Chapter 4
EVALUATING CLASSIFICATION STUDIES

EVALUATING MEDICAL UTILITY

A laboratory study has medical utility if it meets a clinical need as well as or better that the alternative approaches used to address that need. To determine if a particular test has utility as a classification study, it is necessary to find out how well it performs its role as a clinical classifier and how that performance compares with the performance of the other means used to achieve the classification. The investigation of the classification performance of a laboratory study is referred to as a performance evaluation.

Reference classification

A complete report of a performance evaluation includes the seven components listed in Table 4.1. The necessary start to the report is a clear statement of the diagnostic classes or prognostic classes meant to be distinguished by use of the study and of the criteria used to assign study subjects to the classes. Ideally, the method employed for the ultimate classification of subjects, the reference method, should be a perfect classifier, a so-called gold standard. In reality, of course, reference methods usually fall short of perfection. Often, reference methods for diagnostic classification are completely specific but not completely sensitive. This is true, for example, for methods based upon pathologic examination. They are not completely sensitive because a mild form of a disorder or a small focus of disease can be missed. Reference methods for prognostic classification can be perfect classifiers, such as when the prognostic groups are "dead in five years" and "alive in five years". However, in many situations the methods are not completely accurate. Consider, for example, when "disease recurrence at five years" and "no disease recurrence at five years" are the prognostic groups. Even the best reference method can be expected to misclassify some patients in whom recurrent disease is present but not yet clinically detectable. Imperfect reference methods that are completely specific give a correct estimate of the sensitivity of a diagnostic study but lead to underestimation of its specificity (Statquet *et al.* 1981). Similarly, for two-group prognostic studies, an imperfect reference method that is completely accurate in classifying members of the non-event group leads to a correct estimate of the fraction correctly classified in the event group but yields an underestimate of the fraction correctly classified in the non-event group (Table 4.2).

Table 4.1
Components of a Performance Evaluation Report

1. Definition of the diagnostic or prognostic classes
2. Description of the reference method or technique used to assign subjects to the diagnostic or prognostic classes
3. Definition of the clinical setting
4. Description of the study population
5. Description of the analytic procedures and mathematical techniques used
6. Description of performance
7. Description of validation study

Sometimes the reference methods that are used are neither completely specific nor sensitive, when evaluating a diagnostic study, or are inaccurate in classifying the members of both the event and non-event groups, when evaluating a prognostic study. This is often the case when more definitive reference methods are unduly invasive, painful, expensive, or inconvenient. Also there are disorders for which no widely accepted gold standard classification method exists. Estimates of the classification performance measures determined in an evaluation using such reference methods are subject to error and thus, if uncorrected, must be considered rough approximations (Walter and Irwig 1988). Correction of the estimates is sometimes possible, however. For instance, if the performance measures of a reference method have been determined at some other time by comparison with a true gold standard, those values can be used to calculate the corrected estimates of the performance measures of the method being evaluated (Statquet *et al.* 1981). In addition, corrected performance measure estimates can be derived using other more elaborate evaluation designs such as repeat testing (Yanagawa and Gladen 1984, Schulzer *et al.* 1991), testing with multiple studies (Yang and Becker 1997, Torrance-Rynard and Walter 1997, Qu *et al.* 1996), and testing in two

Table 4.2
Effects of Common Deficiencies in the Design of Performance Evaluations

Deficiency	Sensitivity or fraction correctly classified event group	Specificity or fraction correctly classified non-event group
Reference classification		
Imperfect reference method		
sensitivity<1 fraction correctly classified, event group<1	correct	under
specificity, sensitivity<1 fraction correctly classified, both groups<1	incorrect	incorrect
Ascertainment bias	incorrect	incorrect
Diagnostic-review bias	over	over
Incorporation bias	over	over
Study population spectrum		
Inappropriate population	(not applicable)	over
Inadequate heterogeneity	over	over
Work-up bias	over	under
Selection bias	over	under
Analytical methodology		
Inaccurate or imprecise method	under	under
Test-review bias	over	over

clinical populations, assuming that the study performance will be equivalent in both (Hui and Walter 1980).

Other forms of systematic error, or bias, in reference classification of the study population have been recognized. These include ascertainment bias and diagnostic-review bias. Ascertainment bias results from unequal vigor in application of the reference classification method among all the study subjects (Wasson *et al.* 1982). For instance, if the reference technique is invasive, subjects who are less ill will often be spared the procedure but may still be assigned to the disease-free group or the favorable prognosis class. Similarly, if long-term monitoring of subjects is necessary for their final classification, subjects at low risk or with less severe symptoms may be included in the analysis even though they are lost to follow-up.

When the results of the method under study are known to the investigator making the reference class assignments, and when such knowledge can influence the classification made, there is a risk that, consciously or unconsciously, the classification will be biased in favor of agreement with the study method results. This kind of bias, called diagnostic-review bias (Ransohoff and Feinstein 1978), leads to overestimation of study performance. It can be avoided by "blind" interpretation of the reference method results. A similar form of bias, incorporation bias (Ransohoff and Feinstein 1978), arises when the reference method includes as one of its criteria the results of the method whose performance is being studied. Such circular reasoning is especially likely to arise when continuing controversy concerning the precise clinical utility of a diagnostic or predictive test motivates researchers to evaluate its performance after it has been incorporated into clinical practice.

The following are excerpts from the methods section of an unusually thorough evaluation report detailing the performance of a number of plasma enzymes for the diagnosis of acute myocardial infarction (Werner *et al.* 1982):

> We investigated patients with acute myocardial infarction and patients in whom this condition was suspected but ruled out. The diagnostic classifications were established after the patient's discharge by a review of all clinical findings, including history, electrocardiographic data, and laboratory data. The document on "Nomenclature and Criteria for Diagnosis of Ischemic Heart Disease" [Report of the Joint International Society and Federation of Cardiology/World Health Organization Task Force on Standardization of Clinical

> Nomenclature. Nomenclature and criteria for diagnosis of ischemic heart disease. Circulation 1979; 59:607-609] was used for the diagnosis of infarction. Twelve-lead electrocardiograms were recorded on admission and repeated every 24h for four days. The development of Q waves with a duration of 0.04s was considered diagnostic of an acute transmural infarct; S-T and T wave changes with evolution were taken to indicate a nontransmural infarction.

In this investigation, subjects were assigned to the diagnostic classes according to a widely accepted set of clinical and laboratory criteria. Although these criteria do not qualify as a gold standard, as there is still no universally acknowledged perfect diagnostic method for the clinical diagnosis of myocardial infarction (Lee and Goldman 1986), they constitute a generally accurate and reproducible diagnostic tool. However, the results from the very laboratory studies under study are included among the criteria used for reference classification. Therefore, some degree of incorporation bias is certain to affect the performance estimates. Still, the absence of a gold standard method of classification necessitates the use of this reference method. It must be kept in mind, however, that the findings of the evaluation are biased toward overestimating the diagnostic performance of the enzyme markers.

Study population

The third component of a performance evaluation report is a definition of the clinical setting in which the study is to be used. This means that the report must indicate the relevant medical history, signs and symptoms, and preliminary clinical and laboratory study results in patients in whom the study under review is to be used. The definition, while specific, should be general enough to include the range of patient presentations actually seen in practice. The clinical setting for use of the laboratory studies evaluated by Werner *et al.* was:

> Our study was designed to arrive at a protocol for the use of enzymes in the diagnosis of myocardial infarct ... under actual clinical circumstances ...

The "actual clinical circumstances" referred to can be deduced from the make-up of the study population; it consisted entirely of patients admitted to a Cardiac Care Unit. The setting, therefore, is inpatient evaluation of acute myocardial infarction. The findings of the evaluation may not apply to the use of the diagnostic studies in different clinical settings such as work-up of acute chest pain in the emergency room (Lee and Goldman 1986). The extent to which they apply to the diagnosis of perioperative myocardial infarction depends upon the representation of such patients among the study population. The wording of the reports suggests that none of the subjects were perioperative. The clinical setting is further defined by the authors in terms of the interval between the occurrence of the infarct and the time of admission and diagnostic testing of the patients:

> Patients for whom the aggregate data suggested that infarction occurred within 24 hour before hospitalization were classified as "early admission," patients in whom infarction occurred within 24-48 hours before hospitalization as "intermediate admissions," and patients in whom infarction occurred within 48-72 hours before hospitalization as "late admissions." Enzyme values were separated into four groups: those obtained from uninfarcted patients, and those obtained from infarcted patient on the first, second, and third day after infarction.

After defining the clinical setting, the subjects who participated in the evaluation are described. Obviously, the study subjects should be a sample of individuals from the stipulated clinical setting. This condition is sometimes not met, though. For instance, patients with completely unrelated disorders or even healthy persons may be used to determine the specificity of a diagnostic study for which the sensitivity has been ascertained in a population limited to patients known to have the diagnosis, such as patients attending a specialty clinic. In such an evaluation, specificity will be overestimated. The subjects should also constitute a representative sample. Thus the spectrum of clinical variability within the study population should be as broad as possible (Ransohoff and Feinstein 1978). Patients with mild and early forms of a disorder or those falling into more favorable prognostic classes should be included in the evaluation, as should patients with unusual clinical features. The population should not consist entirely of patients who are in the advanced stages of a disorder or who have an extreme prognosis. There should also be considerable heterogeneity among the individuals in favorable diagnostic and prognostic groups. Some of the subjects should suffer from related disorders and illnesses that can be confused with the condition under investigation. Indeed, there should be many patients in whom the reference classification cannot be made without use of the reference technique. Also, the subjects should represent a diverse sample in terms of biologic attributes, such as age and gender. Failure to assemble a population with an adequate biologic spectrum usually results in overestimation of study performance.

Sensitivity will be overestimated if the study population consists of subjects selected because their chances of having a particular disorder are great enough to justify the use of an invasive, painful, or costly reference method of classification. This form of bias is called work-up bias (Ransohoff and Feinstein 1978). In the evaluation of a prognostic study, work-up bias originates from the preferential selection of subjects with a high likelihood of belonging to prognostic classes that are serious enough to warrant the use of an impractical or expensive reference method. Work-up bias leads to overestimation of the fraction correctly classified in those prognostic classes. Related to work-up bias is selection bias (also called verification bias) which arises when the results of the classification method under study determine which subjects will undergo reference classification and thereby be included in the evaluation. This form of bias is particularly likely to appear in evaluations of screening tests because these investigations often have a study design in which performance of the reference method is limited to those individuals who test positive when screened. If selection bias exists, the sensitivity of a diagnostic method will be overestimated and its specificity underestimated; similarly, for prognostic studies, the fraction correctly classified will tend be overestimated in classes with poor prognoses and underestimated in classes with good prognoses. There are parametric (Begg and Greenes 1983, Gray *et al.* 1984) and nonparametric (Zhou 1996) methods for obtaining unbiased estimates of study performance when selection bias exists .

The schemes that are employed to sample individuals from a stipulated clinical setting are of three general types, referred to as naturalistic, retrospective, and prospective by Kraemer (1992). Naturalistic sampling is characterized by either random sampling or strict consecutive sampling of the population of interest. Such sampling results in a study population with a prevalence of disease comparable to that of the clinical population. Using this scheme, the estimate of the efficiency of the study is unbiased but the estimates of sensitivity and specificity or fraction correctly classified are biased in inverse proportion to the number of subjects studied and the prevalence of the diagnostic or prognostic class. With retrospective sampling, members of the pertinent clinical population are screened at random or consecutively using the reference method and then random subsets of individuals in each diagnostic or prognostic class are tested using the study under evaluation. This sampling approach yields estimates of sensitivity and specificity or fraction correctly classified that are unbiased. A practical and financial advantage of this approach compared to that of naturalistic sampling is that the study under evaluation needs to be performed in considerably fewer individuals in the diagnostic or prognostic groups that are common. For instance, in the evaluation of a study used to diagnose a disorder with a prevalence of 0.2, if all of the screen-positive individuals are subsequently studied, only one quarter of the screen-negative individuals need to be studied to have the same number of data points for the estimation of specificity as there are for the estimation of sensitivity. The last type of sampling scheme, prospective sampling, is, as its name implies, the inverse of retrospective sampling. Here, the clinical population is screened using the study being evaluated and then subsets of individuals classified as to their diagnostic or prognostic class according to the study are further tested using the reference method. The unbiased performance measures obtained using this scheme are the predictive values of a test result. Sensitivity and specificity or fraction correctly classified must be derived from the respective predictive values and an indirect estimate of prevalence by using Bayes' formula (Choi 1992, Kraemer 1992). When compared to naturalistic sampling, this scheme results in the reference method being performed on fewer individuals in the diagnostic or prognostic groups that are common. This is advantageous when the reference method is expensive or risky. The disadvantage of prospective sampling is that the performance of a study can only be evaluated at a few critical values thereby making a comprehensive performance evaluation impossible.

The subjects comprising the population studied by Werner et al. are described thus:

> The sample represents the aggregate of admissions to the Cardiac Care Unit (GWU Med. Center) during the period for study, and in this sense reflects the prevalence of myocardial disease in a "complaining" population. However, from this sample, only individuals for whom enzyme assays were done on at least two consecutive days were included in this study ... Table 1 lists the patients used in the former analyses by disease state, age, and sex.

The authors have attempted to avoid sampling bias in the assembly of the study population by using a naturalistic sampling scheme with enrollment of all of the patients admitted to their Cardiac Care Unit. Unfortunately, not all of the patients had plasma enzyme studies performed on at least two consecutive days, so some of the potential subjects were excluded from the evaluation. Because the patients in whom repeat diagnostic testing was not pursued were probably not selected in a random fashion from the Unit's population, their exclusion does introduce a bias into the evaluation. If the excluded patients were those in whom the initial clinical findings indicated only a small likelihood of myocardial infarction, the bias is of the work-up type. The information that the investigators collected regarding the clinical and biologic spectrum represented in the study population is summarized in Table 1 of the report. There are subjects in all of the indicated clinical subgroups but for some, such as infarct-free patients with prior myocardial infarction, the number of subjects studied is small. This makes subgroup differences difficult to demonstrate by statistical analysis. More importantly, though, small numbers necessarily limit the degree of biologic variability among the subjects and thereby lessen the reliability of the performance estimates.

Analytic methodology

The final component of study design considered in the evaluation report is the description of the analytic procedures used to make the study measurements. Although it often happens that little attention is given to this description, the procedures chosen can affect the clinical utility of the study dramatically. Patient preparation, the manner of specimen collection and handling, and the analytic methodology, including instrumentation, need to be specified.

The use of inaccurate methods, especially those suffering from poor analytic specificity, or imprecise methods will lead to underestimation of study performance. As was mentioned in the description of diagnostic-review bias, study performance will be overestimated if there is a bias in favor of having study results agree with reference classifications. In the case of diagnostic-review bias, this can happen when study results are known at the time the reference classifications are made. A similar bias may arise if the test under evaluation has a subjective element to its interpretation and the reference classifications of the subjects are known to the individual reviewing the test results at the time the interpretations are made. This is called test-review bias (Ransohoff and Feinstein 1978). "Blind" interpretation of study results protects against this bias.

Information about the analytic methods should be made available in the evaluation report either in the form of summary statements of their technical performance attributes or by referencing separate technical method evaluations. Werner *et al.* use the latter approach:

> All enzymes were assayed at 37° C. A mechanized system (System TR; Beckman Instruments, Inc., Fullerton, CA 92634) was used to perform the following kinetic assays: the CK assay of Oliver

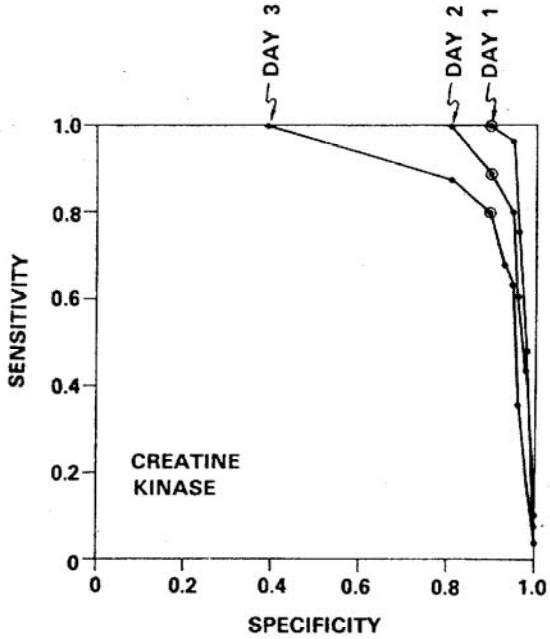
The mathematical techniques used for data exploration and analysis are crucial methodological elements of a performance evaluation. They should be identified and, when necessary, the appropriateness of their use should be discussed (Wasson *et al.* 1985, Concato *et al.* 1993, Simon and Altman 1994). It is especially important that statistical assumptions that may be violated by the data be addressed. An example would be the need to demonstrate the approximate normality of the data if statistical methods based upon normal distributions are used. When a multivariate analytic approach such as discriminant or logistic regression is used for interpreting study result combinations, the goodness-of-fit of the derived classification rule should be assessed (Hosmer *et al.* 1991, Harrell *et al.* 1996, Hosmer et al. 1997).

If the evaluation recommends the use of a certain critical value, the performance of the method at that value should be subjected to a statistical test of significance. This amounts to asking the statistical question: to a stipulated level of confidence (again, usually 95 percent certainty), are the number of misclassifications at the critical value less than would be expected on the basis of chance alone? If the critical value has been specified prior to reviewing the performance data, the appropriate statistical tools are the Fisher exact test, for small sample numbers, and the chi-square test, for large sample numbers. If, however, the critical value has been selected after examining the data, as happens when it is chosen so as to maximize study efficiency, statistical significance should be assessed by using p values corrected for *post hoc* selection of the critical value or by using a statistical test designed for such circumstances such as the one developed by Gail and Green (1976).

Werner *et al.* report the statistical method, stepwise discriminant function analysis, and the computer program, BMDP7M, they employed for their investigation of combination testing. The use of stepwise regression as a tool in the analysis of multivariate rules is very common even though the method can be problematic (Harrell *et al.* 1985, Diamond 1989, Simon and Altman 1994). Difficulties associated with the method include inconsistency in the selection of the significant variables, bias in the estimation of the regression coefficients, and overestimation of the statistical significance of the coefficients. A validation study (*vide infra*) is an absolutely essential component of any performance evaluation in which this analytic technique is used.

Presentation of findings

The presentation of the performance data, the sixth component of the evaluation report, should be in as complete a form as possible. Ideally, the reference result frequency distributions should be given. From these, readers can construct the ROC curve and the likelihood ratio curve for the study. The ROC curve describes the performance of a study in the form appropriate for comparing with alternative studies and the likelihood ratio curve describes the performance in the form needed for the Bayesian assessment of classification probabilities. It is desirable, of course, that the ROC and likelihood ratio curves be presented in the report rather than having the readers generate them (Jaeschke *et al.* 1994). In the evaluation of a multivariate diagnostic or prognostic rule, it is impossible to present the joint reference result frequency distributions if more than two studies are involved. It is possible to display the reference result frequency distributions defined by the rule and this should be done. The ROC and likelihood curves for the rule derived from these distributions should also be presented. Werner *et al.* present their performance data as ROC curves, one for each of the clinical settings considered in the article:



The method for graphing ROC curves is not standardized. In this example, the horizontal and vertical axes are, respectively, specificity and sensitivity. This agrees with the convention used in this book. Graphs in which the horizontal axis is one minus specificity (usually called the "false positive rate") and the vertical axis is sensitivity are often found. They give curves that are left-to-right mirror images of those obtained when the horizontal axis is specificity. The practice of identifying at least some of the points on the curve with the corresponding critical values is not a standard practice either. Here, the point associated with 120 U/L is circled.

Unfortunately, one does not always find a complete presentation of the performance results. Not uncommonly, study evaluations report the classification performance of the study at a single critical value. In that case, the performance description should state explicitly the basis for the authors' selection of the value. The three most frequently cited reasons are that the value (1) is used by other researchers, (2) yields a specificity of 0.95 (for diagnostic studies), or (3) yields maximum efficiency among the subjects studied. If the first criterion applies, the results of the evaluation can be compared to those reported by others. However, it may not permit ready comparison of the performance of the method with that of alternative diagnostic studies. The second criterion permits comparisons among alternative studies also evaluated at critical values yielding a specificity of 0.95. The third criterion is problematic as it often makes it so that the performance findings cannot be compared to the findings from other evaluations concerned with the same or alternative studies.

The statistical treatment of performance results includes calculation of the confidence limits for the performance estimates. Confidence limits define the range of values within which, to a stipulated level of confidence, the true value of the estimate lies. For proportions (such as sensitivity, specificity, and fraction correctly classified), confidence limit calculations are based upon the properties of the binomial distribution. For a proportion derived from a large number of subjects (N more than 50), the approximate confidence limits of the estimate are,

$$\frac{estimate + \frac{z_c^2}{2N} \pm z_c\sqrt{\frac{estimate\,(1 - estimate)}{N} + \frac{z_c^2}{4N^2}}}{1 + \frac{z_c^2}{N}}$$

where z_c is the confidence coefficient as found with the standard normal distribution; z_c equals 1.96 for a 95% confidence level and 1.645 for a 90% confidence level. For proportions derived from a small number of subjects (fewer than 50), the calculation of confidence limits is mathematically involved, so they are usually taken from a table or graph. The 95% confidence limits for proportions derived from samples of 10, 20, 30, and 50 subjects are shown in Figure 4.1. If the result frequency distributions are modeled, the performance measure estimates and associated confidence limits as computed from the model parameters should also be presented. As an example, based on the result frequency distributions for transferrin saturation as a classification study in the diagnosis of iron deficiency in 1-year-olds as reported by Dallman *et al.* (1981), the empirical specificity for a transferrin saturation of 10% is 0.75. As there were 110 iron-replete infants studied, the 95% confidence limits for the empirical specificity of the study result are 0.662 and 0.822. Based on lognormal models of the result frequency distributions, the specificity at a transferrin saturation of 10% is 0.79 with 95% confidence limits, 0.719 and 0.844. The width of the confidence interval for the estimate derived from the distribution model will always be smaller than the interval for the empirical estimate (White and James 1996). Here there is a 22% difference in the interval widths. The ROC curve for transferrin saturation (see Chapter 3) shown in Figure 4.2 indicates the 95% confidence intervals for the empirical sensitivity and specificity estimates.

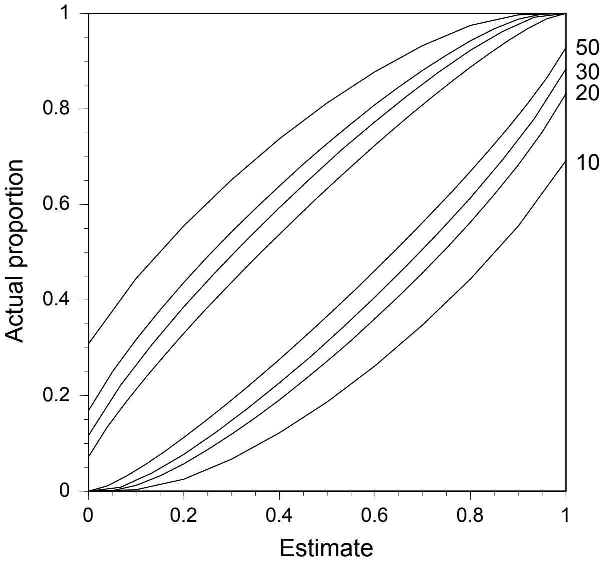


Figure 4.1 95% confidence limits for performance estimates. The number of study subjects is indicated.

Confidence limits should also be calculated for likelihood ratios (Fleiss 1981). Figure 4.3 shows the likelihood ratio curve for transferrin saturation (see Chapter 3) with the approximate 95% confidence intervals for the ratios indicated. Werner *et al.* state that:

> We estimated the uncertainties of these measurements by analogy with binomial distribution ... The uncertainty (standard deviation) ranges for sensitivity estimates from 0.70 ± 0.08 to 0.80 ± 0.07 and for specificity estimates from 0.96 ± 0.02 to 0.99 ± 0.01.

Some researchers also present data for the predictive value of study results (Linnet 1988). When these calculations are based upon appropriate epidemiologic estimates of the prevalence of the

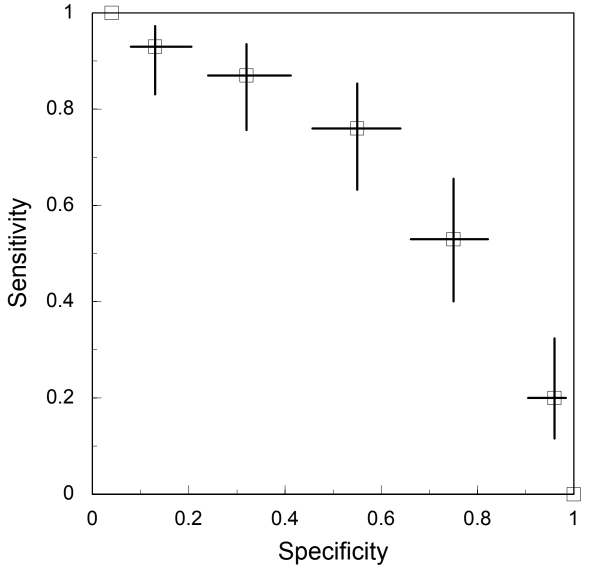


Figure 4.2 ROC curve for transferrin saturation with 95% confidence intervals indicated. The intervals for the specificity estimates are shown as lines parallel to the specificity axis and the intervals for the sensitivity estimates are shown as lines parallel to the sensitivity axis.

diagnostic or prognostic classes, the predictive values are meaningful. Studies employing properly performed naturalistic and retrospective sampling schemes include an epidemiologic inquiry into class prevalence, so the reported predictive values are valid for the site at which the evaluation was performed. The values may not apply at other sites because the prevalence of diagnostic or prognostic classes can differ between locations and institutions.

The problem of varying class prevalence at different sites also arises with classification rules based on logistic functions. Because logistic functions include terms for class prevalence, the posterior probability calculations reflect the class prevalences at the site evaluating the classification study. The probability estimates will be erroneous at sites where the class prevalences are different unless the classification rules are corrected for local class prevalence (Poses *et al.* 1986, Morise *et al.* 1996).

Validation

The final component of a performance evaluation report is the description of the methods and results of a validation study of the performance findings. The most convincing way to demonstrate the validity of the findings is to perform an identical investigation in new group of subjects and to arrive at the same performance estimates. This task is usually left to other researchers. What is done instead is to confirm the reported findings among subjects from the original study population. This is called cross-validation. The simplest design for a cross-validation study, and the one most often seen in the medical literature, is to perform the evaluation using only some of the subjects, the training sample, having selected them at random. The evaluation is then repeated using the remainder of the subjects, the validation sample. Concurrence of the performance estimates in the two groups indicates that the findings are valid. A particularly powerful way to demonstrate concurrence is to use the likelihood ratio estimates derived from the training sample to predict the probability of class membership among the individuals in the validation sample. The predicted probabilities are then compared to the observed probabilities, i.e. class membership frequencies, by plotting them as a calibration curve (e.g., Figure 4.4). Points for the curve are generated by binning the predicted probability results into subgroups each of which has roughly the same number of data. If the performance estimates are valid, the calibration curve will closely follow the line of identity.

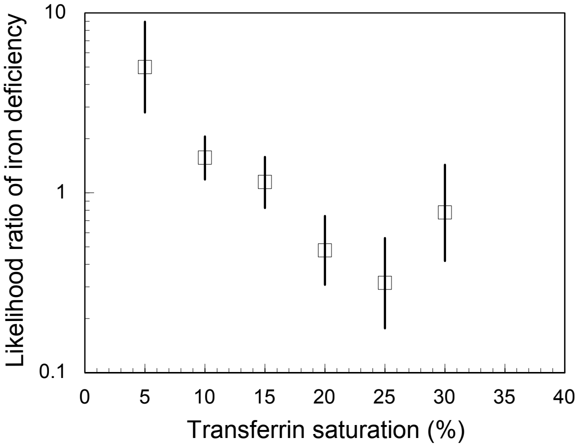


Figure 4.3 The likelihood ratio of iron deficiency as a function of transferrin saturation with 95% confidence intervals indicated.

Another way to demonstrate the validity of a performance evaluation is to show that the findings do not change as a consequence of reasonable variation in the analytic and reference methods or in the composition of the study population. Such an analysis can be provided for as part of the evaluation design, for example, by using a design that permits calculation of the magnitude of analytic variability,

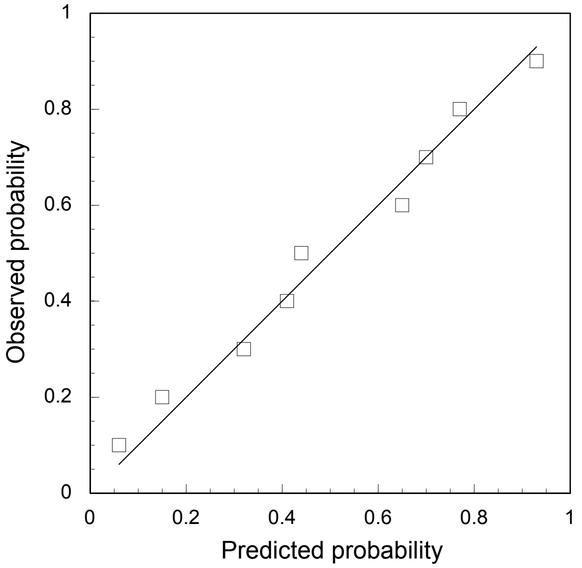


Figure 4.4 Hypothetical calibration curve from a validation study. The squares represent the mean values of the subgroup probabilities and the continuous line is the line of identity.

or it may be performed retrospectively, for example, by using regression analysis to evaluate the effects of demographic variables.

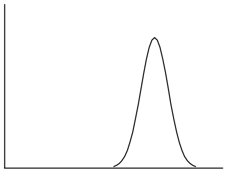
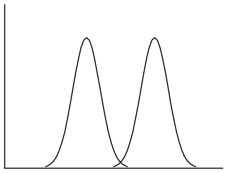
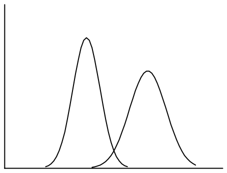
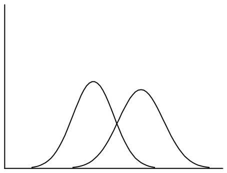
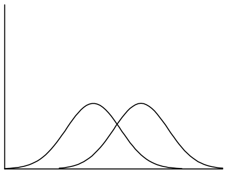
META-ANALYSIS

Most promising laboratory studies are the subject of numerous performance evaluations within a short time following their initial descriptions. Especially interesting studies are likely to engender a daunting bibliography. Given a profusion of reports, how does one integrate the findings to arrive at an accurate appraisal of the performance of the study? Clearly, a well-defined, systematic approach is called for, one that is able to deal in a quantitative way with the numerical data generated in performance evaluations. The field of study concerned with this question is the discipline of meta-analysis (Jenicek 1989). Research in meta-analytic techniques has been conducted for barely 30 years and applications in clinical medicine have appeared only recently. To date, most medical meta-analyses have dealt with the assessment of treatment effectiveness and cause-effect relationships. However, some work has been done in developing a meta-analytic approach to the assessment of classification performance (Irwig *et al.* 1994).

Performance evaluation meta-analysis proceeds in three steps: 1) the assembly of the pertinent evaluation reports, 2) qualitative meta-analysis, and 3) quantitative meta-analysis (Jenicek 1989). The first step, retrieval of the relevant literature, is not a trivial aspect of meta-analysis. Indeed, literature retrieval is itself an area of research within the discipline of medical informatics. Because of shortcomings in any single approach to searching the literature, it is recommended that multiple methods be employed, including searching computerized literature databases, reviewing appropriate journals, and consulting expert practitioners and laboratorians (Irwig *et al.* 1994). When relevant research remains unpublished because the findings are negative, i.e. show poor study performance, meta-analysis will overestimate study performance. This is a form of publication bias (Dickersin and Berlin 1992). It is clear that publication bias is a common problem in the meta-analysis of therapeutic research but the extent of this difficulty in the meta-analysis of classification study performance is not known.

Qualitative meta-analysis consists of the categorization of study reports according to the design of the performance evaluation and the assessment of the quality of the individual evaluations. Nierenberg and Feinstein (1988) have proposed the five category scheme of diagnostic study design shown in Table 4.3. Each successive category in the scheme is characterized by an increase in the breadth and rigor of the evaluation until, in category V, an ideal evaluation is achieved. Category IV evaluations are those that fall somewhat short of ideal, usually because of some limitations in the spectrum of the study populations. Category V and category IV evaluations are the ones upon which further meta-analysis should be performed. The findings of the exploratory evaluations constituting categories I, II, and III must be considered preliminary or provisional so these studies should not be included in the meta-analysis. A three category scheme for prognostic study design has been suggested by Simon and Altman (1994). In their scheme, category 1 consists of early exploratory evaluations, comparable to Nierenberg and Feinstein's categories I, II, and III. Category 2 represents evaluations of the clinical performance of a study as a means of classifying prognostic groups and category 3 consists of clinical evaluations designed to identify subsets of patients who will benefit from a given therapy. Depending upon the clinical spectrum and number of patients studied, the evaluations in these two categories correspond to Nierenberg and Feinstein's categories IV or V.

Table 4.3
Categories of Performance Evaluation Design

Category	Characteristics	Example study result distributions
I	performance of procedures cases: typical spectrum of disease controls: none	
II	coarse distinctions cases: typical spectrum of disease controls: healthy	
III	more subtle distinctions cases: expanded spectrum of disease controls: healthy	
IV	preliminary clinical application cases: include appropriate comorbidity controls: include appropriate comorbidity	
V	definitive clinical application cases: full spectrum controls: full spectrum	

The quality of category IV and V diagnostic performance evaluations and category 2 and 3 prognostic performance evaluations is assessed according to the design standards described in the preceding section of this chapter. Some studies may be found to have one or more serious design flaws which bring into question the validity of the evaluation findings. These studies should be considered unacceptable and they should be excluded from further analysis. Studies with less serious flaws may be grouped and analyzed separately or, if they are included in the quantitative meta-analysis, may have their contribution to the analysis weighted by some index of quality. No standard quality weighting index exists, to date. All well-designed performance evaluations should be included in the quantitative meta-analysis.

Quantitative meta-analysis attempts to combine the quantitative findings of performance evaluations in a way that will give a more complete and presumably more accurate representation of the performance of a laboratory study. This is done by combining the report results so as to generate aggregate performance data. How this is accomplished depends upon the presentation of the individual evaluation results. When complete performance data are available in the form of result frequency distributions, the frequency data can be combined so as to produce aggregate result frequency distributions from which aggregate ROC and likelihood ratio curves can be constructed. This is ideal. If result frequency distributions are not given but ROC curves are presented, ordinal regression methods can be used to model an aggregate ROC curve (Tosteson and Begg 1988). In addition, an aggregate likelihood ratio curve can be modeled using logistic regression techniques (Irwig 1992).

If each evaluation reports only one or a few sensitivity and specificity pairs or, in the case of a two-group prognostic study, only a few of the fraction correctly classified pairs, the findings from all the evaluations should be plotted together generating an aggregate ROC curve. The data can also be modeled to yield a summary ROC curve (Littenberg and Moses 1993, Irwig et al. 1994). Data pairs that lie at some distance from a fitted summary ROC curve are outliers. Explaining such outliers is an essential component of a quantitative meta-analysis. A thorough and systematic examination of the methods employed in the evaluations must be conducted to identify the methodological differences that resulted in outlying findings (Charlson *et al.* 1987).

Sometimes it is not possible to generate aggregate performance data as part of a meta-analysis because the reported findings are not consistent with the assumption of a shared underlying classification performance. In that case, methodological review of the evaluations should reveal the causes of the variability in the data. It can also happen that the assembled data do not combine in such a way as to yield a complete description of the diagnostic or predictive performance of a study. This happens, for instance, when the evaluations are concerned with study performance only in a restricted range, such as when describing the performance of a diagnostic study only at critical values for which the specificity is near 0.95. Then all that can be done is to average the data to produce a single performance pair—not one associated with a stipulated critical value but, rather, one associated with a predetermined value of one of the members of the pair.

REFERENCES

Begg CB and Greenes RA. 1983. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 39:207.

Charlson ME, Ales KL, Simon R, and MacKenzie R. 1987. Why predictive indexes perform less well in validation studies. Arch Intern Med 147:2155.

Choi BC. 1992. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. J Clin Epidemiol 45:581.

Concato J, Feinstein AR, and Holford TR. 1993. The risk of determining risk with multivariate models. Ann Intern Med 118:201.

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.

Diamond GA. 1989. Future imperfect: the limitations of clinical prediction models and the limits of clinical prediction. J Am Coll Cardiol 14:12A.

Dickersin K and Berlin JA. 1992. Meta-analysis: state of the science. Epidemiol Rev 14:154.

Fleiss JL. 1981. *Statistical Methods for Rates and Proportions*. 2nd edition. John Wiley and Sons, New York.

Gail MH and Green SB. 1976. A generalization of the one-sided two-sample Kolmogorov-Smirnov statistic for evaluating diagnostic tests. Biometrics 32:561.

Gray R, Begg CB, and Greenes RA. 1984. Construction of receiver operating characteristic curves when disease verification is subject to selection bias. Med Decis Making 4:151.

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.

Harrell FE, Lee KL, Matchar DB, and Reichert TA. 1985. Regression models for prognostic prediction: advantages, problems, and suggested solutions. Cancer Treat Rep 69:1071.

Hosmer DW, Hosmer T, Le Cessie S, and Lemeshow S. 1997. A comparison of goodness-of -fit tests for the logistic regression model. Stat Med 16:965.

Hosmer DW, Taber S, and Lemeshow S. 1991. The importance of assessing the fit of logistic regression models: a case study. Am J Public Health 81:1630.

Hui SL and Walter SD. 1980. Estimating the error rates of diagnostic tests. Biometrics 36:167.

Irwig L. 1992. Modelling result-specific likelihood ratios. (Letter). J Clin Epidemiol 45:1335.

Irwig L, Tosteson ANA, Gatsonis C, Lau J, Colditz G, Chalmers TC, and Mosteller F. 1994. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 120:667.

Jaeschke R, Guyatt G, and Sackett DL. 1994. III. How to use an article about a diagnostic test A. Are the results of the study valid? JAMA 271:389; B. What are the results and will they help me in caring for my patients? JAMA 271:703.

Jenicek M. 1989. Meta-analysis in medicine. J Clin Epidemiol 42:35.

Kraemer HC. 1992. *Evaluating Medical Tests*. Sage Publications, Newbury Park CA.

Lee TH and Goldman L. 1986. Serum enzyme assays in the diagnosis of acute myocardial infarction. Recommendations based on a quantitative analysis. Ann Intern Med 105:221.

Linnet K. 1988. A review of the methodology for assessing diagnostic tests. Clin Chem 34:1379.

Littenberg B and Moses LE. 1993. Estimating diagnostic accuracy from multiple conflicting reports. Med Decis Making 13:313.

Nierenberg AA and Feinstein AR. 1988. How to evaluate a diagnostic marker test. JAMA 259:1699.

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.

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.

Qu Y, Tan M, and Kutner MH. 1996. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. Biometrics 52:797.

Ransohoff DF and Feinstein AR. 1978. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 299:925.

Schulzer M, Anderson DR, and Drance SM. 1991. Sensitivity and specificity of a diagnostic test determined by repeated observations in the absence of an external standard. J Clin Epidemiol 44:1167.

Simon R and Altman DG. 1994. Statistical aspects of prognostic factor studies in oncology. Br J Cancer 69:979.

Staquet M, Rozencweig M, Lee YJ, and Muggia FM. 1981. Methodology for the assessment of new dichotomous diagnostic tests. J Chron Dis 34:599.

Torrance-Rynard VL and Walter SD. 1997. Effects of dependent errors in the assessment of diagnostic test performance. Stat Med 16:2157.

Tosteson AN and Begg CB. 1988. A general regression methodology for ROC curve estimation. Med Decis Making 8:204.

Walter SD and Irwig LM. 1988. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. J Clin Epidemiol 41:923.

Wasson JH, Sox HC, Neff RK, and Goldman L. 1985. Clinical prediction rules. Applications and methodological standards. N Engl J Med 313:793.

Werner M, Brooks SH, Mohrbacher RJ, and Wasserman AG. 1982. Diagnostic performance of enzymes in the discrimination of myocardial infarction. Clin Chem 28:1297.

White DB and James L. 1996. Standard error and sample size determination for estimation of probabilities based on a test variable. J Clin Epidemiol 4:419.

Yanagawa T and Gladen BC. 1984. Estimating disease rates from a diagnostic test. Am J Epidemiol 119:1015.

Yang I and Becker MP. 1997. Latent variable modeling of diagnostic accuracy. Biometrics 53:948.

Zhou XH. 1996. A nonparametric maximum likelihood estimator for the receiver operating characteristic curve area in the presence of verification bias. Biometics 52:299.