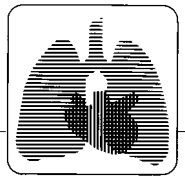

exercise and the heart

Lessons Learned From Studies of the Standard Exercise ECG Test*

Victor F. Froelicher, MD; William F. Fearon, MD; Cynthia M. Ferguson, MD; Anthony P. Morise, MD; Paul Heidenreich, MD; Jeffrey West, MD; and J. Edwin Atwood, MD

(CHEST 1999; 116:1442–1451)

Key words: coronary artery disease; diagnostic techniques; exercise testing

Abbreviations: CAD = coronary artery disease; CI = confidence interval; CKG = cardiokymography; EBCT = electron beam CT; HR = heart rate; LVH = left ventricular hypertrophy; MI = myocardial infarction; ROC = receiver operator curves; SPECT = single-photon emission CT

Motivated by recent systematic reviews of diagnostic tests for coronary artery disease (CAD) and a renewed interest in applying tests for screening healthy individuals, we felt it timely to review the lessons learned from past experience. Since the standard exercise test has been studied for some time, it provides a wealth of experience in this regard. Four major mistakes have been made when evaluating the diagnostic characteristics of the exercise test: (1) choosing subjects for test evaluation who represent a limited challenge to the diagnostic performance of the test; (2) not limiting the amount of workup bias in identifying patients for test evaluation; (3) utilizing soft end points instead of hard end points; and (4) using surrogates instead of an appropriate "gold standard." We will step through each of these errors and provide illustrations for each. In a closing section, we will compare most of the diagnostic techniques that are available for CAD to the exercise test.

*From the Cardiology Division, Veterans Affairs Palo Alto Health Care System, Stanford University (Drs. Froelicher, Fearon, Ferguson, Heidenreich, West, and Atwood), Palo Alto, CA; and the West Virginia University Medical Center (Dr. Morise), Morgantown, WV.

Manuscript received March 24, 1999; revision accepted May 25, 1999.

Correspondence to: Victor Froelicher, MD, Cardiology Division (111C), Veterans Affairs Palo Alto Health Care System, 3801 Miranda Ave, Palo Alto, CA 94304; e-mail: vicmd@aol.com

Limited Challenge

Limited challenge means that a group of healthy or least-diseased patients are compared to patients with severe disease (Fig 1). Limited challenge is present in studies of diagnostic tests when patients without cardiac catheterization but with low risk are compared to patients with demonstrated CAD by cardiac catheterization. Although it is appropriate as the first step in evaluating a new test or measurement, study groups formed by limited challenge are not appropriate for evaluating or demonstrating true test characteristics. Actual test characteristics are only defined in consecutive patients having a complaint that requires testing (*ie*, chest pain). When healthy or least-diseased patients are studied, the specificity of the test should be very high, usually $> 90\%$. When most-diseased patients are studied, the sensitivity should be very high, often $\geq 90\%$ (Fig 2). When receiver operator curves (ROC) are calculated from results from these two disparate groups, a relatively large area will be obtained. It is only when the test or measurement is applied in consecutive patients having a complaint that requires testing, (*ie*, patients presenting with chest pain), that we see the actual test characteristics. Usually the sensitivity and specificity are both much $< 90\%$. Therefore, while limited-challenge groups are easier to obtain for study than consecutive, unselected groups, test characteristics that are generated using limited challenge should not be considered definitive in evaluating the accuracy of any diagnostic test.

An argument could be made that limited challenge does not matter if only test measurements or scores are being compared. However, limited challenge can cause differences in other factors that cause the measurement or scores to be different. For instance, heart rate (HR), systolic BP, and exercise capacity are markedly different in healthy

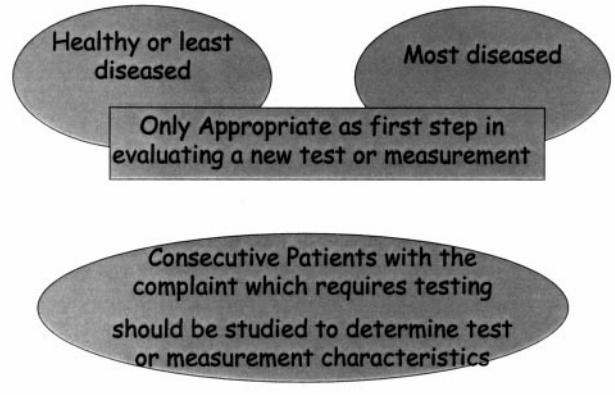

FIGURE 1. Limited challenge means that rather than studying the test in consecutive patients, a group of healthy or least-diseased patients are compared to patients who have severe disease.

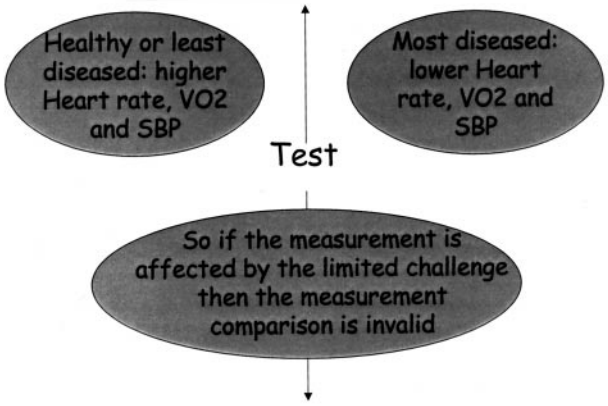

FIGURE 3. HR, systolic BP, and exercise capacity are markedly different in healthy normal subjects compared to those with severe disease (*ie*, limited-challenge populations); VO2 = ventilatory oxygen consumption.

normal subjects compared to those with severe disease. The discriminatory capacity of any ST-segment measurement divided by HR (*ie*, the ST/HR index) is exaggerated when compared in samples with limited challenge (Fig 3).

Workup Bias

Another problem with most of the studies that have evaluated the diagnostic characteristics of the exercise test has been the failure to limit workup bias. Consider the scenario in Figure 4: patients with chest pain that are seen in your office are in the left upper circle. Normal clinical practice results in an exercise test being done, with only certain patients being selected for further evaluation or workup. Cardiac catheterization would be chosen particularly for those with a low exercise capacity and/or abnormal ST-segment response. Others will be catheterized, but the population will be selected to favor these responses. The patients who are excluded from cardiac catheterization after the exercise test will be those with a high exercise capacity and a normal ST-segment response. Others will also be excluded, but in the majority, these characteristics will predominate. Figure 5 shows the results.

Most of the studies that have evaluated the characteristics of the exercise test using the appropriate "gold standard" of cardiac catheterization have some degree of workup bias. In such populations, sensitivity usually is about 70% and specificity is about 70%. What we would really like to know is how the test functions in the population of patients who present

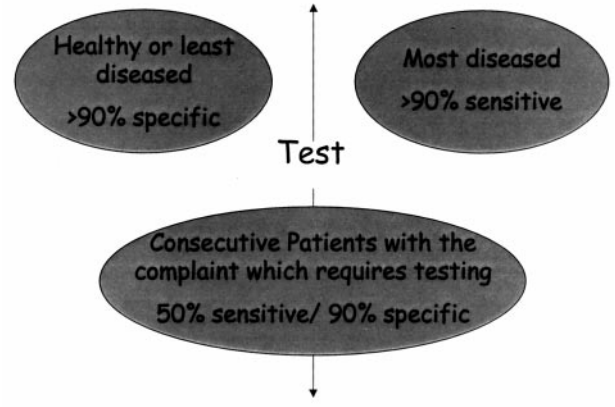

FIGURE 2. Illustrates the results with evaluating a test in a limited-challenge population.

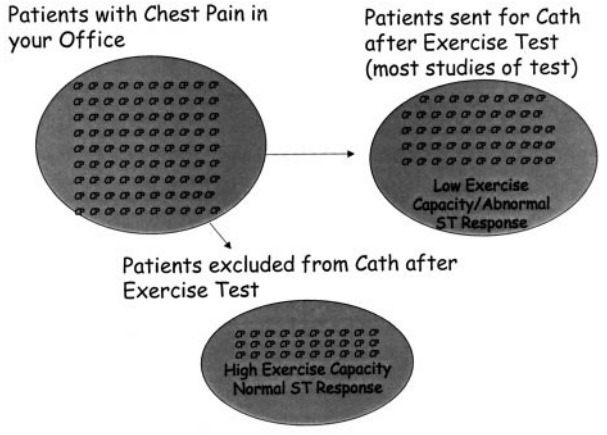

FIGURE 4. Another problem with most of the studies has been the failure to limit workup bias. Patients with chest pain being seen in your office are in the left upper circle. Normal clinical practice then results in an exercise test being done, with only certain patients being selected for further workup. Cardiac catheterization would be chosen particularly for those with a low exercise capacity and an abnormal ST-segment response. Cath = catheterization.

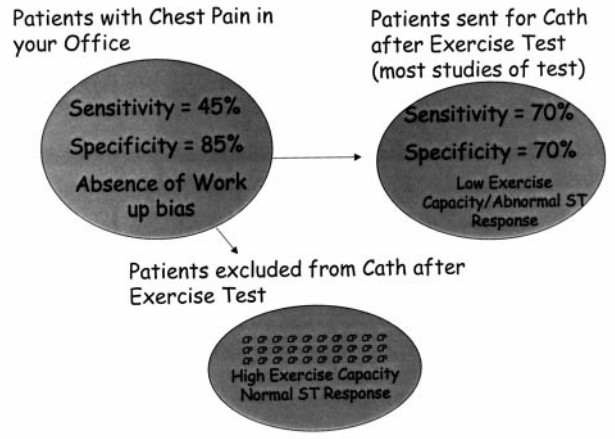

FIGURE 5. Test characteristics in populations with and without workup bias. See Figure 4 legend for expansion of abbreviation.

to the office in the upper left circle. In the few studies that have limited workup bias by protocol, or have had a lower degree of workup bias because of clinical practice (where the exercise test result is largely ignored), different test characteristics are shown: the sensitivity is roughly 40% and the specificity is 85%. These are the characteristics of test performance in the typical office setting. This fall in sensitivity results from the inclusion of patients with CAD who had negative exercise tests (*ie*, false-negatives); the rise in specificity results from the inclusion of patients without CAD who only had negative exercise tests (*ie*, true-negatives).

The meta-analysis of 50 studies that have performed tests with angiographic correlates have been reanalyzed considering the percent of abnormal exercise-induced ST-segment depression in each study.[1] We can assume that there is less workup bias in studies with a relatively lower percentage of patients with abnormal exercise test results, and more workup bias in studies with a higher percentage of abnormal exercise test results. As you can see in Figures 6 and 7, there is a strong correlation between the percent of abnormal test results and specificity and sensitivity. Neither were related to the prevalence of disease. Specificity is higher with less workup bias and sensitivity is lower. This is consistent with the studies that have removed workup bias by protocol.

A recent study[2] included a pilot population and a study population. In the pilot population, investigators were allowed to follow their usual clinical process, while in the actual study population, workup bias was reduced by only entering patients presenting with chest pain who agreed to both the exercise test and cardiac catheterization. All other procedures were performed in the same way during both the pilot and the actual study. Table 1 shows the results.

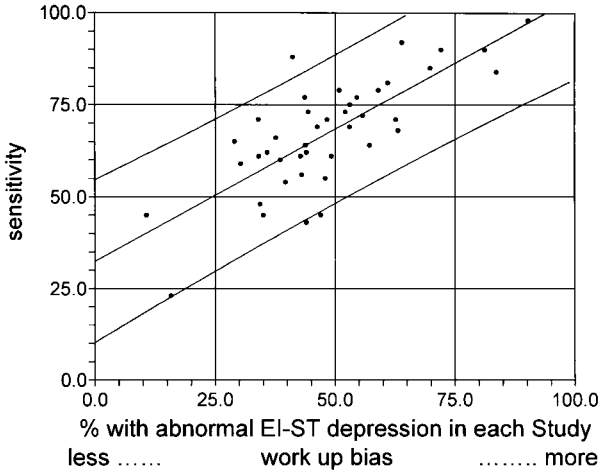

FIGURE 6. The relationship between sensitivity and the percent of abnormal tests in each of the 50 studies of patients with coronary angiography. Patients with a history of MI were excluded. There is a good correlation between the percent of abnormal test results and specificity and sensitivity. Specificity is higher with less workup bias, and sensitivity is lower; EI-ST = exercise-induced ST-segment.

Even though the disease prevalence was the same, the lower percentage of patients with an abnormal treadmill test was due to the reduction in workup bias that altered the test diagnostic characteristics.

In summary, workup bias is due to the fact that because of good clinical practice, only a portion of patients who are seen with chest pain and undergo exercise tests also undergo cardiac catheterization. For the most part, patients with high exercise capacity and normal ST-segment responses are excluded, and patients with low exercise capacity and abnormal ST-segment responses are selected for angiography. Though this is not 100% in any of the studies,

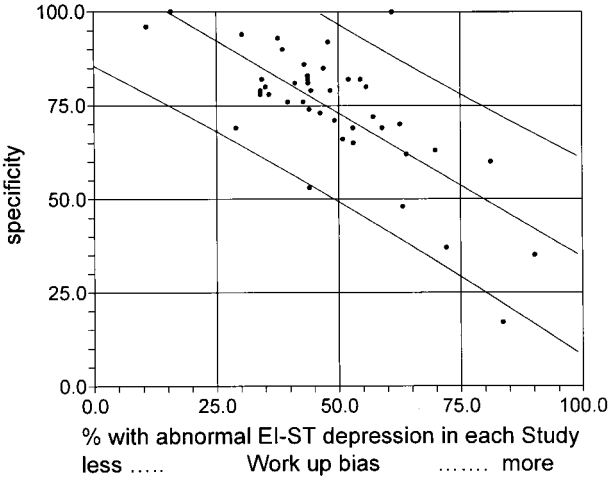

FIGURE 7. The relationship between specificity and the percent of abnormal exercise test results in each of the 50 studies of patients with coronary angiography. See Figure 6 for abbreviation.

Table 1—*The Results in QUEXTA and Its Pilot Population

Population Type	Population, No.	Prevalence of CAD, %	Abnormal TMT, %	Sensitivity, %	Specificity, %
Pilot	321	49	47	65	69
Study	842	51	29	45	85

*QUEXTA = quantitative exercise testing and angiography; TMT = treadmill test.

tendencies for this to occur vary from study to study, and that is why different test performance characteristics have been obtained with the exercise test. In the studies that have removed workup bias by protocol, these differences are very clearly seen. As you can see in the Table 2, approximately 12,000 patients have been included in the 50 studies, with varying degrees of workup bias. The mean sensitivity has been 67%, and mean specificity has been 72%. The two studies that have removed workup bias by protocol have included 2,000 patients and have considerably different test characteristics.[3]

An argument could be made that the clinician does not want to perform a cardiac catheterization on everyone. That is not the point. To determine the discriminatory characteristics of a test, a study protocol must be followed to catheterize and test all patients presenting with chest pain. Then the practicing physician can trust the results from the study to know how the test behaves in his or her office practice, and to better make decisions as to who needs further evaluation. Therefore, studies that fail to limit workup bias are doomed to overestimate sensitivity and underestimate specificity.

Hard vs Soft End Points

A third lesson learned from studies of the standard exercise ECG test is the importance of the end points considered when using data other than the coronary angiogram. Hard end points are myocardial infarction (MI) and death, and soft end points include chest pain and coronary interventions. The best example of the problem of using soft end points instead of hard end points comes from the screening studies. As previously reviewed, there are 12 studies that used the exercise test to screen asymptomatic individuals for cardiac disease.[4] Patients were screened for silent heart disease using the exercise test and were followed for 5 to 10 years for cardiac events. Considerably different results have been obtained in these studies, according to the end points considered. When angina is included as an end point, nonspecific symptoms in a subject with an abnormal test are more likely to be called coronary disease during the follow-up period. Hard end points, like death or MI, eliminate this misclassification and are more appropriate. The first screening studies included angina as an end point, and the last four studies used only hard end points. In Table 3, you can see that the first studies tested 5,000 subjects and ranged in size from 113 to 1,390 individuals. Sensitivity was 50%, specificity was 90%, the risk ratio was nine times, and the predictive value of a positive response was 25%. That means that one out of four patients with abnormal test results went on to have a cardiac event. Remember that some of these events will be angina, which was probably not truly due to cardiac disease or truly angina. The last four studies were larger in size and included only hard end points. The sensitivity of the test has been about 25%, specificity about 90%, the risk ratio was four times, and the predictive value of a positive response was only 5%. That means that only 1 of 20 people with abnormal test results went on to a cardiac event.

Table 2—*Angiographic Correlative Studies for Diagnosis of any CAD*

Studies (No.)	Patients, No.	Sensitivity, %	Specificity, %
With workup bias (58)	12,000	67	72
Without workup bias (2)	2,000	45	90

Because of this very limited predicted value in any asymptomatic population, screening has not been recommended. It can cause more harm then good, and it can lead to unnecessary tests. Several studies have even tried to raise the pretest probability by considering risk factors, and have not been able to do so to a level that limits the false-positives and improve the predicted value. Theoretically, this should be possible by using a risk factor score.

The argument could be made that we should be able to predict which patients are going to get milder forms of coronary disease than death or MI (*ie*, angina). Certainly we would like to do this, but the problem is that the test result is creating end points. Thus, using soft end points exaggerates the sensitivity and predictive value of the test. This could be avoided by blinding all parties to the test result, but this has been considered unethical. Since some of the asymptomatic individuals that develop chest pain really have angina due to coronary disease, the sensitivity probably lies between the 25% and 50% that was obtained in the studies that used hard, and

Table 3—*The 12 Screening Studies**

Studies	Patients, No. (range)	Sensitivity, %	Specificity, %	Risk Ratio, X	PV+, %
First eight studies†	5,526 (100–1,390)	50	90	9	25
Last four studies‡	12,212 (> 2,000)	25	90	4	5

*PV+ = positive predictive value.
†With soft end points included.
‡Hard end points only.

hard plus soft end points, respectively. While soft end points are very much appropriate for randomized intervention trials, they can result in important prediction errors in studies of diagnostic procedures.

Screening studies have other population selection considerations than diagnostic studies. First, the population should truly be asymptomatic and should represent a random sample of the target population. Volunteers are not appropriate because they usually represent the extremes of the population: the most healthy, and those who are concerned about their health for personal reasons, such as family history or symptoms they choose to deny. Volunteers represent a subtle form of limited challenge.

Another misleading soft end point is using interventions as cardiac events. Investigators often find that with modern treatment, there are an inadequate number of cardiovascular deaths and infarctions in most populations studied to obtain statistically significant results. This is particularly the case in studies in which patients with congestive heart failure (*ie*, those who have a much higher mortality than patients with ischemia alone) are excluded. Therefore, in order to have enough end points, follow-up studies have often included bypass surgery or percutaneous transluminal coronary angioplasty as an end point. In fact, very often the majority of the end points are interventions. This is problematic because the test result often determines who gets these procedures, and so it is not really valid to include them as events predicted by the test. A built-in bias favors the test predicting the result. While some investigators have suggested that they can detect this bias by time intervals between the test and the intervention, or by record review, their results are still suspect. Therefore, test evaluation studies that include soft subjective end points that could be influenced by the results of the test under study should be considered suboptimal.

Misleading Surrogates

A final limitation observed in studies of diagnostic tests has been the use of surrogates for various measurements and outcomes. An example is using the same criteria for computer measurement of the ST segments that is used in the visual analysis. Applying strict one millimeter and horizontal slope criteria with a computer is not the same as when visually analyzing the exercise ECG. This is a totally different classification of abnormal ST-segment responses than when done visually. As it turns out, truly horizontal slopes are rare with a computer, whereas our eye flattens things routinely in reading the ST segment. In addition, one millimeter by computer includes visual measurements to approximately 0.75 mm. Visual assessment involves a certain degree of rounding up to the next millimeter of ST-segment depression. Fewer patients are classified as abnormal using standard criteria if a computer does the classification, rather than visually using classic criteria. If comparing computer and visual analysis, one has to adjust for this "fuzz factor" that is inherent in the visual analysis.

A problematic surrogate is using nuclear imaging instead of angiography as a "gold standard." It is well known that nuclear imaging has limitations in predicting coronary disease and cannot be used to replace the best "gold standard" that we have. Nuclear imaging could only be used if we are trying to predict the results of nuclear testing. Even though angiography has limitations, it is the best standard we have for obstructive coronary disease. Therefore, surrogates for standards should be carefully considered and justified only when they perform equal to or better than the standard itself. Ease of attainment over the standard should not be the only consideration.

Comparison With Other Tests

While studies of the standard exercise test have been helpful in illustrating the problems in demonstrating test characteristics, newer technologies have often been evaluated by studies with the same limitations. Nonetheless, it is appropriate to compare the newer diagnostic modalities with the standard exercise test because it is a mature, established technology. The equipment and personnel for performing it are readily available. Exercise testing equipment is relatively inexpensive, so that replace-

ment or updating is not a major limitation. The exercise test can be performed in the doctor's office and does not require injections or exposure to radiation. It can be an extension of the medical history and physical examination, providing more than simply diagnostic information. Furthermore, it can determine the degree of disability and impairment to quality of life, and it can be the first step in rehabilitation and altering a major risk factor (physical inactivity).

Some of the newer add-ons or substitutes for the exercise test have the advantage of being able to localize ischemia as well as diagnose coronary disease when the baseline ECG negates ST-segment analysis (more than one millimeter ST-segment depression; left bundle branch block; Wolfe-Parkinson-White). The substitutes for exercise also have the advantage of not requiring the patient to exercise, which is particularly valuable clinically for those who cannot walk. However, while the newer technologies appear to have better diagnostic characteristics, this is not always the case, particularly when more than the ST segments from the exercise test are used in scores.

Test evaluation has been advanced by the writings of Philbrick et al,[5] Reid et al,[6] and Guyatt,[7] and so we are now in a better position to evaluate studies of test characteristics. A number of researchers have applied these guidelines along with meta-analysis to come to consensus on the diagnostic characteristics of the available tests for angiographic CAD.[8,9] Table 4 presents some of the results from meta-analysis and from multicenter studies. The techniques listed include electron beam CT (EBCT), a fast radiographic technique that can make a quantitative measurement of coronary artery calcification.[10,11] The cardiokymography (CKG) is a simple motion sensor that permits recording of the movement of the left ventricle using a small transducer placed on the chest wall. Signal averaging has recently enhanced this older technology.[12] Nuclear perfusion imaging includes both the early studies (mainly using thallium radiographic images) and the more modern use of single-photon emission CT (SPECT), which requires computer enhancement of the emissions of thallium and other agents.

Since sensitivity and specificity are inversely related and altered by the chosen cut point for normal/abnormal, the predictive accuracy (the percentage of patients correctly classified as normal and abnormal) is a convenient way to compare tests. For instance, while the sensitivity and specificity for exercise testing and EBCT are nearly opposite, the predictive accuracy of the tests are similar. This means that by altering their cut points (*ie*, lowering the amount of ST-segment depression or raising the calcium score) would result in similar sensitivities and specificities. Since predictive accuracy can be thought of as the number of individuals correctly classified out of 100 tested, simply subtracting predictive accuracy provides an estimate of how many more patients are classified by substituting one test for another test. However, this does assume a disease prevalence of 50% that is the intermediate probability for the appropriate use of diagnostic tests (that is, predictive accuracy is affected by disease prevalence).

While the nonexercise stress tests are very useful, the results shown below are probably better than their actual performance because of patient selection. For studies of diagnostic characteristics, patients with a prior MI should be excluded because the diagnosis of coronary disease is not an issue in them.

Exercise Test Scores

The exercise testing studies that have considered additional information in addition to the ST-segment

Table 4—*Comparison of Exercise Testing and Add-ons or Other Test Modalities*

Grouping	Studies, No.	Total Patients, No.	Sensitivity, %	Specificity, %	Predictive Accuracy, %
Meta-analysis of standard exercise ECG	147	24,047	68	77	73
Excluding MI patients	41	11,691	67	72	69
Limiting workup bias	2	>1,000	50	90	69
Meta-analysis of exercise test scores	24	11,788			80
Thallium scintigraphy	59	6,038	85	85	85
SPECT without MI	27	2,136	86	62	74
Exercise echocardiography	58	5,000	84	75	80
Exercise echocardiography excluding MI patients	24	2,109	87	84	85
Nonexercise stress tests					
Persantine thallium	11	< 1,000	85	91	87
Dobutamine echocardiography	5	< 1,000	88	84	86
CKG	1	617	71	88	79
EBCT	5	2,373	90	45	61

response have been reviewed and demonstrate the improved test characteristics obtained using this approach.[13] Recent publications have extended the DUKE prognostic score to diagnosis,[14] and a consensus approach that uses a number of equations appears to make the scores more portable to other populations.[15]

Nuclear Perfusion and Echocardiography

Investigators from the University of California San Francisco reviewed the contemporary literature in order to compare the diagnostic performance of exercise echocardiography and exercise nuclear perfusion scanning in the diagnosis of CAD.[16] Their work included studies published between January 1990 and October 1997 that were identified from a MEDLINE search, bibliographies of reviews and original articles, and suggestions from experts in each area. Articles were included if they discussed exercise echocardiography and/or exercise perfusion imaging with thallium or sestamibi (largely SPECT) for detection and/or evaluation of CAD; if data on coronary angiography were presented as the reference test; and if the absolute numbers of true-positive, false-negative, true-negative, and false-positive observations were available or derivable from the data presented. Studies were excluded if they were performed exclusively in patients after MI, after percutaneous transluminal coronary angioplasty, after coronary artery bypass grafting, or with recent unstable coronary syndromes. Two reviewers used a standardized spreadsheet to independently extract clinical variables, technical factors, and test performance. Discrepancies were resolved by consensus. Forty-four articles met the inclusion criteria: 24 articles reported exercise echocardiography results in 2,637 patients (weighted mean age, 59 years; 69% were men; 66% had angiographic coronary disease; and 20% had prior MI) and 27 articles reported exercise SPECT in 3,237 patients (70% were men; 78% had angiographic coronary disease; and 33% had prior MI). In pooled data weighted by the sample size of each study, exercise echocardiography had a sensitivity of 85% (95% confidence interval [CI], 83 to 87%) with a specificity of 77% (95% CI, 74 to 80%). Exercise perfusion yielded a similar sensitivity of 87% (95% CI, 86 to 88%) but a lower specificity of 64% (95% CI, 60 to 68%). In a summary receiver operating characteristic model that compared exercise echocardiography performance to exercise perfusion, exercise echocardiography was associated with significantly better discriminatory power when adjusted for age, publication year, and a setting including known CAD for perfusion studies. In models comparing the discriminatory abilities of exercise echocardiography and exercise perfusion to exercise testing without imaging, both echocardiography and perfusion performed significantly better than the exercise ECG.

A similar meta-analysis from DUKE that considered 58 studies (performed over 15 years) of the diagnostic characteristics of exercise echocardiography has been reported in abstract form. The average sensitivity was 84% and the specificity was 75%. These studies agree that exercise echocardiography (specificity of 84%) has better specificity than SPECT (specificity of 62%) but not the exercise ECG (specificity of 90%). The earlier meta-analysis of thallium perfusion imaging demonstrated better test characteristics with planar radiographic imaging than the currently used SPECT.

CKG

A multicenter study has demonstrated the diagnostic accuracy of CKG 2 to 3 min after exercise in 617 patients undergoing cardiac catheterization.[17] There were 12 participating centers using a standardized protocol. Obtaining adequate CKG tracings (as accomplished in 82% of patients) was dependent on the skill of the operator and on certain patient characteristics. Of the 327 patients without prior MI who had technically adequate CKG and ECG tracings, 166 patients (51%) had coronary disease. Both the sensitivity and specificity of CKG (71 and 88%, respectively) were significantly greater than the values for the exercise ECG (61 and 76%, respectively). The CKG is a simple, inexpensive add-on to the standard exercise test that only requires placing a small transducer on the chest before and after exercise. It appears to detect wall motion abnormalities nearly as well as echocardiography. It has not been widely adapted because it was never reimbursable.

EBCT

Of the angiographic correlative studies of EBCT, we selected the five with > 200 subjects without overlapping populations. One hundred sixty men and women with coronary disease (age range, 45 to 62 years old; 138 with obstructive CAD; and 22 with normal coronary arteries) and 56 age-matched healthy control subjects underwent double-helix CT.[18] Sensitivity in detecting obstructive CAD was high (91%); however, specificity was low (52%) because of calcification in nonobstructive lesions. A multicenter study evaluated patients who were referred for angiography.[19] Four hundred ninety-one symptomatic patients underwent coronary angiography and EBCT at five different centers between 1989 and 1993. The area under the ROC curve was

0.75 for the coronary calcium score. In this group, sensitivity of any detectable calcification by EBCT as an indicator of significant stenosis (> 50% narrowing) was 92% and specificity was 43%. When these CT images were reinterpreted in a blinded and standardized manner, however, specificity was only 31%. In another multicenter study[20] of 710 enrolled patients, 427 had significant angiographic disease, and coronary calcification was detected in 404, yielding a sensitivity of 95%. Of the 283 patients without angiographically significant disease, 124 had negative EBCT studies, for a specificity of 44%. Ultrafast CT was used to detect and quantify coronary artery calcium levels in 584 subjects (mean age, 48 years old), 19% of whom had clinical CAD.[21] Total calcium scores were calculated based on the number, areas, and peak Hounsfield computed tomographic numbers of the calcific lesions detected. Sensitivity, specificity, and predictive values for clinical CAD were calculated for several total calcium scores in each decade. For the age groups of 40 to 49 years and 50 to 59 years, a total score of 50 resulted in a sensitivity of 71% and 74% and a specificity of 91% and 70%, respectively. For the age group of 60 to 69 years, a total score of 300 gave a sensitivity of 74% and a specificity of 81%. Three hundred sixty-eight symptomatic patients underwent coronary angiography and EBCT at four different centers between April 1989 and December 1993.[22] Coronary risk factors were obtained in all 368 patients. One hundred fifty-eight patients (43%) had angiographically obstructive CAD (> 50%), and 297 patients (81%) had coronary calcification. At the bivariate level, only male gender and log-transformed coronary calcification were predictive of angiographic disease. It appears that even the best studies of EBCT suffer from limited challenge and workup bias, so that the true characteristics of this procedure are not known. However, the five studies averaged in the Table demonstrated a high sensitivity and a low specificity, with a predictive accuracy of about 61%. Although adjusting the cut point for calcium density can alter the sensitivity and specificity, the EBCT is not more diagnostic for angiographic CAD than the standard exercise test.

Cost Efficacy

There have been numerous attempts to evaluate the cost efficacy of various diagnostic strategies for CAD.[23–25] These rely on the test characteristics (*ie*, sensitivity and specificity) used in the calculations that should be derived from meta-analysis. Unfortunately, they usually do not consider the societal costs of the implementation of the technology. For instance, if EBCT was recommended as a "screening test" superior to the standard exercise test, such a recommendation would imply the need for more than the 42 units that are available in the United States at a cost of from $1 to $2 million each. These costs, plus the associated yearly maintenance of $150,000 per device, make EBCT extremely more expensive than any other technique.

Discussion

The American Heart Association and other organizations have expressed renewed interest in applying screening tests in asymptomatic individuals because of the data showing the ability of the statins to lower cholesterol and decrease cardiac events.[26] The thought is that a marker of atherosclerosis could help to decide who should or should not be started on a statin, rather than just relying on a risk factor score or threshold. Because of the association of radiographically imaged calcification with atherosclerotic plaque burden, the new technology of EBCT has been proposed as such a screening test.[27] Unfortunately, studies avoiding the above-mentioned lessons have not been performed with EBCT, and we do not yet know the true unbiased test characteristics of EBCT. An obvious result of some of the screening studies of EBCT is that a finding of calcium leads to catheter and surgical interventions in asymptomatic individuals.[28] Unfortunately, this creates end points and exaggerates the discriminatory power of the test. Other tests that are considered for screening are the carotid ultrasound, the ankle-brachial index, the resting ECG, and the standard exercise test.

The exercise test guidelines have not recommended using the standard exercise test for screening because of the problem with false-positives in populations with a low prevalence of coronary disease.[29] Given the characteristics of the standard exercise test, in a low prevalence population, roughly 9 of 10 individuals with an abnormal test will have false-positive results; that is, they will not have coronary disease. This leads to insurance and employment problems, and it can even start a healthy person through the cardiologist's diagnostic cascade. If the results of a screening test would be reported as a probability (for instance, as a treadmill or calcium score) rather than a negative or a positive, and only used to tailor the preventive treatment, then any screening test could be helpful. However, even this approach remains a hypothesis until the following study is performed: an asymptomatic population is identified and then randomized to either receiving or not receiving the screening test and a standardized strategy of response. Such a study has been accomplished for mammography; whether or not it can be accomplished for coronary disease remains to be seen.

Considering Bayes theorem, we would like to increase the pretest probability of disease prior to applying a screening test by using risk factors to identify asymptomatic individuals with a higher prevalence of coronary disease. While this seems reasonable, the only study to attempt this did not document a better predictive value. The Lipid Research Clinics Coronary Primary Prevention Trial investigators attempted to predict coronary heart disease morbidity and mortality in hypercholesterolemic men from an exercise test.[30] To study whether the test was more predictive for hypercholesterolemic men (*ie*, increasing the pretest probability for disease), data from 3,806 asymptomatic hypercholesterolemic men were analyzed. The prevalence of an abnormal test result was 8.3%. During a mean 7-year follow-up, the mortality rate from coronary heart disease was 6.7% (21 of 315 patients) in men with abnormal test results and 1.3% (46 of 3,460 patients) in men with negative test results. The age-adjusted rate ratio for an abnormal test result compared to a negative test result was nearly seven times in the placebo group and five times in the cholestyramine group. While ST-segment depression yielded excellent risk ratios for the development of coronary disease, the predictive value of an abnormal response was still low.

For a screening test to be worth the additional expense, it must add significantly to the ability of standard risk factors in identifying asymptomatic individuals with subclinical disease. The method with which the risk is estimated with the risk factors must also be considered for such a comparison. A simple adding of risk factors, as recommended by the Joint National Committee or the National Cholesterol Education Program, is not as accurate as using the logistic regression equations that were developed from the Framingham data.[31] In an asymptomatic population, the Framingham score calculates an estimate of the 5-year incidence of cardiovascular events, using age, smoking, diabetes, standing systolic BP, ECG-left ventricular hypertrophy (LVH), and the levels of high density lipoprotein and total cholesterol.[32] The most important factor in the score was the ECG diagnosis of LVH. The most recent version of the score removed LVH because the prevalence of LVH has declined with the improved treatment for hypertension.[33] Theoretically, a level of this score should be able to identify asymptomatic individuals with a risk for coronary events similar to patients symptomatic with CAD who should be receiving statin drugs. Rather than starting the patient on a treatment regimen for life, perhaps an exercise test could be used to make this decision (*ie*, only those with an abnormal test result would be treated). However, this is complicated by the fact that probably those at high risk should be treated anyway (since they are likely to develop disease, even if they do not have an abnormal test result), and only those with an intermediate risk and an abnormal test result should be treated. The cut points for these levels of risk using the Framingham score are somewhat arbitrary and are dependent on how portable the score is to other populations.

Summary

The above lessons that were learned from studies of exercise testing are pertinent today, particularly as we investigate new diagnostic technologies. The key role of population selection in assessing the validity of studies evaluating diagnostic and screening tests for CAD is now very apparent. It is interesting to note the evolution of test evaluation that has occurred over the history of exercise testing. A number of articles that state the rules for evaluating studies of diagnostic tests have targeted the exercise test as well as other procedures. Others have used these standards in meta-analysis of tests. Thus, the criteria for demonstrating the diagnostic characteristics of tests are now common knowledge and should be considered when reviewing such studies. Particularly when applied with scores that consider more than the ST-segment response, the diagnostic characteristics of the exercise test are similar to those of newer technologies that are more expensive and less well studied.

Appendix

Definitions and Calculations of the Terms Used to Quantify Test Diagnostic Accuracy

- Sensitivity = (TP/TP + FN) × 100
- Specificity = (TN/FP + TN) × 100
- TP = those with an abnormal test result and disease (true-positives)
- TN = those with a normal test result and no disease (true-negatives)
- FP = those with an abnormal test result and no disease (false-positives)
- FN = those with a normal test result and disease (false-negatives)
- TP + TN + FP + FN = total population
- + Likelihood ratio = ratio that a positive response is likely to have disease vs a negative response (sensitivity/[1-specificity])
- − Likelihood ratio = ratio that a negative response is not likely to have disease vs a positive response (1-sensitivity/[specificity])
- Predictive value positive (PV+) = the percentage of those with an abnormal (positive) test result who have disease

- Predictive value negative (PV−) = the percentage of those with a negative test result who do not have disease
- Risk ratio = the ratio of disease rate in those with a positive test result compared to those with a negative test result (PV+/[FN/FN + TN])
- Predictive accuracy = the percentage of correct classifications, both + and −
- ROC = range of characteristics curve; plot of sensitivity vs specificity for the range of measurement cut points

References

1 Gianrossi R, Detrano R, Lehmann K, et al. Exercise-induced ST: depression in the diagnosis of coronary artery disease; a meta-analysis. Circulation 1989; 80:87–98

2 Froelicher VF, Lehmann KG, Thomas R, et al. The electrocardiographic exercise test in a population with reduced workup bias: diagnostic performance, computerized interpretation, and multivariable prediction; Veterans Affairs Cooperative Study in Health Services #016 (QUEXTA) Study Group. Quantitative Exercise Testing and Angiography. Ann Intern Med 1998; 128(12 Pt 1):965–974

3 Morise AP, Diamond GA. Comparison of the sensitivity and specificity of exercise electrocardiography in biased and unbiased populations of men and women. Am Heart J 1995;130:741–747

4 Froelicher VF, Quaglietti S. Handbook of Exercise Testing. Boston, MA: Little, Brown Publishers, 1995; 101–103

5 Philbrick JT, Horwitz RI, Feinstein AR. Methodological problems of exercise testing for coronary artery disease: groups, analysis and bias. Am J Cardiol 1980; 46:807–812

6 Reid M, Lachs M, Feinstein A. Use of methodological standards in diagnostic test research. JAMA 1995; 274:645–651

7 Guyatt GH. Readers' guide for articles evaluating diagnostic tests: what ACP Journal Club does for you and what you must do yourself [editorial]. ACP J Club 1991; 115:A-16

8 Gianrossi R, Detrano R, Columbo A, et al. Cardiac fluoroscopy for the diagnosis of coronary artery disease: a meta-analytic review. Am Heart J 1990; 120:1179–1188

9 Detrano R, Janosi A, Marcondes G, et al. Factors affecting sensitivity and specificity of a diagnostic test: the exercise thallium scintigram. Am J Med 1988; 84:699–710

10 Wexler L, Brundage B, Crouse J, et al. Coronary artery calcification: pathophysiology, epidemiology, imaging methods, and clinical implications; a statement for health professionals from the American Heart Association. Writing Group. Circulation 1996; 94:1175–1192

11 Fiorino AS. Electron-beam computed tomography, coronary artery calcium, and evaluation of patients with coronary artery disease. Ann Intern Med 1998; 128:839–847

12 Gehring J, Koenig W, Donner M, et al. The diagnostic value of signal-averaged stress cardiokymography compared with exercise electrocardiography. J Noninvas Cardiol 1998; 5:32–41

13 Yamada H, Do D, Morise A, et al. Review of studies utilizing multi-variable analysis of clinical and exercise test data to predict angiographic coronary artery disease. Prog Cardiovasc Dis 1997; 39:457–481

14 Shaw LJ, Peterson ED, Shaw LK, et al. Use of a prognostic treadmill score in identifying diagnostic coronary disease subgroups. Circulation 1998; 98:1622–1630

15 Dat Do, Jeffrey A West, Anthony Morise, et al. A consensus approach to diagnosing coronary artery disease based on clinical and exercise test data. Chest 1997; 111:1742–1749

16 Fleischmann KE, Hunink MG, Kuntz KM, et al. Exercise echocardiography or exercise SPECT imaging? A meta-analysis of diagnostic test performance. JAMA 1998; 280:913–920

17 Weiner DA. Accuracy of cardiokymography during exercise testing: results of a multicenter study. J Am Coll Cardiol 1985; 6:502–509

18 Shemesh J, Apter S, Rozenman J, et al. Calcification of coronary arteries: detection and quantification with double-helix CT. Radiology 1995; 197:779–783

19 Detrano R, Hsiai T, Wang S, et al. Prognostic value of coronary calcification and angiographic stenoses in patients undergoing coronary angiography. J Am Coll Cardiol 1996; 27:285–290

20 Budhoff MJ, Georgiou D, Brody A, et al. Ultrafast computed tomography as a diagnostic modality in the detection of coronary artery disease: a multicenter study. Circulation 1996; 93:898–904

21 Agatston AS, Janowitz WR, Hildner FJ, et al. Quantification of coronary artery calcium using ultrafast computed tomography. J Am Coll Cardiol 1990; 15:827–832

22 Kennedy J, Shavelle R, Wang S, et al. Coronary calcium and standard risk factors in symptomatic patients referred for coronary angiography. Am Heart J 1998; 135:696–702

23 Rumberger JA, Behrenbeck T, Breen JF, et al. Coronary calcification by electron beam computed tomography and obstructive coronary artery disease: a model for costs and effectiveness of diagnosis as compared with conventional cardiac testing methods. J Am Coll Cardiol 1999; 33:453–462

24 Pilote L, Pashkow F, Thomas JD, et al. Clinical yield and cost of exercise treadmill testing to screen for coronary artery disease in asymptomatic adults. Am J Cardiol 1998; 81:219–224

25 Mattera JA, Arain SA, Sinusas AJ, et al. Exercise testing with myocardial perfusion imaging in patients with normal baseline electrocardiograms: cost savings with a stepwise diagnostic strategy. J Nucl Cardiol 1998; 5:498–506

26 Downs JR, Clearfield M, Weis S, et al. Primary prevention of acute coronary events with lovastatin in men and women with average cholesterol levels: results of AFCAPS/TexCAPS. Air Force/Texas Coronary Atherosclerosis Prevention Study. JAMA 1998; 279:1615–1622

27 Detrano R, Froelicher V. A logical approach to screening for coronary artery disease. Ann Intern Med 1987; 106:846–852

28 Arad Y, Spadaro LA, Goodman K, et al. Predictive value of electron beam CT of the coronary arteries: 19-month follow-up of 1173 asymptomatic subjects. Circulation 1996; 93:1951–1953

29 Gibbons RJ, Balady GJ, Beasley JW, et al. Guidelines for exercise testing: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee on Exercise Testing). J Am Coll Cardiol 1997; 30:260–311

30 Ekelund LG, Suchindran CM, McMahon RP, et al. Coronary heart disease morbidity and mortality in hypercholesterolemic men predicted from an exercise test: the Lipid Research Clinics Coronary Primary Prevention Trial. J Am Coll Cardiol 1989; 14:556–563

31 Grover SA, Coupal L, Hu XP. Identifying adults at increased risk of coronary disease: how well do the current cholesterol guidelines work? JAMA 1995; 274:801–806

32 Anderson, P. An updated risk factor profile. Circulation 1991; 83:356–362

33 Wilson PW, D'Agostino RB, Levy D, et al. Prediction of coronary heart disease using risk factor categories: Circulation 1998; 97:1837–1847