Better Decisions Through Science—Case in Point: Exercise Testing Scores

Euan Ashley, MRCP and Victor Froelicher, MD, Department of Cardiovascular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK and the Cardiology Division, Veterans Affairs Palo Alto Health Care System, Stanford University, Palo Alto, CA

The application of common statistical techniques to clinical and exercise test data has the potential to become a useful tool for assisting in the diagnosis of coronary artery disease, assessing prognosis and reducing the cost of evaluating patients with suspected coronary disease. Since generalists increasingly function as gatekeepers and decide which patients must be referred to the cardiologist, they need to make optimal use of the basic tools they have available: history, physical exam and the exercise test. Scores derived from multivariable statistical techniques considering clinical and exercise data have demonstrated superior discriminatory power when compared with simple classification of the ST response. In addition, by stratifying patients as to probability of disease and prognosis, they provide a management strategy. While computers, as part of information management systems, can run complicated equations and derive these scores, physicians are reluctant to use them. Thus, scores have been represented as nomograms or simple additive tables so physicians are comfortable with their application. Their results have also been compared to physician judgment and found to estimate the presence of coronary disease and prognosis as well as expert cardiologists and often better than non-specialists. Scores can empower the clinician to assure the cardiac patient access to appropriate and cost-effective cardiological care.

Coronary artery disease continues to be the leading cause of morbidity and mortality in the United States and the prevalence is expected to increase due to the increasing proportion of the population that is elderly. Meanwhile, in spite of efforts to control costs, health care spending increased by a greater amount in 1999 than in any other year of the last decade. With half of the increases explained by pharmaceuticals that can decrease heart disease interventions and events, the next target of cost containment must be expensive diagnostics and interventions. Small improvements in our ability to pick out those likely to have disease or benefit from therapy can translate into enormous cost savings when implemented population wide. This makes it important to implement clinically cost-effective strategies that direct the appropriate patients to the optimal procedures through clinical risk prediction. There is a growing awareness of the need to apply statistical techniques to improve decision-making. Here we provide by example a blueprint for this process.

Diagnosis: Gathering the Data

The first step in considering a diagnostic test is an evaluation of its validity. Critical to this process is that only consecutive, non-diagnosed patients from a representative population are used to evaluate the test or score. In addition, patients must agree at the outset to undergo both the diagnostic test in question and the gold standard (the choice of gold standard is clearly a separate issue). Administering the gold standard test (here, angiography) only to those positive for the test in question (the exercise test) creates "work-up bias." Another error, "limited challenge" (Table 1) occurs when a test is evaluated by comparing patients with severe disease to apparently healthy individuals.

Developing the Tool

The next step is to convert the raw data into a useful clinical prediction tool. Theoretical considerations lead investigators to focus on the variables most likely to predict the result or outcome. These variables are then tested using mathematical techniques. Regression analysis is especially attractive, since it makes possible the derivation of complex regression functions directly from a database. Logistic regression has been preferred since it models the relationship to a sigmoid curve (which is often the mathematical relationship between a probability variable and an outcome) and its output is between zero and one, this representing the probability of disease being present. The variables then found to have discriminating power are combined to form an algorithm for estimating the probability of disease.

An initial evaluation of a score of measurement can be made by graphing how much the score differs among those with and without the disease. These measurements could be ST segment depression, calcium score from electron beam computed tomography perfusion scan values or echocardiographic wall motion estimates. Figure 1 consists of actual data from over 1000 male veterans who underwent both exercise testing and coronary angiography. As illustrated in the figure, the values for the score, for those with and without disease, usually greatly overlap. The cut-point of 50 is a practical choice for the treadmill score so that those above 50 are considered to have disease and those below are considered free of CAD. However, as can be seen, this is not really the case. Figure 2 separately considers the two curves with the cut-point permitting calculation of sensitivity (bottom curve of population distribution) and specificity (top curve).

Score Evaluation (ROC Curves)

The ability of the model to separate is assessed formally by means of a receiver-operating-characteristic (ROC) curve. For a given score or measurement, several possible cut-off

1062-1458/02/$22.00
PII S1062-1458(02)00598-6

Table 1. Pre-Test Probability of Coronary Disease by Symptoms, Gender and Age

Age	Gender	Typical/Definite Angina Pectoris	Atypical/Probable Angina Pectoris	Non-Anginal Chest Pain	Asymptomatic
30–39	Males	Intermediate	Intermediate	Low (<10%)	Very low (<5%)
	Females	Intermediate	Very low (<5%)	Very low	Very low
40–49	Males	High	Intermediate	Intermediate	Low
	Females	Intermediate	Low	Very low	Very low
50–59	Males	High (>90%)	Intermediate	Intermediate	Low
	Females	Intermediate	Intermediate	Low	Very Low
60–69	Males	High	Intermediate	Intermediate	Low
	Females	High	Intermediate	Intermediate	Low

There are no data for patients younger than 30 or older than 69, but it can be assumed that coronary artery disease prevalence increases with age. High = >90%; intermediate = 10–90%; low = <10%; very low = <5%.

criteria could be used to separate results into positive and negative groups. However, the sensitivity and specificity will be different for each criterion. The ROC curve is a plot of the sensitivity against specificity for the full range of possible cutpoints of a score. The area under the curve ranges from 0 to 1, with 0.5 corresponding to no discrimination (random performance) and 1.0 to perfect discrimination. The shape of the curve demonstrates the inverse relationship between sensitivity and specificity at different cutpoints. Figure 3 is a ROC plot of our simple treadmill score ranging from 0 to 100 illustrating two other cutpoints, 40 and 60. In certain circumstances, different cutpoints could be appropriate. For example, a high specificity is needed when screening healthy people, whereas a high sensitivity is required for ruling out ischemia after presentation for chest pain. Figure 4 shows a comparison of the characteristics of four diagnostic tests for coronary artery disease: the Morise pre-test clinical score, ST analysis measured visually or by computer and our simple treadmill score. The four curves allow for a comparison of the diagnostic value of these techniques. The treadmill test clearly adds to the discriminatory value of clinical data while computer analysis is similar to expert visual analysis of the ST segments.

Combining Pre-Test and Test Data

It is common following development of a new test to assess it in isolation, overlooking the basic information available in history and examination. Pre-test classification is, however, an important part of the work up. The classification of pretest probability is enabled through a table considering age, gender and chest pain characteristics using the Diamond-Forrester tabular method (Table 1). Morise et al. proposed a pre-test score for categorizing patients with suspected coronary disease and normal resting electrocardiograms that is possibly superior to the method advocated by the guidelines. We have validated this score in a large sample of male veterans.

Regardless of the method used, it has long been known that combining clinical and exercise parameters along with the ST responses can improve the accuracy of the test. As a result, many clinical investigators have published studies proposing multivariable equations to enhance the accuracy of the standard exercise test. We reviewed 24 studies attempting to predict the presence of any angiographic disease and listed the 30 equations created. Despite methodological shortcomings, it was clear that scores more accurately identified those with disease than ST diagnostic criteria alone.

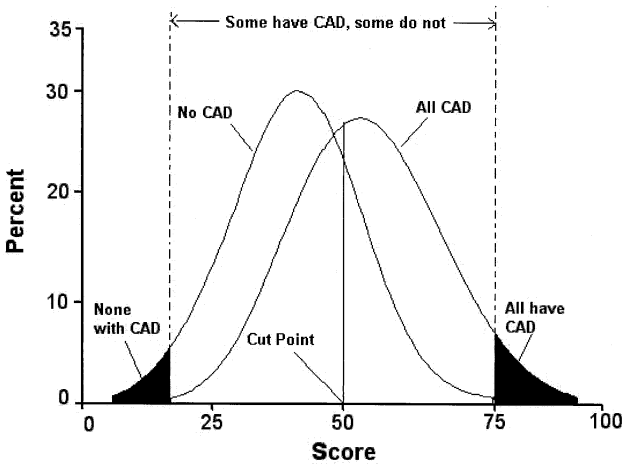


Fig. 1. Range of characteristics plots for the simple treadmill score for those with and those without angiographic coronary disease.

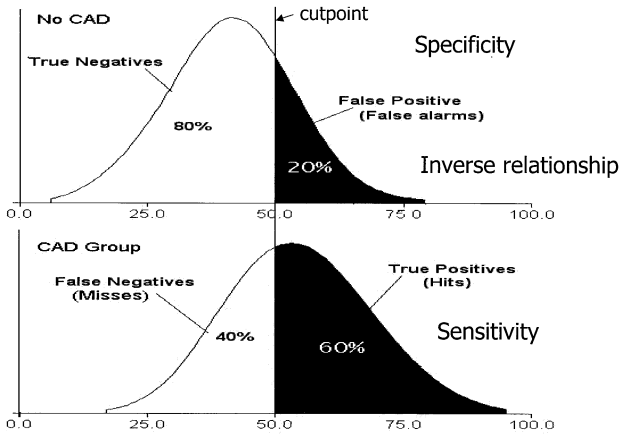


Fig. 2. Separate frequency plots indicating the four test responses that enable calculation of test characteristics (true positives, true negatives, false positives, false negatives).

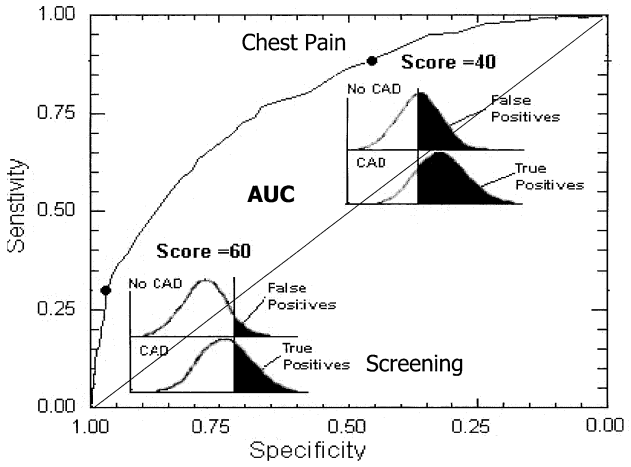


Fig. 3. Range of characteristics plot of the simple treadmill score showing how different cut points can be chosen according to the specific use of the test.

Consensus of Scores

One potential drawback of diagnostic scores is their reliance on the uniformity of the populations in which they were developed and tested. In an attempt to make scores portable to different populations, we considered a consensus approach. This is a strategy adopted in diverse fields including space travel that involves using several equations validated in different populations. The equations are assessed together and a result is valued if consensus between scores is achieved. We reasoned that if it works for spacecraft trajectories, then it could work for coronary disease. We used the Detrano and Morise equations along with our own equation. A probability score was calculated for each patient using the three equations. Thresholds were set in each equation and if a patient was "high probability" in at least two of the three then he or she was considered to have a high probability of disease, similarly if low in at least two of the equations where the label was "low risk." All others would be intermediate. Since the patients in the intermediate group would be sent for further testing and would eventually be correctly classified, the sensitivity of the consensus approach was 94% and specificity was 92%. The percent of correct diagnoses increased from 67% for standard exercise ECG analysis and from 77% for multivariable predictive equations alone to greater than 90% correct diagnoses for the consensus approach. This was a significant success and compares favorably with the best (much more expensive) tests in cardiac medicine. This approach, however, requires a computer program and, despite the increasing number of physicians carrying handheld devices capable of carrying out these calculations, this can limit its clinical application.

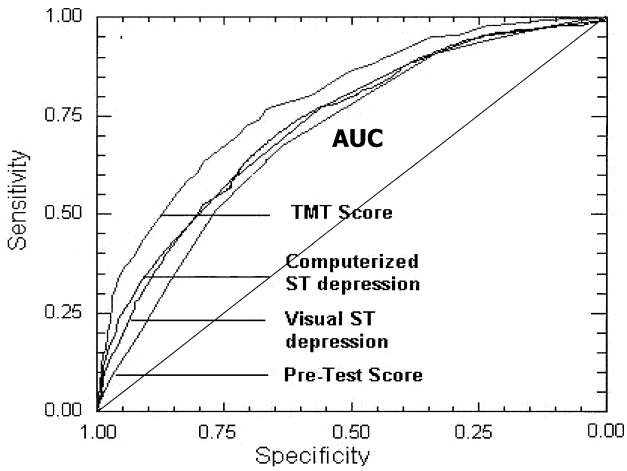


Fig. 4. Range of characteristics plots comparing the discriminating power of a pre-test score, ST measurements and a simple treadmill test (TMT) score.

Variable	Circle response	Sum
Age	Men<40, Women<50 = 3	
	Men40-55, Wom50-65=6	
	Men>55, Women>65 =9	
Estrogen Status	Positive=-3	
	Negative=+3	
Diabetes	Yes = 2	
Obesity?	Yes = 1	
Family History?	Yes = 1	
Hypercholesterolemia?	Yes = 1	
HBP?	Yes = 1	
Smoking?	Yes = 1	
	Total Score	

Choose only one per group

<=8 low prob

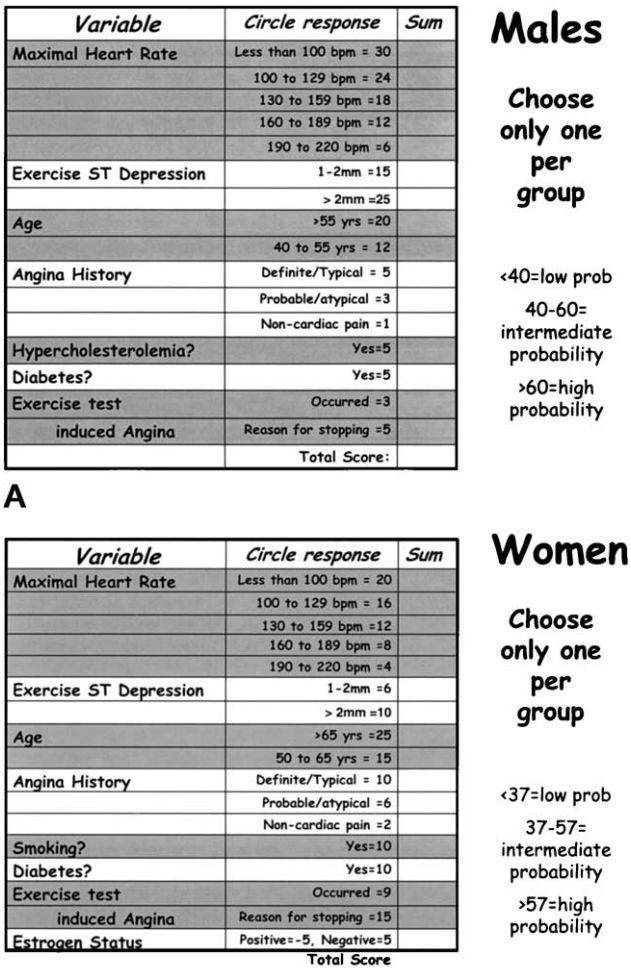
9-15 = intermediate probability

>=16 high probability

Fig. 5. Calculation of the simple clinical score for angiographic coronary disease.

"Simplified" Score Derivation

Simplified scores derived from multivariable equations have been developed for pre-test estimates of disease and for prognosis. They require physicians only to add points, and as such are available at the point of care without recourse to technology (Figures 5 and 6). To decrease the complexity of the predictive equations, we used the variables chosen in logistic regression to derive a simple linear score. We first coded all variables with the same number of intervals so that the coefficients would be proportional. Then we coded the bin with the larger value to associate it with higher probability of disease. For instance, if 5 is the chosen interval, dichotomous variables are 0 if not present and 5 if present and continuous variables like age and heart rate are coded in 5 bins by ranges. All codes are then directly related to probability and the smallest coefficient is associated with the least important variable. The other coefficients were set to their proportional weight or importance by dividing each coefficient by the smallest. This made the relative importance of the selected variables obvious. This approach results in a very simple linear score in which the health care provider merely compiles the variables, multiplies by the appropriate number and then adds up the products. Surprisingly, these simple linear scores have the same ROC areas as the more complicated equations requiring the calculation of exponentials.



Variable	Circle response	Sum
Maximal Heart Rate	Less than 100 bpm = 30	
	100 to 129 bpm = 24	
	130 to 159 bpm =18	
	160 to 189 bpm =12	
	190 to 220 bpm =6	
Exercise ST Depression	1-2mm =15	
	> 2mm =25	
Age	>55 yrs =20	
	40 to 55 yrs = 12	
Angina History	Definite/Typical = 5	
	Probable/atypical =3	
	Non-cardiac pain =1	
Hypercholesterolemia?	Yes=5	
Diabetes?	Yes=5	
Exercise test induced Angina	Occurred =3	
	Reason for stopping =5	
	Total Score:	

Fig. 6. **(A)** Calculation of the simple score for angiographic coronary disease in men. **(B)** Calculation of the simple score for angiographic coronary disease in women.

Management Strategy

Diagnostic scores, in giving more sophisticated estimates of likelihood of disease, also allow a more sophisticated management strategy (Table 2). Specifically, the more detailed information provided and the greater confidence a practitioner can have in prediction allows patients to be placed in categories of risk, rather than being limited by a simple positive or negative dichotomy. For coronary disease, patients graded as low risk would need no further testing at that time, high-risk patients would need an invasive study and intermediate-risk patients would require another noninvasive study.

Prognosis

For diagnostic tests, the choice of gold standard can be difficult. Specific to coronary disease, some practitioners have suggested that angiography is so poor at predicting unstable disease that diagnosis is secondary to the prediction of prognosis. Here also, statistical tools help identify the patients with most to gain from intervention.

Table 2. Paradigm for the Clinical Reaction to the Estimated Probability of CAD

Probability for clinically significant CAD (≥50% occlusion)	
Low probability	Patient reassured symptoms most likely not due to CAD
Intermediate probability	Require other tests such as stress echo, nuclear, or angiography to clarify diagnosis; anti-anginal medications tried
High probability	Anti-anginal treatment indicated; intervention clinically appropriate; angiography may be required

Nine studies have incorporated multiple exercise variables into simple prognostic scores without the use of complex regression formulas. Table 3 lists the number of times the major prognostic variables were chosen as significantly and independently predictive of time to death in the published prognostic studies. The most widely used prognostic score is the Duke Treadmill score since it can also be used for diagnosis. This score has been validated in other populations including women and when the resting ECG exhibits ST depression. The nomogram for the Duke Treadmill score can be found in all of the exercise test guidelines.

Some methodological points are specific to prognosis such as which mortality index to use. Consideration of all-cause mortality instead of cardiovascular mortality may explain why the ischemic variables included in the Duke score that clearly had diagnostic power do not predict death. While all-cause mortality has advantages over cardiovascular mortality as an end point, the Duke score was generated using the end points of infarction and cardiovascular death. The use of interventions as end points falsely strengthens the association of ischemic variables with end points. While some investigators have justified their use by requiring a time period to expire after the test before using the intervention/procedure as an end point, this still influences the associations between test responses and end

Table 3. Frequency of Clinical and Exercise Test Variables Chosen as Significantly and Independently Associated With Time Until Death in Nine Previous Prognositic Studies

Variable	Out of Nine Studies
Clinical	
Age	2
CHF	2
MI by history or Q waves	1
Resting ST depression	1
Exercise responses	
Exercise capacity (METs)	7
Angina	5
ST depression	4
Maximal heart rate	3
Maximal SBP	2
ST elevation	1
PVCs	1
Maximal double product	1

points. In addition, the relative importance of ischemic variables can be minimized by not censoring on interventions for ischemia (removing intervened patients from observation when the intervention occurs in follow up).

Previous prognostic studies focused on specific subsets of patients. We analyzed all patients referred for evaluation at our exercise lab between 1987 and 2000 in order to develop a prognostic score. There were over 6000 patients who had standard exercise ECG treadmill tests over the study period, with a mean 6-year follow up. Twenty percent died over the follow-up period and the average annual mortality was 2.6%. We used the Cox hazard function which ranked the following variables in order as independently and significantly associated with time to death: METs less than 5, age greater than 65, a history of congestive heart failure and a history of myocardial infarction or presence of a diagnostic Q wave. The simplified prognostic score, derived by simply adding these variables, classified patients into three risk groups as shown in Figure 7.

Comparing Scores and Physicians

Though scores based on exercise testing data have been advocated for years, only three previous studies have compared them to physician estimates of disease. The first study derived a score for estimating probabilities of significant and severe coronary disease and then validated and compared it with the assessments of cardiologists. The score performed at least as well as the clinicians when the latter knew the identity of the patients. The clinicians were more accurate when they did not know the identity of the subjects but worked from tabulated objective data. A second study validated two scores by comparing their diagnostic accuracy to that of cardiologists. The scores outperformed the cardiologists. A third study considered scores for prognosis (rather than diagnosis) with 100 patients sent to five senior cardiologists at one center. Again the scores outperformed the cardiologists.

We performed a study that was larger and included different groups of physicians, once again showing that scores can predict angiographic results and prognosis as well as physicians. Clinical/treadmill test reports were sent to expert cardiologists and to two other groups including randomly selected cardiologists and internists who classified them as high, low or intermediate probability of disease in addition to estimating a numerical probability from 0 to 100%. Over 150 physicians returned over 600 patient evaluations. When probability estimates were compared, the scores were superior to all the physician groups. In a subsequent analysis, we found the scores to predict prognosis as well or better than physicians.

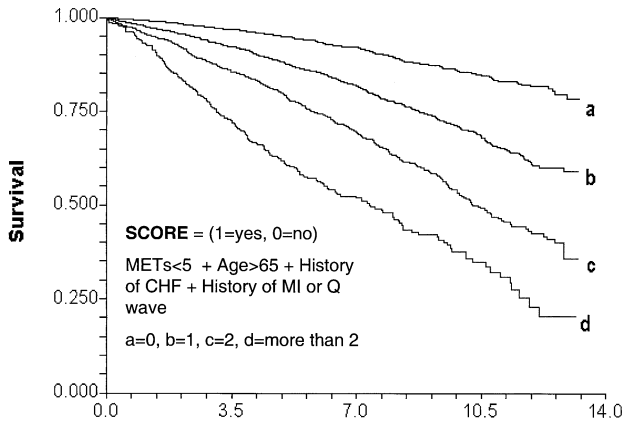


Fig. 7. Kaplan-Meier Survival curves for the "all-comers" prognostic score.

Conclusions

Physicians should not reduce their diagnostic assessments to blindly using and memorizing prediction rules. However, in spite of the methodological limitations of the available studies, the scores make possible better decisions. Statistical approaches cannot make counterintuitive leaps of tangential thinking but can excel at that which humans do not: considering vast quantities of information perfectly, then categorizing and analyzing it without bias. Making use of statistics as we described gives clinicians a powerful second opinion and allows them to concentrate on what the computer can never do: look after patients as individuals. In particular, scores make available the experience of the specialist clinician to generalists. Generalists have to cover a wide range of specialties and they cannot be equally up to date in each. We have shown that scores can, in certain cases, equal the diagnostic reasoning of specialist physicians. Making these opinions available to the generalist would allow resources to be concentrated on those who need it the most. Scores can help diagnose, thereby avoiding expensive, unnecessary invasive investigations and their associated risk. They help with prediction of prognosis, allowing optimal use of secondary prevention measures. Since Laennec's invention of the stethoscope, doctors have worked to develop tools to aid clinical assessment. In this technological age, clinical scores represent the natural extension of this historic tradition.

Questions and Answers

1. How does selection bias affect test characteristics?
   Falsely raises sensitivity and lowers specificity.
2. Should patients with prior MI be used in studies defining the diagnostic characteristics of a functional study?
   No, MI patients have diagnosed coronary disease for the most part.
3. Are interventional events like bypass surgery valid end points for studies developing prognostic tools?
   No, an abnormal test result leads to the intervention.
4. Do scores substitute for clinical judgment making a poor clinician equivalent to a good one?
   No, scores are best thought of as a second opinion.

5. Should continued symptoms from a patient judged to be low risk from a score be ignored?
Absolutely not. Further testing is required.

Suggested Reading

Swets JA, Dawes RM, and Monahan J. Better decisions through science. Scientific American 2000;October:82–7.

Reid M, Lachs M, Feinstein A. Use of methodological standards in diagnostic test research. JAMA 1995;274:645–51.

Morise A. Comparison of the Diamond-Forrester method and a new score to estimate the pretest probability of coronary disease before exercise testing. Am Heart J 1999;138:740–5.

Raxwal V, Shetler K, Do D, Froelicher V. A simple treadmill score. Chest 2001;119:1933–40.

Froelicher VF, Myers J. Exercise and the heart. Philadelphia: Saunders-Mosby, 1999.

Shaw LJ, Peterson ED, Shaw LK, et al. Use of a prognostic treadmill score in identifying diagnostic coronary disease subgroups. Circulation 1998;16:1622–30.

Alexander K, Shaw L, Delong E, et al. Value of exercise treadmill testing in women. J Am Coll Cardiol 1998;32:1657–64.

Fearon W, Lee D, Froelicher V. The effect of resting ST segment depression on the diagnostic characteristics of the exercise treadmill test. J Am Coll Cardiol 2000;35:1206–11.

Kwok JM, Miller TD, Christian TF, Hodge DO, Gibbons RJ. Prognostic value of a treadmill exercise score in symptomatic patients with nonspecific ST-T abnormalities on resting ECG. JAMA 1999;282:1047–53.

Froelicher VF, Morrow K, Brown M, Atwood E, Morris C. Prediction of artherosclerotic cardiovascular death in men using a prognostic score. Am J Cardiol 1994;73:133–8.

Prakash M, Myers J, Froelicher VF, et al. Clinical and exercise test predictors of all-cause mortality: results from >6,000 consecutive referred male patients. Chest 2001;120:1003–13.

Gauri AJ, Raxwal VK, Roux L, Fearon WF, Froelicher VF. Effects of chronotropic incompetence and beta-blocker use on the exercise treadmill test in men. Am Heart J 2001;142:136–41.

Ashley EA, Myers J, Froelicher V. Exercise testing in clinical medicine. Lancet 2000; 356:1592–7.

Address correspondence and reprint requests to Victor Froelicher, MD, Cardiology Division (111C), VA Palo Alto Health Care System, 3801 Miranda Ave., Palo Alto, CA 94304.