Evaluating the Yield of Medical Tests

Frank E. Harrell, Jr, PhD; Robert M. Califf, MD; David B. Pryor, MD;
Kerry L. Lee, PhD; Robert A. Rosati, MD

• A method is presented for evaluating the amount of information a medical test provides about individual patients. Emphasis is placed on the role of a test in the evaluation of patients with a chronic disease. In this context, the yield of a test is best interpreted by analyzing the prognostic information it furnishes. Information from the history, physical examination, and routine procedures should be used in assessing the yield of a new test. As an example, the method is applied to the use of the treadmill exercise test in evaluating the prognosis of patients with suspected coronary artery disease. The treadmill test is shown to provide surprisingly little prognostic information beyond that obtained from basic clinical measurements.

(*JAMA* 1982;247:2543-2546)

NEW MEDICAL tests have increased physicians' diagnostic capabilities, but they have also increased the cost of medical care. Information obtained from many tests is often unnecessarily redundant. To delineate the appropriate use of the expanding array of tests, a method of assessing the information yield of additional testing for an individual patient is required. This report outlines such a method, using an evaluation of the treadmill exercise test as an example.

Medical tests traditionally have been judged by their capability to establish a diagnosis. The management of a chronic disease, however, requires more information than is necessary for diagnosis alone. For example, the location and extent of ischemic heart disease and its effects on cardiac function and life-style is a dynamic process occurring over many years. When evaluating a medical test for patients with a chronic disease, the prognostic value of the test should be emphasized.

Physicians assimilate a variety of data from the history and physical examination to guide clinical decision-making. However, most studies of test utility have not explicitly considered such information because of the complexity of describing this assimilation process. This report emphasizes the importance of considering all prior relevant clinical information in determining the value of a new or additional test.

METHODS

Patient Population and Information Base

The computerized information system for patient data has been described previously.[1] Information on all patients undergoing cardiac catheterization at the Duke University Medical Center, Durham, NC, is collected prospectively. Table 1 lists the most important measurements obtained from the routine clinical evaluation, which form the baseline clinical assessment used in the analysis.

To determine the value of the treadmill test in the noninvasive setting, treadmill test descriptors were added to this basic set. To examine what useful information is provided by the treadmill test when the results of the cardiac catheterization are known, a selected set of important measurements from the cardiac catheterization (Table 2) were added to those from the initial patient assessment.

From 1969 to mid-1981, 5,461 consecutive patients have been catheterized for suspected ischemic heart disease at the Duke University Medical Center. This figure excludes patients with previous cardiac surgery, congenital heart disease, hypertrophic cardiomyopathy, or valvular heart disease other than mitral insufficiency thought to be secondary to ischemic heart disease. There were 3,833 medically treated patients, of whom 2,232 had treadmill exercise tests. Thirty-four patients were excluded because of uninterpretable exercise recordings due to conduction defects or baseline artifacts, and 33 patients were excluded because one or more descriptors listed in Tables 1 and 2 were not known. The remaining 2,165 patients constitute the study population. The criteria and protocol for exercise testing have been described in detail previously.[2] In general, all patients underwent exercise testing except those with physical disability, severe symptoms, or congestive heart failure. Of the 2,165 patients, 1,051

Table 1.—Basic Clinical Descriptors

- History
  - Age
  - Sex
  - Type of chest pain (typical angina, atypical angina, nonanginal)
  - Nocturnal chest pain
  - Progressive chest pain
  - Preinfarctional angina
  - Chest pain frequency per week
  - Congestive heart failure severity (New York Heart Association classification)
  - History of myocardial infarction
  - History of peripheral vascular disease
  - History of cerebral vascular disease
  - History of diabetes
  - History of smoking
  - Family history of ischemic heart disease
  - Serum cholesterol value
- Physical Examination
  - Systolic BP
  - Ventricular gallop
- Chest X-ray Film
  - Cardiomegaly
- ECG
  - Significant Q waves
  - Resting ST-T wave abnormalities
  - Resting premature ventricular contractions

From the Department of Community and Family Medicine, Division of Biometry (Drs Harrell, Lee, and Rosati), and the Department of Medicine, Division of Cardiology (Drs Califf and Pryor), Duke University Medical Center, Durham, NC.

Reprint requests to PO Box 3337, Duke University Medical Center, Durham, NC 27710 (Dr Rosati).

Table 2.—Catheterization Descriptors

Descriptors of Left Ventricular Performance
Diffuse abnormal contraction
Aneurysm
Anterior asyneresis
Inferior asyneresis
Apical asyneresis
Ejection fraction
End-diastolic pressure
Mitral insufficiency (graded 0-4)
Coronary Stenosis—Maximum % of Diameter Narrowing
Left main coronary artery
Left anterior descending coronary artery
Left circumflex coronary artery
Right coronary artery

were eventually found to have significant coronary disease (ie, diameter narrowing of at least 75% in one or more major arteries) by cardiac catheterization. Follow-up has been obtained at six months, at one year, and at yearly intervals thereafter; follow-up is 99% complete.

Exercise Protocol

The graded multistage protocol of Bruce[3] was used. The criteria for a positive ST-segment interpretation were (1) horizontal or downsloping ST-segment change of 0.1 mV or greater when no pretest ST abnormalities of at least 0.05 mV were present; (2) ST-segment elevation of 0.1 mV or more than the control tracing in any lead except aV_R; (3) additional horizontal or downsloping ST depression of 0.2 mV or greater with preexisting ST depression of at least 0.05 mV. The criteria for a negative result were failure to meet positive criteria in addition to the achievement of 85% of the maximum predicted heart rate.[4] Patients who did not meet positive criteria and who failed to reach 85% of the maximum predicted heart rate were considered to have indeterminate test results. Results for patients with baseline ST-segment depression greater than 0.05 mV who had additional ST-segment depression of at least 0.1 mV but not 0.2 mV were also considered indeterminate.

Method of Evaluating a Medical Test

The value of a test for prognosis or diagnosis should be evaluated in a four-step process.

Step 1: Assessing Prior Information.—Simply relating a test result by itself to prognosis is not an adequate analysis of the utility of a test. The first step in the analysis should be to gather all information about the patient that will be available before the test is ordered. In the evaluation of the cardiac patient, the history, physical examination results, resting ECG, and standard chest x-ray film are almost always obtained before additional testing. Since additional tests add more expense and risk, they will be useful only if they contribute new information.

Step 2: Coding Test Results.—All pertinent measurements from the test being evaluated should be used in assessing the information provided by the test. Regression methods discussed below are useful in evaluating and combining multiple test measurements. Whenever possible, the main test result should be quantified rather than classified as "positive" or "negative."

Step 3: Testing for Additive Information in a Test.—For an additional test to be clinically useful, it must provide *statistically* significant information that is independent of *all* information already obtained, assuming the analysis is based on an adequate number of patients for the type II error to be small. Clinical significance will be addressed later, but if statistical significance is not obtained, the question of clinical significance is a moot point. When considering prognosis, the Cox[5] regression model (see "Statistical Calculations") can assess the usefulness of the information available prior to a test in predicting a patient's survival time and determine if the accuracy of the prediction is improved by the new test. (Survival time is the number of days a patient lives after the baseline clinical evaluation.)

Step 4: Quantifying Additional Information.—Because statistical significance is only a prerequisite for clinical significance, we desire to know if the information added by the test is useful to the physician. In a traditional statistical assessment, the test variables are evaluated to see if they provide *any* predictive information. But we need to devise a relevant measure of *how much* prognostic information the test actually provides. To obtain such a measure, we propose calculating an index of correlation (described subsequently) between the baseline routine clinical data and the outcome (length of survival). Then, by examining the increment in this correlation index when information from the test is added to the basic set of clinical descriptors, we can quantify the additional prognostic information provided by the test.

An alternate method for assessing the added prognostic information from a new procedure provides a different perspective on the problem. This method consists of quantifying how much the results of the procedure change the best estimates. of prognosis, rightly or wrongly, for individual patients. Two Cox regression models are developed to predict the probability that each patient will survive for five years from the date of evaluation; one model makes use of the additional information from the test, and one does not. For example, suppose that the predicted probability of surviving five years is .75 for a certain patient. If the posttest model predicts a survival probability of .71 after a positive test result is known, the positive test actually lowered the prognostic estimate very little. To summarize how the test result changed prognostic estimates, the percentage of patients for which the test information changed the five-year probability of survival by more than .10 will be presented. Although this method of quantifying added prognostic information is simple to apply, it does have drawbacks. It does not tie prognostic predictions to individual patients' actual outcomes, and it does not actually quantify how well any given set of information can actually predict the outcome. It also requires one to make a choice as to the time period and change in probability of interest.

Statistical Calculations

The Cox regression model[5] is a generalization of the multiple linear regression model in which the response variable usually represents the time until some event, and its value for some subjects can be censored, ie, for some patients still alive, the survival time is only known to exceed the current follow-up time. For the analyses presented in this article, only cardiovascular deaths are considered as events. For patients suffering noncardiovascular deaths, the follow-up time is censored at the time of the nonrelated death.

Any set of clinical measurements for a patient can be denoted by $X_1, X_2, \ldots X_p$. The Cox model assumes that the probability that the patient survives past time t is given by So(t) raised to the exp $(B_1X_1+B_2X_2+\ldots+B_pX_p)$ power, where So(t) is the survival probability for a "standard" patient, B_i is the coefficient or weight for the ith descriptor variable, and exp is the exponential function. A "standard" patient is a hypothetical one for whom each descriptor variable has a value equal to the mean for that descriptor variable over the entire patient sample. Using the model, the set of weights $B_1, \ldots B_p$ are estimated from the data and a prognostic score $PS=B_1X_1+B_2X_2+\ldots+B_pX_p$ can be calculated so that PS is optimally related to survival time. The standard survival function So(t) is also estimated from the data, and no particular shape is assumed for this function. Using the estimate for So(t) in conjunction with the prognostic score, the probability of surviving until time t can be estimated for each patient.

The prognostic score arising from the Cox model captures and condenses the information in any given set of measurements. The predictive value of a set of patient measurements can be quantified by computing a measure of correlation between survival time and the prognostic score based on that set. A complication arises because only 188 of the 2,165 patients have had cardiovascular deaths to date. The survival times of the remaining 1,977 patients are censored. An index of correlation like the commonly used linear correlation coefficient does not allow for censored values and also makes the restrictive assumption that survival time can be predicted by a purely linear equation.

We chose a measure of correlation that would have a simple, clinically meaningful interpretation, yet still be applicable in follow-up studies. The measure is similar to that of Brown et al.[6] Draw a pair of patients and determine which patient lived longer from his baseline evaluation. Survival times can be validly compared either when both patients have died, or when one has died and the other's follow-up time has exceeded the survival time of the first. If both patients are still alive, which will live longer is not known, and that pair of patients is not used in the analysis. Otherwise, it can be determined whether the patient with the higher prognostic score (ie, the weighted combination of baseline and test variables used to predict survival) also had the longer survival time. The process is repeated until all possible pairs of patients have been examined. Of the pairs of patients for which the ordering of survival times could be inferred, the fraction of pairs such that the patient with the higher score had the longer survival time will be denoted by c.

The index c estimates the probability that, of two randomly chosen patients, the patient with the higher prognostic score will outlive the patient with the lower prognostic score. Values of c near .5 indicate that the prognostic score is no better than a coin-flip in determining which patient will live longer. Values of c near 0 or 1 indicate the baseline data virtually always determine which patient has a better prognosis. The c index measures a probability; many clinicians are more used to dealing with a correlation index that ranges from −1 to +1. A Kendall or Goodman-Kruskal type of correlation index[7] can easily be constructed by calculating $\gamma=2(c-.5)$, where γ is the estimated probability that the prognostic score correctly orders prognosis for a pair of patients minus the probability that it incorrectly orders prognosis. When the prognostic score is unrelated to survival time, γ is zero. When $\gamma=.5$, the relationship between the prognostic score and survival time is halfway between a random relationship and a perfect relationship, and the corresponding c value is .75.

Evaluation of Treadmill Exercise Test

Step 1: Summarizing Routine Clinical Information.—When routine noninvasive descriptors are combined to form a prognostic score, the index c is .82 and γ is .63. In other words, one could correctly determine which of two patients would live longer 82% of the time based on routine data. If, in addition, invasive measurements from a cardiac catheterization such as ventricular contraction patterns and degree of narrowing in the major coronary arteries were also incorporated into the score, the index of prognostic information is c=.87 with γ=.74, meaning that one can correctly order survival times 87% of the time. The addition of information obtained by catheterization changed five-year prognostic estimates by more than 0.10 in 12.8% of the patients.

Step 2: Coding Exercise Test Information.—Ideally, the exercise test should be coded for the degree of ST-segment abnormality, the duration of exercise, the maximum heart rate achieved, the occurrence of angina during stress, the frequency and type of ventricular arrhythmias, and changes in BP. Because of the difficulty in reliably coding the degree of ST-segment change recorded by different observers, and because rhythm and BP data have been recorded in the computed data bank only since 1974, only the qualitative ST-segment interpretation, duration of exercise, heart rate, and occurrence of angina were used in this analysis. The following discussion describes how the available information was coded.

Clinically, we thought that with regard to prognosis, an indeterminate test result should be interpreted as halfway between a negative and a positive result (this assumption was verified statistically). The other measurements of major interest were the duration of exercise, which was coded as the number of minutes of exercise, and the maximum heart rate. A positive result should be associated with worse prognosis if the ECG changes occur earlier or at a lower heart rate. The most abnormal results should have a significant ST-segment change at the beginning of exercise. Thus, six terms were evaluated: ST-segment interpretation (coded as 0, 1, 2 for negative, indeterminate, and positive), duration of exercise, maximum heart rate, presence of angina during exercise, the product of interpretation and duration, and the product of interpretation and heart rate (to allow for the added importance of a positive test if the duration, heart rate, or both, were less).

Step 3: Testing for Additive Information in Exercise Test.—Using the Cox regression model, the six treadmill variables were found to add significant independent prognostic information to the other noninvasive clinical descriptors (χ^2=33.8 with 6 *df*, $P<.0001$). The most important treadmill variables were the interpretation and the product of interpretation and duration. The χ^2 statistic does not quantify the actual clinical utility of treadmill information, although one can gain perspective by comparing 33.8, the extra χ^2 contributed by the treadmill variables, to the χ^2 value of 257.4 due to all other basic noninvasive clinical descriptors.

Step 4: Quantifying Information Added by Exercise Test.—As stated before, the c value based on routine noninvasive data is .82. Adding treadmill information raises c to .83 (the values for γ are .63 and .67, respectively). This finding indicates that with treadmill data, one can determine correctly which of two patients will live longer 83% of the time. Although the additional information provided by the treadmill is statistically significant, the question of whether the test is clinically useful when it adds only 1% to the ability to correctly order patients' prognoses must be raised. The incorporation of treadmill test results changed five-year survival probability estimates by more than .10 in only 6.8% of patients. Consistent with these findings, the treadmill test was also found to provide less prognostic information by itself (c=.73) than the basic measurements provided alone (c=.82).

Assessing Yield of Exercise Test for Catheterized Patients

When invasive information is known, a Cox test of the additional prognostic information of the exercise test yields χ^2=4.7 with 6 *df* (P=.58). In other words, if a patient is going to undergo catheterization, the exercise test provides no additional prognostic information. For illustration, however, we could still calculate the indices of test utility. Inclusion of treadmill information means that one can correctly order patients' prognoses 87.1% of the time rather than 86.8% of the time; γ is increased from .735 to .743. Treadmill test information changed five-year prognostic estimates by more than .10 in only one of the 2,165 patients.

COMMENT

The foregoing analysis demonstrates that if one utilizes all previously obtained noninvasive findings, the exercise test provides some additional prognostic information. However, when the amount of additional information is quantified, it is found to be small. After a patient is catheterized, the incorporation of exercise treadmill results adds no measurable amount of prognostic information. Thus, the use of exercise test results in making clinical decisions may actually be detrimental to a patient. For example, if a patient with a less severe form of coronary disease and good ventricular function has an early positive treadmill test result (significant ST-segment deviations occurring in the first six minutes of exercise), the patient still has an excellent prognosis. In the 310 medically treated patients with one- or two-vessel disease and normal left ventricular function without an early positive test result, the survival rate is 95% at four years. However, in the 57 patients with one- or two-vessel disease and normal ventricular function having an early positive test result, the four-year survival is 96%. The physician may be unduly concerned about the patient having an early positive test result, not realizing that once coronary anatomy and ventricular function are described, the treadmill exercise test results are of no value in assessing prognosis.

The limitations of the exercise test in relation to diagnosis of coronary disease have been described.[8] No previous work has outlined the limitations of the exercise test with regard to prognosis in the noninvasive evaluation. Certain cautions must be observed in interpreting these results. The addition of the degree of ST-segment change,[9] the occurrence of ventricular arrhythmias,[10] changes in BP with exercise,[11] or alternate measures of exercise workloads may make the exercise test a more potent predictor of outcome. In our opinion, the incremental gain in prognostic information from these measurements will be small. For example, a recent study[12] described a group of men with suspected coronary disease who had excellent prognoses despite the occurrence of profound ST changes with exercise. Of course, exercise testing may be used for purposes other than estimating prognosis, such as evaluating work capacity or obtaining objective evidence of ischemia for patients in whom the history is unclear.

Another potential problem affecting our analyses is that the determination of the type of angina can sometimes be influenced by previous tests the patient may have had. To address this problem, we repeated all analyses without using the type of angina in the baseline clinical assessment. The c indices changed only in the third decimal place. We also repeated the analyses, including only those patients found to have significant coronary disease at catheterization, to evaluate the utility of the treadmill test in assessing prognosis in a more tightly defined patient group. No substantive changes resulted.

When diagnosis is considered as a binary end point—the presence or absence of disease—patients can be lumped conveniently into two categories. If outcomes are known, the average prognosis of each patient with the diagnostic label can be determined. However, this information is of limited value to the physician caring for a particular patient with a complex chronic disease. Patients with the same disease will have markedly different prognoses based on individual characteristics. For example, in ischemic heart disease, patients will have different prognoses based on left ventricular function, anatomy, and other information. A patient with stable angina, three-vessel coronary disease, and normal left ventricular function has a two-year survival of 93%, while a patient with progressive angina, three-vessel disease, and abnormal ventricular function has a two-year survival of 68%.[13] Both patients have coronary artery disease (even of the same severity), yet the degree of illness is markedly different. Once prognosis becomes a central focus and the variety of outcomes is realized, the concept of simple categorization of a disease state changes. Coronary disease is usually thought of as a common disease, but because of its wide prognostic spectrum, patients cannot readily be categorized.

We have demonstrated the use of the proposed method to determine the prognostic value of the treadmill exercise test for individual patients evaluated for coronary disease. The same method can be used to assess the independent value of any new test in patients with any chronic disease and can be used to assess the incremental value of each test in a battery of tests. No longer should tests be considered in the absence of the clinical information available to the physician at the time tests are ordered. Diagnosis is important, but most patients and physicians ultimately must be more concerned with prognosis. Further insights from clinical knowledge and developments in statistical techniques should allow continual improvements in the evaluation of medical tests.

This work was supported by grant HS 03834 from the National Center for Health Services Research and the National Center for Health Care Technology, Hyattsville, Md; research grant HL-17670 from the National Heart, Lung, and Blood Institute, Bethesda, Md; training grant LM 07003 and grants LM 03373 and LM 00042 from the National Library of Medicine, Bethesda, Md; and grants from the Prudential Insurance Company of America, Newark, NJ, the Kaiser Family Foundation, Palo Alto, Calif, and the Andrew W. Mellon Foundation, New York.

Eugene A. Stead, Jr, MD, offered us valuable insights and ideas.

References

1. Rosati RA, McNeer JF, Starmer CF, et al: A new information system for medical practice. *Arch Intern Med* 1975;135:1017-1024.

2. McNeer JF, Margolis JR, Lee KL, et al: The role of the exercise test in the evaluation of patients for ischemic heart disease. *Circulation* 1978;57:64-70.

3. Bruce RA: Exercise testing of patients with coronary heart disease. *Ann Clin Res* 1971; 3:323-332.

4. Robinson S: Experimental studies of physical fitness. *Arbeitsphysiologie* 1938;10:251-323.

5. Cox DR: Regression models and life tables. *J R Stat Soc* 1972;34:187-202.

6. Brown BW, Hollander M, Korwar RM: Nonparametric tests of independence for censored data, with applications to heart transplant studies, in Proschan F, Sarfling RJ (eds): *Reliability and Biometry.* Philadelphia, Social and Industrial Applied Mathematics, 1974, pp 327-354.

7. Kendall DG: *Rank Correlation Methods*, ed 3. London, Charles Griffin & Co, 1962, chap 1.

8. Fisher LD, Kennedy JW, Chaitman BR, et al: Diagnostic quantification of CASS (Coronary Artery Surgery Study): Clinical and exercise test results in determining presence and extent of coronary artery disease. *Circulation* 1981;63:987-1000.

9. Diamond GA, Hirsch M, Forrester JS, et al: Applications of information theory to clinical diagnostic testing. *Circulation* 1981;63:915-921.

10. Udall JA, Ellestad MH: Predictive implications of ventricular premature contractions associated with treadmill exercise testing. *Circulation* 1977;56:985-989.

11. Morris SN, Phillips JF, Jordan JW, et al: Incidence and significance of decreases in systolic blood pressure during graded treadmill exercise testing. *Am J Cardiol* 1978;41:221-226.

12. Podrid PJ, Graboys TB, Lown B: Prognosis of medically treated patients with coronary-artery disease with profound ST-segment depression during exercise testing. *N Engl J Med* 1981;305:1111-1116.

13. Harris PJ, Harrell FE, Lee KL, et al: Survival in medically treated coronary artery disease. *Circulation* 1979;60:1259-1269.