions must be multivariate normal. As shown in Figure 3.7, quadratic discriminant functions yield maximum separation of the result frequency distributions of the diagnostic classes using a curved line or a curved surface.

Diagnostic ratios. Diagnostic ratios are a multivariate approach to result interpretation when only two laboratory studies are concerned. Positivity rules based on diagnostic ratios separate the diagnostic space into two regions using a straight line that passes through the origin (Figure 3.8). Patients are classified by plotting their test results or by calculating the value of the diagnostic ratio,

$$\textit{diagnostic ratio} = \frac{\textit{result 1}}{\textit{result 2}}$$

and comparing it to the critical value for the ratio.

Ratios have proved most useful when the values of the analytes change in opposite directions in response to disease. The ratio of the two magnifies the changes and thereby increases the diagnostic

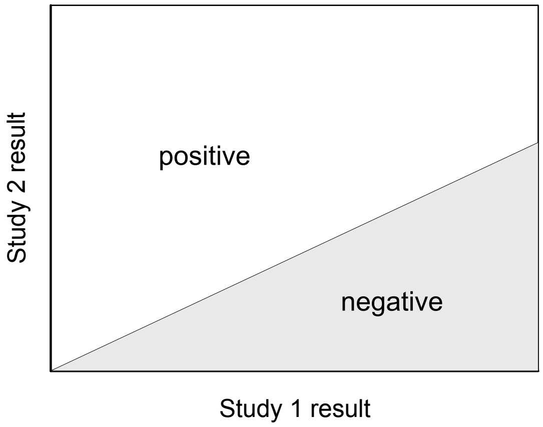

Figure 3.8 Diagnostic ratio positivity rule for the interpretation of a two-test combination.

resolution. Transferrin saturation, which is the ratio of plasma iron concentration to plasma total iron-binding capacity, is a more reliable marker of iron deficiency than either measure taken separately because as the plasma iron concentration declines with iron deficiency, the total iron-binding capacity increases. Consequently, the ratio of the two diminishes markedly.

Ordinarily, diagnostic ratios are not the best way to achieve maximum diagnostic discrimination because the line separating the diagnostic classes must necessarily pass through the origin. In contrast, the line of separation defined by a linear discriminant function is free to have a nonzero intercept and, therefore, has the positional flexibility to optimally separate the diagnostic classes. Thus, linear discriminant functions make for better positivity rules than diagnostic ratios.

If the result frequency distributions of the diagnostic classes are bivariate lognormal, the discriminant function, which is linear in the log-transformed diagnostic space, can be written as a ratio in the untransformed diagnostic space

$$\textit{discriminant ratio} = \frac{\textit{result } 1^{b1}}{\textit{result } 2^{-b2}}$$

Because of their superior diagnostic accuracy, discriminant ratios are preferable to diagnostic ratios as the bases for positivity rules.

Diagnostic plots. Positivity rules for two-study test combinations can be very effectively presented graphically in what are called diagnostic plots. In such plots, the result combinations that are considered negative are indicated by one enclosed region and the result combinations that are positive are indicated by another. The diagnostic regions are generally nonoverlapping.

In addition to simplifying the application of individual positivity rules, diagnostic plots can incorporate multiple positivity rules, a feature that is extremely useful when the same pair of laboratory studies is used to diagnose a multiplicity of clinical conditions. Perhaps the premier example of this functionality are diagnostic plots of acid-base status based on blood pCO_2 and hydrogen ion concentration.

THE PROBABILITY OF DISEASE IN AN INDIVIDUAL PATIENT

There is usually some degree of uncertainty in the diagnostic classification of a patient. It arises from an inability to separate completely the presence of a condition from its absence on the basis of clinical or laboratory findings. This means that the presence of a disorder can be expressed only as a probability: “It is quite likely that you are affected,” “There is a fifty-fifty chance you have this disorder,” “You could be suffering from,” and so forth. When expressed quantitatively, the probability of a diagnosis being correct has a value between zero, which means that the condition is definitely not present, and one, which means that the condition is unarguably present.

Bayes' formula

Formal approaches exist for estimating the probability of a diagnosis. One of these, the Bayesian approach, is a well studied general method for making decisions in the face of uncertainty. Even in its formal realization, it is clinically practicable and it is one of the most common methods used in computer-based medical decision support. The Bayesian approach has informal counterparts in the diagnostic decision making of many clinicians, although not all clinicians agree that it is a good model of, or for, the diagnostic process (for example, Feinstein 1979).

Prior probability. The prior probability of a disorder in a patient, P[pre], is the estimate of the probability of the disorder arrived at prior to the performance of a stipulated diagnostic study. Thus, it is the pre-test probability. At the time of initial presentation of a patient, the prior probability is equivalent to the disease prevalence in the clinical population to which the patient belongs. That population is defined by the symptoms elicited by history taking, the signs revealed by physical examination, and additional pertinent historic and demographic data such as age, gender, disease history, disease exposure, and, in the evaluation of heritable disorders, family history and geoethnic lineage. Once diagnostic tests have been performed, the prior probability is equal to the prevalence of the disorder in the clinical subpopulation that is characterized by the results of those studies.

Dallman *et al.* (1981) found that the prevalence of iron deficiency in the apparently healthy 1-year-olds they studied was approximately 0.09. Thus, at presentation, the prior probability of iron deficiency in this population was 0.09. Infants who were subsequently found to have a low blood hemoglobin concentration had a number of additional diagnostic studies performed. The prevalence of iron deficiency in this (low hemoglobin) subpopulation was found to be 0.35 so, in terms of further testing, these infants had a prior probability of 0.35.

Posterior probability. Diagnostic laboratory studies are ordered with the intent of adjusting the estimate of the probability of a disorder in a patient based upon the study results. The revised estimate of the disorder's probability is called the posterior probability, P[post]. It is the post-test probability. The method of adjusting probability estimates to be discussed here is based upon Bayes' formula for inverting a conditional probability. The method possesses a great intuitive appeal, and in addition, the formulation is provable from the axioms of probability theory.

Consider the case of a 1-year-old who has a low blood hemoglobin concentration and who, on subsequent testing, is found to have a transferrin saturation of 7.5%. According to the findings of Dallman *et al.* (1981), this child's prior probability of iron deficiency is 0.35. What is her posterior probability of iron deficiency given the measured transferrin concentration? The critical value for transferrin saturation is 10% so the test is positive for iron deficiency. There are two ways in which this patient could have a positive result: she could have a true positive test result or she could have a false positive result. The probability of a true positive study result, P[true positive], equals the product of her prior probability of iron deficiency times the probability of registering a positive test result in a patient with iron deficiency. The latter probability equals the sensitivity of the study. So,

$$P[\textit{true positive}] = P[\textit{pre}]\ \textit{sens}$$

In this case, the prior probability is 0.35 and the sensitivity is 0.53 (from Table 3.3), so the probability of a true positive result is 0.19. The probability of a false positive result, P[false positive], equals the product of the pretest probability that the patient does not have iron deficiency (1 minus the prior probability) times the probability of having a positive result given that she is not iron deficient (1 minus the specificity). Therefore,

$$P[\textit{false positive}] = (1\text{-}P[\textit{pre}])\ (1\text{-}\textit{spec})$$

The prior probability is 0.35 and the specificity is 0.25 (Table 3.3), so the probability of a false positive result is 0.16.

The patient's posterior probability of iron deficiency, meaning the probability that she had a positive result because she is truly iron deficient, is equal to the probability of a true positive result divided by the total probability of having a positive result,

$$P[\textit{post}] = \frac{P[\textit{true positive result}]}{P[\textit{true positive}] + P[\textit{false positive}]}$$

$$= \frac{P[\textit{pre}]\ \textit{sens}}{P[\textit{pre}]\ \textit{sens} + (1 - P[\textit{pre}])\ (1 - \textit{spec})}$$

This is Bayes' formula. Using it, the posterior probability of iron deficiency in this patient is 0.54.

This form of Bayes' formula should only be used when dichotomous interpretation of study results is obligatory because of the qualitative nature of the study. When results are quantitative, as for most laboratory studies, categorizing the result into a binary classification results in loss of diagnostic information and unnecessarily restricts the values that the posterior probability can take.

The use of likelihood ratios based upon the frequency distributions of results in the pertinent reference populations allows incorporation of all the available diagnostic information in the calculation of the posterior probability of a diagnosis and broadens the range of values the probability can assume (Radack 1986). The likelihood ratio is the ratio of the frequency of a study result in one diagnostic group to the frequency of the result in another. In the example being considered here, it is the ratio of the frequency of a transferrin saturation of 7.5% in patients with iron deficiency to the frequency of that result in patients who are iron replete. Examination of Figure 3.1, the reference frequency histograms

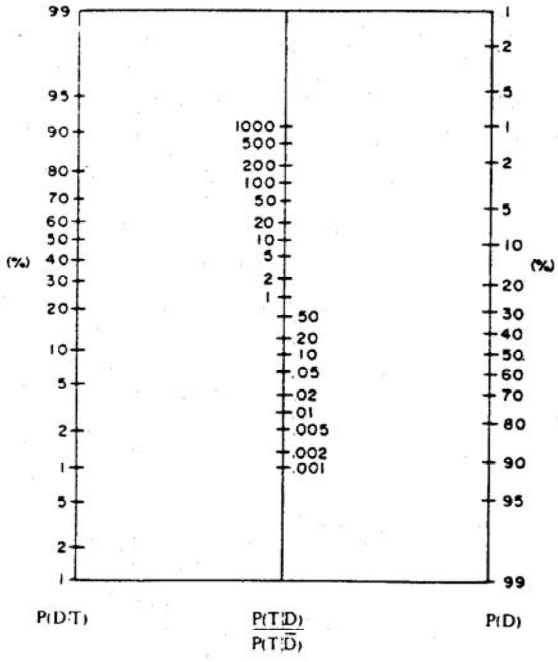

Figure 3.9 Nomogram for Bayes' formula. P(D), prior probability; P(T/D)/P(D/T), likelihood ratio; P(D/T), posterior probability. Reprinted from Fagan TJ. 1975. Nomogram for Bayes' theorem. N Engl J Med 293:257.

for testing with transferrin saturation, shows that in patients with transferrin saturations between 6 and 10% the likelihood ratio of iron deficiency is 1.57. Using the likelihood ratio form of Bayes' formula (Albert 1982),

$$P[post] = \frac{P[pre]\ likelihood\ ratio}{P[pre]\ likelihood\ ratio + (1 - P[pre])}$$

or the nomogram shown in Figure 3.9 (taken from Fagan 1975), the posterior probability of iron deficiency in the patient is 0.46. This estimate is lower than the one obtained from dichotomous interpretation of the study result. It is also more accurate. It quantitatively reflects the fact that a transferrin saturation in the range 6 to 10% occurs with only a slightly greater frequency in patients with iron deficiency than in patients who are not iron deficient, 33 percent versus 21 percent, respectively. This very relevant diagnostic information is lost when the study results are interpreted simplistically, in a dichotomous fashion, resulting in too high an estimate of the posterior probability of disease.

When likelihood ratios are derived from continuous distribution models of the frequency data the ratios take on a continuous range of values. This is illustrated in Figure 3.10 (line) which shows the likelihood ratio for iron deficiency as a function of transferrin saturation. The ratios have been calculated using lognormal distribution models of the frequency data of Dallman *et al.* (1981). Also shown on the figure (squares) are the empirical likelihood ratios derived from the reference frequency histograms. The likelihood ratio for a transferrin saturation of 7.5%, as determined using the model-based curve, is 1.88 which yields a posterior probability of iron deficiency of 0.50 for the example patient. This estimate is more accurate than that based on the empirical likelihood ratio because the empirical estimates are derived from binned data (all results between 6 and 10%).

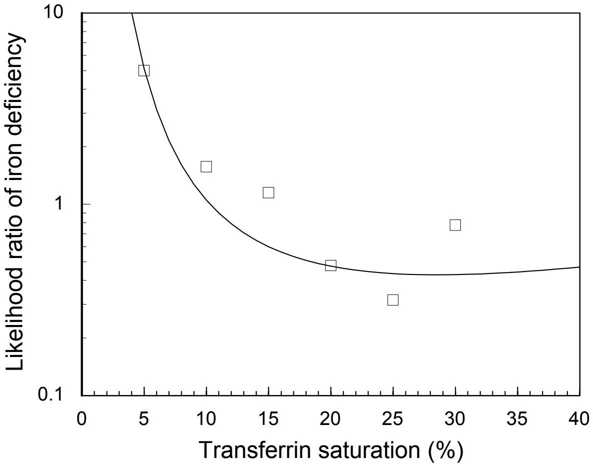

Figure 3.10 The likelihood ratio of iron deficiency as a function of transferrin saturation. The squares represent the points constructed from the observed frequency data (Figure 3.1). The continuous line is the curve constructed from the lognormal frequency distribution models of the data.

Remember that likelihood ratios will usually vary widely among different reference populations. Careless application of a likelihood ratio that is not appropriate to the actual clinical situation can be expected to result in erroneous posterior probability calculations and subsequent diagnostic inaccuracy.

Multiple study results. Clinicians rarely limit diagnostic testing to a single study. Instead, multiple studies are usually used. The clinical challenge, therefore, is the interpretation of a series or combination of study results.

Serial study interpretation will be discussed first because diagnostic evaluation most often proceeds in a sequential fashion. First, certain facts are uncovered by the history and physical; next, the results of the preliminary laboratory studies are obtained; and then, over a period of hours to weeks, the results of additional laboratory studies ordered by the clinician become available. As each new study result is received, the clinician is able to reassess the probability of the competing diagnoses using Bayes' formula. The posterior probability calculated from the preceding study result serves as the prior probability for the computation of the probability of a diagnosis based upon the current study results.

This approach is correct as long as there is no result correlation among the studies, that is, as long as the segregation of patients into subgroups according to study results does not affect the result frequency distributions and, hence, likelihood ratios, for any of the studies. When there is appreciable result correlation—and there usually is—this approach will generate probability estimates that are exaggerated; low probability estimates will be too low and high probability estimates will be too high. Indeed, as the number of study results becomes large, the probability estimate will approach either one or zero even though the true probability has an intermediate value (Russek *et al.* 1983).

In the presence of result correlation, conditional likelihood ratios must be used in Bayes' formula. A conditional likelihood ratio is the likelihood ratio for a study result calculated from reference populations who have identical results for the preceding studies. This ratio may be greater than, less than, or equal to the ratio that would be calculated from reference populations assembled without this restriction.

When multiple study results are analyzed in combination rather than serially, the following form of Bayes formula can be used, but only if there is no result correlation among the studies,

$$P[post] = \frac{P[pre]\ \Pi\ likelihood\ ratio_i}{P[pre]\ \Pi\ likelihood\ ratio_i + (1 - P[pre])}$$

This formula indicates that for a combination of *i* study results, the overall likelihood ratio used in calculating the posterior probability is the product of the likelihood ratios of each of the individual studies. When result correlation is present, the joint likelihood ratio should be used to calculate the posterior probability,

$$P[post] = \frac{P[pre]\ joint\ likelihood\ ratio}{P[pre]\ joint\ likelihood\ ratio + (1 - P[pre])}$$

The joint likelihood ratio is the ratio of the frequency of the combination of study results in the presence of the disorder to that in the absence of the disorder. Although the calculation of joint likelihood ratios is simple in the case of two diagnostic studies, as the number of studies increases the computational burden becomes significant. More importantly, tabulation of the ratios for their ready use clinically becomes nearly impossible, although the growing availability of computer databases may someday make it achievable (Krieg 1988). If the result frequency distributions behave according to a parametric statistical model, an enormous simplification can be realized because only the model parameter values need to be recorded. Specific result combination frequencies and joint likelihood ratios can then be calculated as needed.

Discriminant and logistic functions. If the result frequency distributions for a test combination satisfy the statistical conditions required for linear discriminant regression and are multivariate normal, the likelihood ratios for result combinations can be calculated directly using the discriminant function (Strike 1996),

$$likelihood\ ratio = e^{(Z_D + Z_{DF})/2 - Z}$$

where Z is the discriminant score for the result combination,

$$discriminant\ score = \Sigma\ bi\ result\ i$$

with i indicating the i th study, Z_D is the mean discriminant score among individuals with the disease, and Z_{DF} is the mean discriminant score among individuals who are disease-free. Dividing through by (1-P[pre]) and re-expressing the fraction P[pre]/(1-P[pre]) as an exponential allows Bayes' formula to be written,

$$P[post] = \frac{e^{\log\left\{\frac{P[pre]}{1-P[pre]}\right\} + (Z_D + Z_{DF})/2 - \Sigma\ bi\ result\ i}}{e^{\log\left\{\frac{P[pre]}{1-P[pre]}\right\} + (Z_D + Z_{DF})/2 - \Sigma\ bi\ result\ i} + 1}$$

In this form, which is that of a logistic function, the parameter values can be estimated using logistic regression techniques (Strike 1996). Logistic regression has two advantages over linear discriminant regression. First, the method is much more robust regarding deviations from statistical constraints; in particular, it can be used, cautiously, when the combination result frequency distributions are not multivariate normal and when the variance/covariance structure of the distributions in the two diagnostic classes are not identical. Second, logistic regression allows for the inclusion of qualitative and semiquantitative study results as test combination terms (Liao 1994). This capability is not often needed in the realm of diagnostic discrimination but it is indispensable in prognostic discrimination.

Table 3.4
Distribution of posterior probabilities for an example with imprecise prior probability and likelihood ratio estimates

P[post]	Distribution
0.57	0.0625
0.67	0.2500
0.73	0.0625
0.75	0.3125
0.80	0.1250
0.82	0.1250
0.86	0.0625

Imprecision. Variability in the quality of epidemiologic investigations of disease prevalence and in the performance and interpretation of clinical studies produces estimates of prior probability that are imprecise. Imprecision in estimates of the likelihood ratio arise from limitations in the quality of performance evaluations of laboratory studies and from the presence of inter-laboratory variability in the technical performance of the studies. Consequently, posterior probabilities arrived at using Bayes' formula are also imprecise (Diamond and Forrester 1983, Machin *et al.* 1983).

As a simple example, consider a case in which the estimate for the prior probability is 0.5 and that for the likelihood ratio of 3. Application of Bayes' formula using these estimates yields a posterior probability estimate of 0.75. Now suppose that the estimate of prior probability is not exact but instead consists of a 25 percent chance of a probability of 0.4, a 50 percent chance of a probability of 0.5, and a 25 percent chance of a probability of 0.6. Further suppose that the likelihood ratio estimate actually consists of a 25 percent chance of a ratio of 2, a 50 percent chance of a ratio of 3, and a 25 percent chance of a ratio of 4. If the true prior probability is 0.4 and the true likelihood ratio is 3, then the posterior probability is 0.67. Since the chances of the true prior probability being 4 and the true likelihood ratio being 2 are 0.25 and 0.5, respectively, the chance that both are the true values is 0.125 (the product of the separate chances). The complete distribution of values for the posterior probability (shown in Table 3.4) is obtained by repeating the foregoing calculations for all of the possible combinations of prior probability and likelihood ratio and aggregating the chances that correspond to identical values of the posterior probability (Iversen 1984). The range of values for the posterior probability is 0.57 to 0.86 with a central 87.5 percent confidence interval of 0.67 to 0.82. Notice that there is only a 0.3125 chance that the posterior probability is 0.75, the value calculated using the mean values for the estimates of the prior probability and likelihood ratio; the chances are 0.375 that the actual posterior probability is lower than 0.75 and 0.3125 that it is higher.

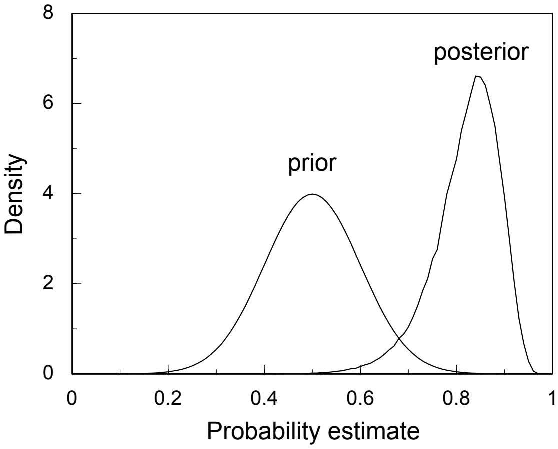

Figure 3.11 Distributions of prior and posterior probability for an example with normal distributions of the estimates of the prior probability and the likelihood ratio.

A somewhat more realistic example of the use of Bayes' formula when there is imprecision in the estimates of the prior probability and the likelihood ratio is illustrated in Figure 3.11. In this example the estimates of the prior probability and the likelihood ratio vary in a continuous fashion according to normal distributions. The mean value and standard deviation are set at 0.5 and 0.1, respectively, for the prior probability and at 5 and 1, respectively, for the likelihood ratio. The distribution of posterior probabilities that results from these inputs is left-skewed with a mode of 0.84. The central 90% confidence interval for the distribution is 0.69 to 0.91.

In practice, the imprecision inherent in the estimation of disease probabilities is rarely explicitly calculated in the foregoing quantitative fashion. Nevertheless, the clinician must always be mindful of such uncertainty, especially when prior probabilities and study performance measures are derived from research investigations with relatively small numbers of patients.

Study results that confirm or exclude a diagnosis

A confirming study result is one that raises the probability of a suspected diagnosis past the threshold probability for acceptance of the diagnosis, P(acceptance); the threshold probability for acceptance being that level of probability at which the physician and patient agree that the diagnosis is established with adequate certainty, given the pros and cons of making the diagnosis and in the knowledge that the diagnosis can subsequently be changed if the future course of the disease or the response to therapy are not typical of the diagnosed illness. Determining a threshold probability explicitly is no easy task and the most widely applied formal method for its calculation, clinical decision analysis (Pauker and Kassirer 1987, Kassirer *et al.* 1987), is still controversial. Still, the notion of a threshold probability is present, albeit in an informal form, in most diagnostic reasoning. Believing that an approximate value of the threshold probability can be identified, study values that confirm a diagnosis are those that yield a posterior likelihood of disease at least equal to the threshold probability. What that means in terms of result likelihood ratios can be appreciated by expressing Bayes' formula in the following form,

$$P[\mathit{acceptance}] = \frac{P[\mathit{pre}]\ \textit{threshold likelihood ratio}}{P[\mathit{pre}]\ \textit{threshold likelihood ratio} + (1 + P[\mathit{pre}])}$$

Rearrangement of this equation yields,

$$\textit{threshold likelihood ratio for acceptance} = \frac{(1 - P[\mathit{pre}])\ P[\mathit{acceptance}]}{P[\mathit{pre}]\ (1 - P[\mathit{acceptance}])}$$

Study results with likelihood ratios greater than the threshold likelihood ratio confirm the diagnosis.

Similar reasoning yields an analogous formula for the threshold likelihood ratio for rejection of a diagnosis in which the threshold probability for rejection of a diagnosis, P[rejection], is that level of probability at which it is agreed that the diagnosis is so unlikely that it can be excluded,

$$\textit{threshold likelihood ratio for rejection} = \frac{(1 - P[\mathit{pre}])\ P[\mathit{rejection}]}{P[\mathit{pre}]\ (1 - P[\mathit{rejection}])}$$

Study results with likelihood ratios less than the threshold ratio exclude the diagnosis.

Usually the threshold probability for acceptance of a diagnosis is different than the threshold probability for rejecting the diagnosis. That means that there exist intermediate probabilities that do not justify acceptance or rejection of the diagnosis. Patients with these intermediate probabilities require additional diagnostic workup. Sometimes, however, the threshold probability for acceptance of a diagnosis is equal to the threshold probability for rejecting the diagnosis; for instance, in situations in which patients in whom the diagnosis is rejected are to be seen at some subsequent time, offering another opportunity to evaluate them for the disorder.

The application of the formulas is illustrated by again using data of Dallman *et al.* (1981). The prior probability of iron deficiency in the screen-positive clinical population is 0.35. The threshold probability for accepting a diagnosis of iron deficiency and instituting oral iron therapy might, for example, be around 0.7, a value twice that of the prior probability. Substituting these values into the formula for the threshold likelihood ratio for acceptance of a diagnosis yields a ratio of 4.33. Figure 3.12 shows that a transferrin saturation of 5.5% is associated with this ratio. Thus, a transferrin saturation of 5.5% or less would be confirmatory of the diagnosis of iron deficiency. If the threshold probability for rejecting a diagnosis of iron deficiency were, for example, 0.1, the threshold likelihood ratio would be 0.21. No value of the transferrin saturation has a likelihood ratio that low so the measurement of transferrin saturation alone could not be used as a tool for the exclusion of a diagnosis of iron deficiency.

Screening for a disorder

A screening study is one used to detect a serious, treatable disorder in persons afflicted with the disorder but who have no clinical findings suggestive of the condition. Such clinically silent conditions are sometimes labeled "occult." It seems reasonable to define the threshold likelihood ratio for followup of a screening test result as the ratio that yields a posterior probability of the disorder equal to the threshold probability for rejecting the diagnosis. For study values associated with likelihood ratios greater than the threshold, the disorder cannot be considered excluded so further evaluation is clearly justified. The applicable formula is,

$$\textit{threshold likelihood ratio for followup} = \frac{(1 - \mathit{prevalence})\ P[\mathit{rejection}]}{\mathit{prevalence}\ (1 - P[\mathit{rejection}])}$$

Notice that, in this usage, the prior probability is the prevalence of the disorder in the screened population. Also notice that the formula reveals that the larger the threshold probability for rejection of a diagnosis, the greater the threshold likelihood ratio must be. This is somewhat unexpected because it means that the less critical the clinician is about excluding a certain disorder, the better the performance required of a screening test. Such a conclusion is at odds with the intuitive notion that less serious illnesses can be screened for casually, with studies of mediocre quality. But, in fact, evidence from a study of high quality is needed to convince a clinician to abandon an impression of health, which is after all the alternate hypothesis in an asymptomatic patient, in favor of the pursuit of a disorder of little clinical moment, especially when additional diagnostic evaluation involves stress, expense, and risks to the patient.

It is possible that some results of a screening study possess large enough likelihood ratios that a diagnosis can be made on the basis of these results alone. For this to be so, the study results must have likelihood ratios that exceed the threshold ratio for acceptance of the diagnosis,

$$\textit{threshold likelihood for acceptance} = \frac{(1 - \textit{prevalence})\, P[\textit{acceptance}]}{\textit{prevalence}\, (1 - P[\textit{acceptance}])}$$

Differences between populations as regards the frequency of risk or protective factors can significantly alter the prevalence of the disorder within the populations and, thereby, affect the value of the threshold likelihood ratio. Hence, the composition of the population subjected to screening is an essential consideration when calculating threshold likelihoods in screening for a disorder.

SELECTING DIAGNOSTIC STUDIES

A physician conducting a diagnostic evaluation usually has available a number of studies and study combinations from which to choose to address specific diagnostic questions. Which study or study combination is the best to order? The first step in answering the question is to decide if "best" means that, over a broad range of possible performance criteria, one study is superior to the alternative studies or if "best" means that the study is the most successful classifier within a specified criterion range.

When one wants to compare study performance over a wide range, the index of classification accuracy that is used is the area under the ROC curve. This index is appealing because it is equivalent, in the case of a diagnostic study, to the probability that, given two individuals, one with a disorder and one without, the study result will be more suggestive of the condition in the individual who has the disorder. Obviously, the larger this probability, the better a classifier the study is. To compare the classification accuracy of laboratory studies, then, one calculates the area under the respective ROC curves and tests the differences between the area estimates for statistical significance. If significant differences are found, the study with the largest area is the best classifier.

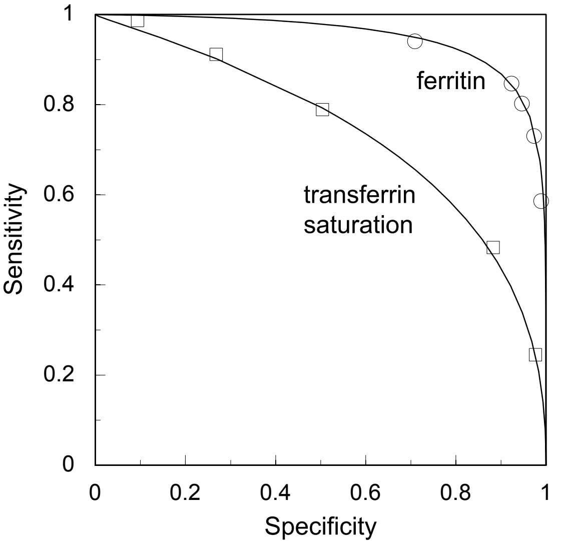

Figure 3.12 ROC curves for ferritin and transferrin saturation. The squares and circles represent the points derived from the empirical frequency data and the continuous lines are the curves derived from the lognormal frequency distribution models of the data.

A method for calculating and comparing the areas under ROC curves is available for data fitting normal distributions (Wieand *et al.* 1989). ROC curves arising from lognormally distributed data can also be analyzed by this method by log transformation of the data into its normally distributed form. Guyatt *et al.* (1992) describe the application of this method to the ROC curves for various markers of iron deficiency in adults with anemia. The empirical curves for plasma ferritin concentration and transferrin saturation from that paper and the curves arising from lognormal frequency distribution models of the data are shown in Figure 3.12. The area under the ferritin ROC curve is 0.95 (95% confidence interval, 0.94 to 0.96). The area under the transferrin saturation ROC curve is 0.74 (95% confidence interval, 0.70 to 0.78). The area under the ferritin curve is significantly larger than that under the transferrin saturation curve, so, overall, ferritin is a better—indeed, much better—study for the diagnosis of iron deficiency. Nonparametric statistical methods (i.e., methods that do not employ parametric data modeling) are also available for analyzing ROC curves (McNeil and Hanley 1984, DeLong *et al.* 1988).

Not infrequently, the comparison of the performance of different laboratory studies is relevant only within a certain range of performance criteria, such as when a study is sought to confirm the presence of disorder in an individual for whom the diagnosis is likely, or to exclude an important but unlikely alternative diagnosis, or to screen for a disorder among asymptomatic individuals. Which studies perform these specific clinical tasks best is revealed by a consideration of the performance characteristics necessitated by each.

In the case of a study to be used to confirm a diagnosis, it was shown that study results associated with likelihood ratios larger than the threshold ratio for acceptance of the diagnosis are considered confirmatory. Many studies may have results that satisfy this performance criterion. Which is the preferred study? It seems reasonable to propose that the study with the greatest sensitivity should be preferred. This assures that the maximum number of patients afflicted by the disorder will have the diagnosis confirmed when the study is performed. When selecting among excluding studies, the study with the largest specificity at the study result giving the threshold likelihood ratio for rejection should be preferred. Then the greatest number of patients free of the condition will have the diagnosis excluded. Because the object of screening studies is to detect a disorder, they must be sensitive. So, the preferred screening study should be the one with the highest sensitivity at the study value yielding the threshold likelihood ratio for followup.

The application of these selection rules for confirming and excluding studies are illustrated by considering the choice between plasma ferritin concentration and transferrin saturation in the diagnostic evaluation of an adult patient who is anemic. If the clinician's mindset is to prove that iron deficiency is not the cause of the anemia, he or she will want to order the study that is the preferred excluding study. If a diagnosis of iron deficiency is sought, so that iron therapy can be initiated promptly, the study that better serves as a confirming study should be ordered. If the prior probability of iron deficiency is assumed to be 0.35 and the threshold probability for accepting a diagnosis of iron deficiency is 0.7, the threshold likelihood ratio for acceptance is 4.33. Figure 3.13 shows the likelihood ratio of iron deficiency as a function of the study value for ferritin concentration and transferrin saturation based on the lognormal modeling of the data reported by Guyatt *et al.* (1992). Both studies have results that yield a likelihood ratio of 4.33, for ferritin it is a concentration of 25 μg/L and for transferrin saturation it is a value of 5.6%. At these values, transferrin saturation has a sensitivity of 0.26 and ferritin has a sensitivity of 0.75 so ferritin is by far the superior test for confirming the diagnosis. Using 0.1 as the threshold probability for rejecting a diagnosis of iron deficiency, the threshold likelihood ratio for rejection is 0.21. There is no value at which transferrin saturation has a likelihood ratio this low, so it cannot be used to exclude the diagnosis. Ferritin has the requisite likelihood ratio at a concentration of 107 μg/L. At that concentration, the specificity of the study is 0.70 making ferritin a very good study for excluding the diagnosis.

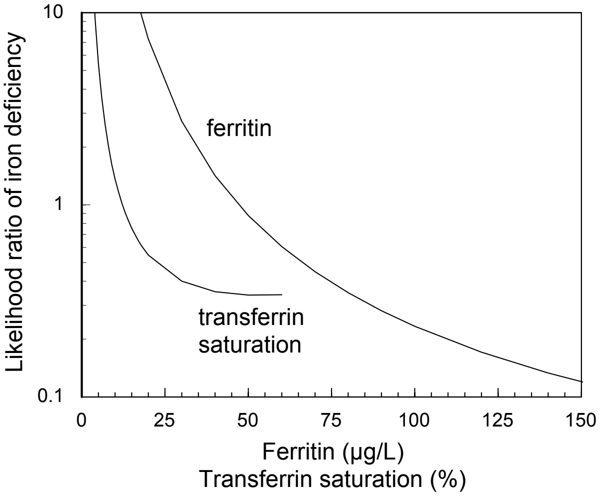

Figure 3.13 The likelihood ratio of iron deficiency as a function of ferritin and transferrin saturation based on lognormal frequency distribution models of the data.

PROGNOSTIC STUDY PERFORMANCE

Prognostic laboratory studies are used in two ways: to aid in predicting the outcome of an illness and to help predict if an individual will develop or relapse from a disorder at some specific time in the

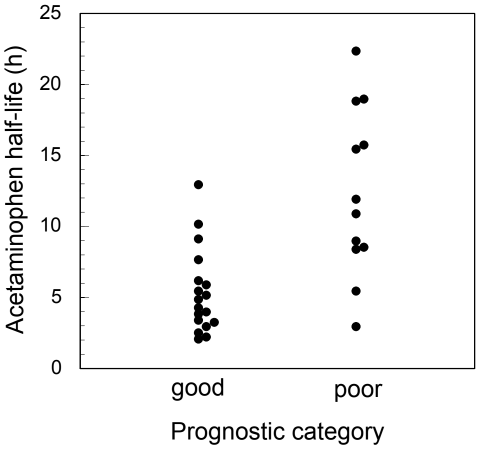

Figure 3.14 Individual values of acetaminophen half-life in acute acetaminophen poisoning with data sorted into two prognostic categories.

future. As both of these uses for prognostication represent exercises in clinical classification, it is possible to describe the performance of prognostic studies to a large extent using the same techniques that have been developed to characterize diagnostic classification.

In diagnostic classification, the fundamental measures of performance are sensitivity and specificity. Because there may be more than two prognostic categories and because outcome and risk categories are not necessarily as antithetical as the diagnostic categories of disease and free from disease, terms other than sensitivity and specificity need to be employed to measure performance in prognostic classification. Unfortunately, no distinctive terms have been invented for this purpose. Instead, the utilitarian phrase, fraction correctly classified, will be used here to quantify the fraction of individuals belonging to a prognostic group who are correctly placed into that group by the results of the laboratory study.

The prognostic group into which an individual is classified is determined by the study result for the individual and the study's critical values. When there are only two prognostic categories, there will be a single critical value. If the study result is smaller than the critical value, the individual will be assigned to one of the prognostic groups and if the result is larger than the critical value he or she will be assigned to the another. Because any study result can potentially function as the critical value, the comprehensive description of the classification performance of the study requires a listing of the fractions correctly classified for every possible choice of critical value. This is conveniently done graphically by means of a performance characteristic curve, i.e., a ROC curve.

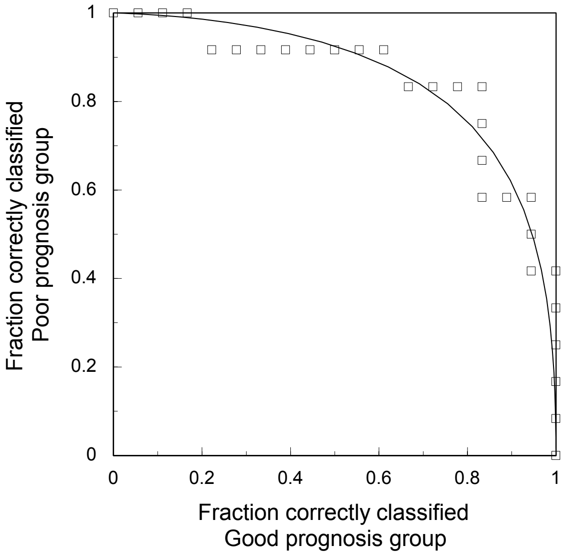

Figure 3.15 ROC curve for acetaminophen half-life. The squares represent the points constructed from the observed data. The continuous line is the curve constructed from frequency distribution models of the data.

For example, in a study comparing the prognostic performance of acetaminophen half-life and the ^{14}C-aminopyrine breath test as outcome predictors in acute acetaminophen poisoning, Saunders *et al.* (1980) report the acetaminophen half-lives shown in Figure 3.14. As pictured here, the good prognosis group consists of those patients who had either no liver damage or mild to moderate liver damage as a consequence of the drug overdose. The poor prognosis group is made up of those patients who either died acutely or who had severe liver damage. The ROC curve for these data is shown in Figure 3.15.

Multiple prognostic categories

When there are more than two prognostic categories, more than one critical value is necessary; the number being one less than the number of categories. Hence, two values are required to separate three categories: one value indicates the separating line between the good prognosis group and the intermediate prognosis group and the second value delimits the intermediate and poor prognosis groups. A complete description of the performance of the study as a prognostic classifier in this case requires tabulation of the fraction correctly classified for each of the prognostic categories for every

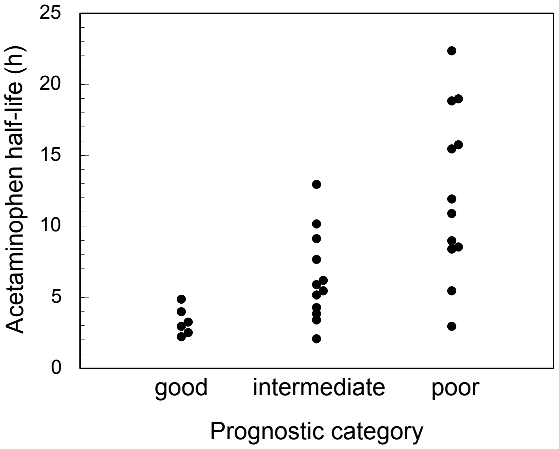

Figure 3.16 Individual values of acetaminophen half-life in acute acetaminophen poisoning with data sorted among three prognostic categories.

possible choice of critical value pairs. This produces a trivariate data set that can be represented graphically as a three-dimensional ROC surface.

Returning to the example from Saunders *et al.* (1980), consider the classification performance of acetaminophen half-life when three prognostic groups are defined—a good prognosis group, here consisting of patients who had no liver damage, an intermediate prognosis group composed of patients with mild to moderate liver damage, and a poor prognosis group made up of patients who had severe liver damage or who died acutely. The acetaminophen half-life data arranged according to this categorization of outcomes is shown in Figure 3.16 and the empirical ROC surface is shown in Figure 3.17.

Just as there are measures of diagnostic performance that incorporate disease prevalence, there are measures of prognostic performance that reflect the quantitative effects of prognostic category prevalence. The most important of these measures is prognostic efficiency, the fraction of individuals in a clinical population who will be correctly classified by the use of a prognostic study,

$$efficiency = \Sigma FCC_i \cdot prevalence_i$$

where FCC_i is the fraction correctly classified for prognostic category i and the summation is carried out over all the prognostic categories.

THE PROGNOSIS IN AN INDIVIDUAL

Prognoses are not like diagnoses. It is not necessary to eventually assign a patient to one or another prognostic group. Instead, it is enough to

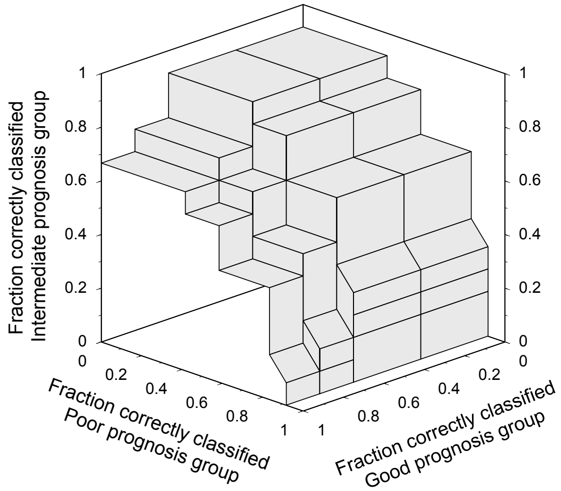

Figure 3.17 Empirical ROC surface for acetaminophen half-life used to classify three prognostic groups.

know, and let the patient know, the relative probabilities of being in each of the relevant prognostic groups. For instance, a clinician does not tell a patient "your cancer will recur" even though the probability of having a recurrence is greater than the probability of remaining recurrence-free. Instead, the patient may be told, "There is a 75 percent chance that you will have a recurrence of your tumor within 5 years."

The probability of an individual patient belonging to a particular prognostic group can be calculated using Bayes' formula. The best way to make the calculation is by using likelihood ratios, in which case (Birkett 1988),

$$P_j\,[post] = \frac{prevalence_j}{prevalence_j + \sum_{i \neq j} \dfrac{prevalence_j}{likelihood\ ratio_{ji}}}$$

where $P_{j\ [post]}$ is the posterior probability of being in the j th prognostic category and $likelihood\ ratio_{ji}$ is the frequency of the study result in group j divided by the frequency in group i. The summation is carried out over all the prognostic categories except category j. If there are only two prognostic categories, the preceding formula has the more familiar appearance,

$$P[post] = \frac{prevalence \cdot likelihood\ ratio}{prevalence \cdot likelihood\ ratio + (1 - prevalence)}$$

Figure 3.18 (left graph) shows the likelihood ratio of a good outcome in acute acetaminophen poisoning as a function of the acetaminophen half-life as

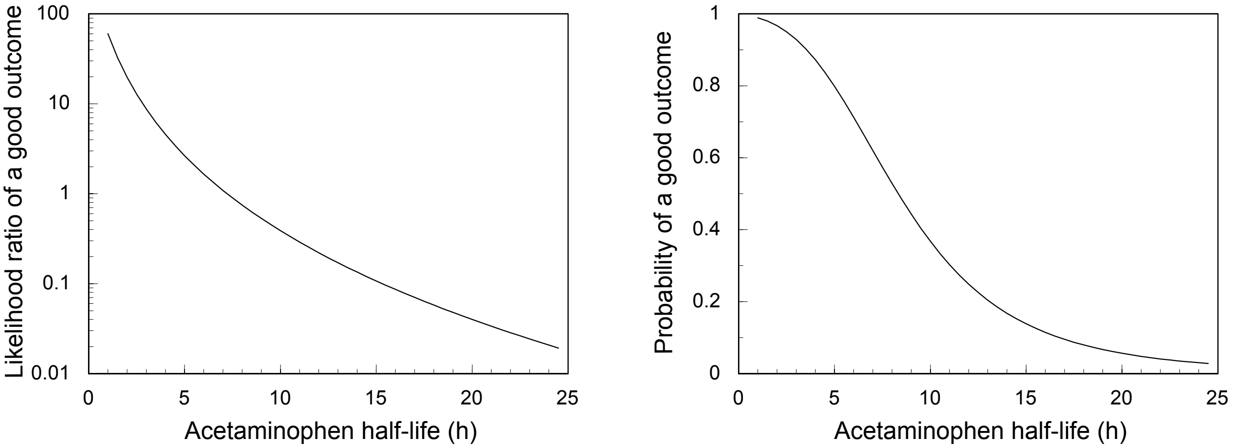

Figure 3.18 Bayesian calculation of posterior probabilities of a good outcome in acetaminophen poisoning. Left graph, the likelihood ratio of a good outcome as a function of acetaminophen half-life as derived from the lognormal frequency distribution models of the data in Figure 3.14. Right graph, the probability of a good outcome as a function of acetaminophen half-life given a value of 0.6 for the prevalence of a good outcome.

computed from lognormal modeling of the data in Figure 3.14. The posterior probability of a good outcome as a function of acetaminophen half-life, as calculated for a prevalence of 0.6, is also shown (right graph). At this prevalence, which was chosen because it is the prevalence in the study of Saunders *et al.* (1980), the study value corresponding to a probability of 0.5 is 8.3 h.

Regrettably, it is not infrequent for the prognostic performance of a laboratory study to be reported in such a way that it is impossible to calculate likelihood ratios for the study. Instead, one is forced to calculate posterior probabilities solely from the reported values of the fractions correctly classified for a limited set of critical values. For two prognostic categories, the form of Bayes' formula that must then be used is,

$$P[post]= \frac{prevalence \cdot FCC_1}{prevalence \cdot FCC_1 + (1-prevalence)(1-FCC_2)}$$

where FCC_1 and FCC_2 are the fractions correctly identified as reported for the critical value closest to the study result. Note the similarity of this formula and that for the calculation of diagnostic probability using sensitivity and specificity.

Logistic functions

The probability curve shown in Figure 3.20 (right graph) was computed by modeling the frequency data of each of the prognostic groups and then using the model parameters to calculate, by Bayes' formula, the probability of the indicated outcome. An alternative computational approach is to estimate the parameters of a model that directly describes the sigmoidal relationship between the value of the study result and the probability of membership in the prognostic group. Such models are called probability models (Liao 1994). By far the most commonly used probability model is the logistic model,

$$P[post] = \frac{e^{b0+b1\ result}}{e^{b0+b1\ result}+1}$$

where *b0* and *b1* are the model parameters.

Probability modeling using logistic regression has a number of very desirable features that explain its great appeal: it can be used to model posterior probabilities for test result combinations, in which case it has the form,

$$p[post] = \frac{e^{b0+\Sigma\, bi\ resulti}}{e^{b0+\Sigma\, bi\ resulti}+1}$$

it can use qualitative and semiquantitative study results and categorical variables either alone or in combination with quantitative study results, it enjoys considerable robustness to departures from the statistical constraints regarding data normality and variance/covariance structure, and it can be used when there are multiple prognostic groups (Strike 1996). Because logistic modeling allows the inclusion of other pertinent demographic and clinical data, logistic functions are the most common way to calculate posterior probabilities from prognostic study results. Care must be taken in their use, however, because logistic functions include the effect of prognostic group prevalence; it is

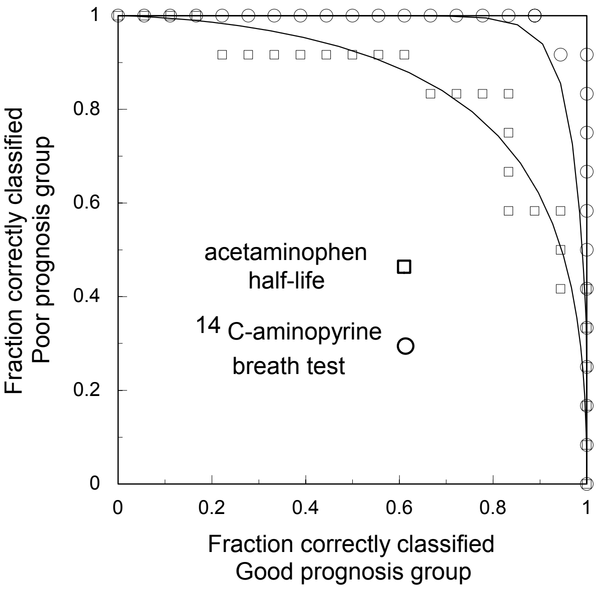

Figure 3.19 ROC curves for acetaminophen half-life and ^{14}C-aminopyrine breath test. The symbols represent the points derived from the observed data and the continuous lines are the curves derived from frequency distribution models of the data.

embedded in the constant exponential term, *b0*. In clinical settings with group prevalences different from those in which the function parameters were determined, the probability estimates will be inaccurate unless a prevalence adjustment is made (Poses *et al.* 1986, Morise *et al.* 1996).

SELECTING PROGNOSTIC STUDIES

Prognoses are usually based on the classification probabilities generated by prognostic models that take into account a combination of demographic, clinical, and laboratory data. Consequently, the question of what laboratory studies to order when determining a prognosis comes down to selecting the prognostic model that has been found to have the best prognostic performance.

When there are only two prognostic categories, the most useful index of classification accuracy in the comparison of prognostic performance is the area under the ROC curve. The methods for calculating the area and its confidence interval are identical to those described earlier for comparing diagnostic study performance. As an example, based on the data from Saunders *et al.* (1980), the ROC curves for acetaminophen half-life and the ^{14}C-aminopyrine breath test as outcome predictors in acute acetaminophen poisoning are as shown in Figure 3.19. The empirical ROC curves and the ROC curves derived from lognormal frequency distribution models of the data are illustrated. The area under the acetaminophen half-life ROC curve is 0.85 with a 95% confidence interval of 0.70 to 1.00. The area under the ROC curve for the ^{14}C-aminopyrine breath test is 0.99 with a 95% confidence interval of 0.96 to 1.00. The area under the ROC curve for the ^{14}C-aminopyrine breath test is larger than that for acetaminophen half-life, suggesting that that study is the superior prognostic tool; however, the confidence intervals overlap so the ^{14}C-aminopyrine breath test cannot be said with certainty to be better.

It might be imagined that a multidimensional extrapolation of the area under the ROC curve would serve as a useful measure of classification accuracy when there are multiple prognostic categories. Regrettably, there is as yet no statistical research concerning this measure or any other measure for performance comparison in the setting of multiple prognostic categories.

REFERENCES

Albert A. 1982. On the use and computation of likelihood ratios in clinical chemistry. Clin Chem 28:1113.

Beck JR and Shultz EK. 1986. The use of relative operating characteristic (ROC) curves in test performance evaluation. Arch Pathol Lab Med 110:13.

Birkett NJ. 1988. Evaluation of diagnostic tests with multiple diagnostic categories. J Clin Epidemiol 41:491.

Cebul RD, Hershey JC, and Williams SV. 1982. Using multiple tests: series and parallel approaches. Clin Lab Med 2:871.

Dallman PR, Reeves JD, Driggers DA, and Lo EYT. 1981. Diagnosis of iron deficiency: the limitations of laboratory tests in predicting response to iron treatment in 1-year-old infants. J Pediatr 99:376.

DeLong ER, DeLong DH, and Clarke-Pearson DL. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837.

Diamond GA and Forrester JS. 1983. Metadiagnosis. An epistemologic model of clinical judgment. Am J Med 75:129.

Fagan TJ. 1975. Nomogram for Bayes' theorem. N Engl J Med 293:257.

Feinstein AR. 1979. Clinical biostatistics. XXXIX. The haze of Bayes, the aerial palaces of decision analysis, and the computerized Ouija board. Clin Pharmacol Ther 21:482.

Guyatt GH, Oxman AD, Ali M, Willan A, McIlroy W, and Patterson C. 1992. Laboratory diagnosis of iron-deficiency anemia. J Gen Intern Med 7:145.

Harrell FE, Lee KL, and Mark DB. 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15:361.

Henderson AR. 1993. Assessing test accuracy and its clinical consequences: a primer for receiver operating characteristic curve analysis. Ann Clin Biochem 30:521.

Iversen GR. 1984. *Bayesian Statistical Inference*. Sage Publications, Newbury Park CA.

Kassirer JP, Moskowitz AJ, Lau J, and Pauker SG. 1987. Decision analysis: A progress report. Ann Intern Med 106:275.

Krieg AF, Beck JR, and Bongiovanni MB. 1988. Evaluating diagnostic performance of clinical tests by spreadsheet modeling. Arch Pathol Lab Med 112:588.

Krieg AF, Wagner CH, and Bongiovanni MB. 1989. Dot diagrams as source documents for evaluations of test performance. Arch Pathol Lab Med 113:746.

Liao TF. 1994. *Interpreting Probability Models. Logit, Probit, and Other Generalized Linear Models*. Sage, Newbury Park CA.

Machin D, Dennis NR, Tippett PA, and Andrews V. 1983. On the standard error of the probability of a particular diagnosis. Stat Med 2:87.

McNeil BJ and Hanley JA. 1984. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med Decis Making 4:137.

Morise AP, Diamond GA, Detrano R, Bobbio M, and Gunel E. 1996. The effect of disease-prevalence adjustments on the accuracy of a logistic prediction model. Med Decis Making 16:133.

Pauker SG and Kassirer JD. 1987. Decision analysis. N Engl J Med 316:250.

Politser P. 1982. Reliability, decision rules, and the value of repeated tests. Med Decis Making 2:47.

Poses RM, Cebul RD, Collins M, and Fager SS. 1986. The importance of disease prevalence in transporting clinical prediction rule. Ann Intern Med 105:585.

Radack KL, Rouan G, and Hedges J. 1986. The likelihood ratio. An improved measure for reporting and evaluating diagnostic test results. Arch Pathol Lab Med 110:689.

Russek E, Kronmal RA, and Fisher LD. 1983. The effect of assuming independence in applying Bayes' theorem to risk estimation and classification in diagnosis. Comput Biomed Res 16:537.

Saunders JB, Wright N, and Lewis KO. 1980. Predicting outcome of paracetamol poisoning by using ^{14}C-aminopyrine breath test. Brit Med J ii:279.

Solberg HE. 1978. Discriminant analysis. CRC Crit Rev Clin Lab Sci 9:209.

Strike PW. 1996. *Measurement in Laboratory Medicine: A Primer on Control and Interpretation*. Buttersworth-Heinemann, Oxford.

Wieand S, Gail MH, James BR, and James KL. 1989. A family of statistics for comparing diagnostic markers with paired and unpaired data. Biometrika 76:585.

Zou KH, Hall WJ, and Shapiro DE. 1997. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. Stat Med 16:2143.

Zweig MH and Campbell G. 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 39:561.