Exercise testing in clinical medicine

Euan A Ashley, Jonathan Myers, Victor Froelicher

Exercise-induced changes in the electrocardiogram have been used to identify coronary artery disease for almost a century. Over the past decade, however, clinicians have increasingly focused on more expensive diagnostic tools believing them to offer improved diagnostic accuracy. In fact, by incorporating historical data, the simple exercise test can in most cases outperform the newer tests. The use of prediction equations and non-staged exercise protocols can improve the test still further, while advances in the use of the test for prognosis, with the discovery of novel risk factors and the addition of gas analysis, may in the future shift the primary emphasis away from diagnosis. Brief, inexpensive, and done in most cases without the presence of a cardiologist, the exercise test offers the highest value for predictive accuracy of any of the non-invasive tests for coronary artery disease.

The exercise test has a long history in cardiovascular medicine. Although Willem Einthoven,1 in 1908, was the first to document changes in the ST segment of the electrocardiogram (ECG) with exercise, it was not until 1932 that Goldhammer and Scherf proposed exercise electrocardiography as a diagnostic tool for angina. Since then, the technique of exercising patients to stress the cardiovascular system has played a central part in the diagnostic work-up of coronary artery disease. However, a low reported sensitivity1 has led some specialists, notably those in the USA, to turn away from the traditional exercise test in favour of more expensive investigations. We propose that this desertion is premature, because optimisation of the test can produce predictive values equal to those of the best and most expensive techniques. Indeed the test should be more widely used.

Diagnosis

Meta-analysis of trials has shown that the exercise test has a specificity of around 80% and a sensitivity of around 70% for obstructive coronary disease confirmed by angiography.4 However, many of these studies had methodological problems of limited challenge and work-up bias. The former takes place when patients already known to have coronary artery disease (eg, those with previous myocardial infarction) are used to challenge a diagnostic test. Work-up bias describes the situation in which the gold-standard test is applied only after a positive result from the test in question. In fact, in the only trial to date to clearly avoid work-up bias5 specificity was 85% and sensitivity 45%. However, the application of computers yields multivariate prediction equations that can, by incorporating data obtained from patients outside the exercise-test setting, make the test as accurate as much more expensive tests (table 1 and panel 1). Although they may seem intimidating, programs are available for desktop computers and on the internet to do the calculations.

One difficulty with such equations is that they are hard to apply to populations other than those in which they were developed. One way of overcoming this drawback is to use a consensus approach, in which a final classification of low or high risk is made on the basis of a consensus among equations validated in different populations (eg, each equation classifies a given patient into low, intermediate, or high risk, and a given classification is accepted only if two of three equations agree).6 Such a method predicts angiographic disease better than do cardiologists.6 A variation on this theme is the use of prognostic scores, the most universal of which is the Duke treadmill score because it can be used both for prognosis7 and diagnosis.8 The score consists of three variables: the amount of ST depression, exercise capacity, and the presence of angina. This score can be calculated by a nomogram or by a computer (some of the commercial treadmill systems automatically calculate it).

A further advantage of the consensus or score approach is that, as opposed to yes or no answers, the separation of patients into groups according to risk suggests a management course: low-risk patients need no further testing at that time; those at high-risk need an invasive study; and those at intermediate-risk need another non-invasive study. On the assumption that the intermediate group is eventually diagnosed, data suggest that this test strategy would produce sensitivity and specificity values of around 90%.6

Another approach to improving test characteristics is to investigate different ECG criteria. Atwood and colleagues9 took 100 computed measurements from digitised exercise ECG recordings and related them to angiographic data on the same patients. They found that computerised measurements at 3·5 min of recovery, from lead V5, 60 ms after the QRS complex (ST60), were better than all other single measurements. Further, prediction equations that included clinical and exercise-test data showed the greatest diagnostic power. In a new approach, Michaelides and colleagues examined 245 patients who underwent exercise testing with standard 12 leads, right ventricular leads, and thallium-201 scintigraphy. They found sensitivities of 66%, 92%, and 93%, and specificities of 88%, 88%, and 82%, respectively, for the detection of “any” coronary artery disease by angiography.10 It should be noted, however, that their study was of a population with an unusually high prevalence of coronary disease. The QRS complex has also been a focus of investigation as a marker of
myocardial ischaemia consequent on its use in scoring systems to predict infarct size. The Athens score is calculated from exercise-induced changes in the QRS complex, and a Dutch group have proposed a score based on composite changes in the Q, R, and S waves in two leads (aVF, V5). These investigators reported a sensitivity of 88-2% and a specificity of 84-8% for coronary artery disease, compared with values for ST depression of 54-9% and 83%, respectively. It appears that within this basic test there could yet be a measurement with even greater diagnostic power.

A question frequently raised is whether the test can predict disease in different sections of the population, such as in women or the elderly. In fact, the diagnostic characteristics are unchanged in women when age and symptoms move them into the intermediate probability of disease level. However, the lower prevalence of severe coronary artery disease and the fact that more women than men fail to reach maximum aerobic capacity during exercise test can localise ischaemia as well as diagnose coronary stenoses in this group would increase the diagnostic power.

Some of the newer add-ons or substitutes for the usefulness of exercise testing is the fundamental incapacity of ST depression to localise ischaemia. Although suggested by the recurrent finding of V5 as the lead of maximum predictive power irrespective of the coronary disease pattern, several direct investigations have corroborated the result. The electrophysiological explanation for this apparent anomaly only came to light in 1998. Li and colleagues showed, with a computer simulation and a sheep model of subendocardial ischaemia, that ST depression is generated at the lateral epicardial boundaries of ischaemic territories are shared, and only those reaching the free wall contribute to external ST depression, this process cannot be said to localise ischaemia.

Some of the newer add-ons or substitutes for the exercise test can localise ischaemia as well as diagnose coronary disease when poor mobility precludes an exercise test or when the baseline ECG prohibits ST analysis (such as in the presence of more than 1 mm of resting ST depression, left bundle-branch block, or Wolf-Parkinson-White syndrome). Table 1 presents some of the results from meta-analysis and multicentre diagnostic studies. Electron-beam computed tomography (EBCT) is a fast radiographic technique that can make a quantitative measurement of coronary artery calcification.

Cardiokymography is a simple motion sensor that allows recording of the movement of the left ventricle with a small transducer placed on the chest wall. Signal averaging has enhanced this older technology. Nuclear perfusion imaging includes both the early thallium radiographic images and the modern single-photon-emitted computed tomography (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 (percentage of patients correctly classified) is a convenient way to compare tests. For instance, although the sensitivity and specificity for exercise testing and EBCT are nearly opposite, their predictive accuracy is similar. Altering their cut points by lowering the amount of ST segment depression or raising the calcium score would result in similar sensitivities and specificities (although this assumes a disease prevalence of 50%). The clear message, however, is that although the newer technologies seem to have better diagnostic characteristics than the basic

**Table 1: Comparison of tests used in the diagnosis of coronary artery disease**

<table>
<thead>
<tr>
<th>Grouping</th>
<th>Number of studies</th>
<th>Total patients</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Predictive accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meta-analysis of standard exercise ECG</td>
<td>147</td>
<td>44 047</td>
<td>68%</td>
<td>77%</td>
<td>73%</td>
</tr>
<tr>
<td>Excluding MI patients</td>
<td>41</td>
<td>11 691</td>
<td>67%</td>
<td>72%</td>
<td>69%</td>
</tr>
<tr>
<td>Limiting work up bias</td>
<td>2</td>
<td>&lt;1000</td>
<td>50%</td>
<td>90%</td>
<td>69%</td>
</tr>
<tr>
<td>Meta-analysis of exercise test scores</td>
<td>24</td>
<td>11 788</td>
<td>-</td>
<td>-</td>
<td>80%</td>
</tr>
<tr>
<td>Probability stratification by exercise test scores</td>
<td>1</td>
<td>2000</td>
<td>95%</td>
<td>90%</td>
<td>92%</td>
</tr>
<tr>
<td>Thallium scintigraphy</td>
<td>59</td>
<td>6038</td>
<td>85%</td>
<td>85%</td>
<td>85%</td>
</tr>
<tr>
<td>SPECT without MI</td>
<td>27</td>
<td>2136</td>
<td>86%</td>
<td>62%</td>
<td>74%</td>
</tr>
<tr>
<td>Exercise echocardiography</td>
<td>58</td>
<td>5000</td>
<td>84%</td>
<td>75%</td>
<td>80%</td>
</tr>
<tr>
<td>Exercise echocardiography excluding MI patients</td>
<td>24</td>
<td>2109</td>
<td>87%</td>
<td>84%</td>
<td>85%</td>
</tr>
<tr>
<td>Non-exercise stress tests</td>
<td>Persantine thallium</td>
<td>11</td>
<td>1000</td>
<td>85%</td>
<td>91%</td>
</tr>
<tr>
<td>Dobutamine echocardiography</td>
<td>5</td>
<td>1000</td>
<td>88%</td>
<td>84%</td>
<td>86%</td>
</tr>
<tr>
<td>Cardiokymography</td>
<td>1</td>
<td>617</td>
<td>71%</td>
<td>88%</td>
<td>79%</td>
</tr>
<tr>
<td>Electron beam computed tomography</td>
<td>5</td>
<td>2373</td>
<td>90%</td>
<td>45%</td>
<td>81%</td>
</tr>
</tbody>
</table>

**Panel 1: Examples of multivariate equations showing weightings for various risk factors**

**Froelicher pretest**

\[ -2.1+0.03 \times \text{age} -(0.4 \times \text{symptoms})+(0.8 \times \text{diabetes})+(0.4 \times \text{hypercholesterolaemia})+(0.01 \times \text{pack-years})+(0.7 \times \text{resting ST depression in mm}) \]

**Froelicher post-test**

\[ -1.2+(3.3 \times \text{pretest})+(0.5 \times \text{exercise ST depression in mm})+(0.6 \times \text{slope})-(0.16 \times \text{MET})-(0.5 \times \text{exercise angina}) \]

**Morise pretest**

\[ -3.6+(0.08 \times \text{age})-(1.3 \times \text{sex})+(0.6 \times \text{symptoms})+(0.7 \times \text{diabetes})+(0.3 \times \text{smoking})-(1.5 \times \text{body surface area})+(0.5 \times \text{oestrogen})+(0.3 \times \text{number of risk factors})-(0.4 \times \text{resting ECG}) \]

**Morise post-test**

\[ -0.12+(4.5 \times \text{pretest})+(0.37 \times \text{mm ST depression})+(1.0 \times \text{slope})-(0.4 \times \text{negative ST})-(0.016 \times \text{maximum heart rate}) \]

**Detrano**

\[ 1.9+(0.025 \times \text{age})+(0.6 \times \text{sex})+(0.7 \times \text{symptoms})+(0.05 \times \text{MET})-(0.02 \times \text{maximum heart rate})+(0.36 \times \text{exercise-induced angina})+(0.6 \times \text{mm ST depression}) \]
showing that peak oxygen consumption ($VO_2$) was a significant univariate or multivariate predictor of risk in heart failure, and this variable is now regarded as one of the most potent prognostic markers in this condition. A landmark 1991 study engendered the notion of a cutpoint for patients who attained values above this point displaying significantly worse survival rates.

Although appealing to clinicians, the idea of a cutpoint associated with prognosis does not bear comparison with methods incorporating other clinical information such as the exercise-test scores.

### Prognosis and novel indications

The overriding paradigm in the diagnosis of cardiovascular disease over the past decades has been the presumptive identification of obstructive coronary disease in order to direct interventional angiography or bypass surgery. However, shortcomings in non-invasive tests such as the work-up bias mentioned above, combined with limitations fundamental to angiography have led some to suggest that the principal place for the exercise test is in the assessment of prognosis (table 2). The appropriate direction of life-saving pharmacological tools in the form of statins, angiotensin-converting-enzyme inhibitors, β-blockers, and aspirin can be enhanced by tests capable of identifying high-risk populations. The financial implications in a novel risk factor age are clear.

### Exercise and heart failure

The rise of exercise training as a treatment for heart failure has led to renewed interest in the assessment of patients with reduced ventricular function. Most of the work in this area has focused on pretransplant patients, and, in particular, the ability of ventilatory and gas-exchange measures to stratify risk and predict outcome. During the 1990s, over 40 studies were published showing that peak oxygen consumption ($VO_2$) was a significant univariate or multivariate predictor of risk in heart failure, and this variable is now regarded as one of the most potent prognostic markers in this condition. A landmark 1991 study engendered the notion of a cutpoint value for peak $VO_2$ (14 mL/kg per min), with patients who attained values above this point displaying 1 year and 2 year survival rates similar to those in transplanted patients, whereas those below this value had significantly worse survival rates.

Although appealing to clinicians, the idea of a cutpoint has been questioned. One study showed that ventilatory variables ($VE/VCO_2$) and a chronotropic index (panel 2) fared better than peak $VO_2$ in predicting death, whereas another study recorded similar discriminatory power for each of seven cutpoints between 10 mL/kg per min and 17 mL/kg per min. In fact, an earlier publication from the same group showed that peak $VO_2$ expressed as a continuous variable outperformed clinical measures, right heart catheterisation data, exercise time, and other exercise test data in predicting outcome. Such observations are very important in the current health-care climate, because they help direct scarce resources (eg, donor hearts) towards patients who are most likely to benefit.

As a result of these studies, the American College of Cardiology and American Heart Association have recommended the inclusion of gas analysis in exercise tests done to assess transplant patients and to differentiate cardiac exercise intolerance or dyspnoea from pulmonary causes. Much evidence suggests an important role for abnormally heightened ventilation, both as a marker of the severity of heart failure and of prognosis. This heightened ventilation has been expressed as an increase in the slope of the relation between minute ventilation and CO$_2$ production ($VE/VCO_2$ slope), a variable that has become common as an expression of the response to exercise in heart failure. Because this ratio is inversely related to cardiac output and parallels the degree of dead-space ventilation, it signifies the extent to which reduced pulmonary perfusion underlies the excess ventilatory response to exercise in part through a mismatching of ventilation and perfusion in the lung.

Data have suggested, however, that these techniques are underused. One survey found that less than 3% of exercise tests done in the Veterans Health Care system in the USA included gas-exchange measurements, despite more than 15% of the patients meeting the class-I criterion (panel 3). The promising data from randomised controlled trials of exercise training in heart failure suggest a class IIb indication for the use of exercise testing with gas analysis in the cardiac patient. Furthermore, the rising prevalence of heart failure demands an increasing role for this method of assessment for diagnosis, prognosis, or assessment of therapy indications.

### Heart rate

Exercise chronotropic incompetence has been known for some time to predict all-cause mortality in healthy populations. When combined with a measure of exercise capacity as the chronotropic index (panel 2) it is predictive of all-cause mortality independent of thallium...
ischaemia. In fact, the risk associated with relative bradycardia was equal to that of a perfusion defect, and the effects were additive. The mechanism for these associations is not clear. Control of exercise heart rate involves both sympathetic and parasympathetic input. Some form of autonomic paresis, which does not necessarily require localised perfusion defects, could accompany the chronic sympathetic overactivation of the heart-failure syndrome. We know for certain that β receptors are downregulated in heart failure, that chronotropic incompetence is predicted by increased exercise-induced dyspnoea or impaired exercise capacity when cause is uncertain.

One study found that a delayed decrease in heart rate during the first minute after graded exercise was a powerful and independent predictor of all-cause mortality in 2428 patients followed for 6 years. A potential difficulty with the measurement of heart-rate recovery for prognosis, however, is the use of a cool-down period, which is thought to negatively impact the diagnostic sensitivity of the test.

**Hypertension**

The clinical significance of an exaggerated diastolic blood-pressure rise to exercise has been recognised and debated for over 15 years. But there is now a clear indication that such a rise represents a risk factor for new-onset hypertension. The Framingham Offspring study showed that an exaggerated diastolic blood-pressure rise was associated with a two-fold to four-fold risk of new-onset hypertension. A raised recovery systolic blood pressure was also predictive of hypertension in men alone. However, the data in relation to angiographic disease and mortality are less clear cut. Some investigators reported a relation between exercise systolic blood pressure and mortality significantly greater than that with resting blood pressure. McHam and colleagues showed that although a delayed decline in systolic blood pressure during recovery is associated with severe angiographic coronary disease, exercise hypertension is not associated with an increased mortality rate. This finding may be partly explained by the different exercise methods used in these studies. Some workers prefer bicycle ergometry to treadmill testing for exercise blood-pressure measurements because the arm is kept more still. McHam and colleagues also showed that, as in heart-failure patients, exercise hypertension is negatively associated with myocardial perfusion abnormalities. What is clear is that for the moment the importance of exercise hypertension remains contentious.

**Exercise capacity**

The measurement of exercise capacity, made more difficult by staged exercise protocols, is often lost in the focus on the ST segment. In fact, exercise capacity turns out to be a strong independent risk factor for all-cause and cardiovascular mortality and one, moreover, that can be altered by training. Blair and colleagues showed that for every minute of increase in maximum treadmill time there was a corresponding 7.9% decrease in the risk of mortality. Another study that examined several factors including SPECT thallium-perfusion scanning, found...
that the strongest predictor of all-cause mortality was estimated fair or poor functional capacity (estimated simply from age and sex). These analyses show exercise capacity to be a powerful prognostic indicator, and one that, by contrast with the measures discussed above in relation to heart rate, implies a mechanism for its own amelioration. Goldstein and Holmboe have emphasised the importance of the potential to modify poor prognostic indicators, if we are to argue for measuring them.

Protocol

Experience and history show that the pervasiveness of an idea, method, or product relies only in part on its intrinsic quality. Rather more, the acceptance of an idea relies on its extrinsic survival potential, a complex attribute with temporal, logistical, and experiential dimensions. In applied exercise testing, these features are apparent in relation to the choice of treadmill protocol. When treadmill and cycle ergometer testing was first introduced into clinical practice, practitioners adopted protocols used by major researchers such as Balke, Astrand, Bruce, and Ellestad (figure 1). However, over time, those described by Bruce came to predominate. A survey of 71 cardiology divisions within the US Veterans Health Care system showed that the Bruce protocol was used by 82%, a number that is broadly in line with the clinical experience of the authors and several other surveys. This use is in contrast with the recommendations of major academic bodies that advocate gradual and individualised protocols. Their reasons are clear: work-rate increments that are too large or rapid result in lower sensitivity for detecting coronary disease, a less reliable test for studying the effects of therapy, and a tendency to overestimate exercise capacity. The advantage of the experience and data acquired with the Bruce protocol is outweighed by the many patients unable to complete even stage 1 (about 5 METS, see panel 2), the difficulty in interpreting gas-exchange measurements below maximum with large and unequal increments in work, and the poorer estimates of exercise capacity achieved compared with other protocols. In addition, since many clinical laboratories do not employ gas analysis routinely, a clear relation between those variables that are measured and oxygen uptake is desirable.

Ramp testing is an alternative to the Bruce treadmill protocol, which was first proposed in 1991 and is now taking on a higher profile (figure 2). This protocol, aided by a pretest activity questionnaire, aims to bring the patient from rest to maximum exercise through a linear increase in work over about 10 min. The rationale for this length of test has been well validated. Studies suggest that tests individualised to last 10 min produce the highest values for oxygen uptake, better differentiation of treatment effects in clinical trials, and a closer relation of work rate to oxygen uptake. In addition, avoidance of fatigue and a focus on total work done rather than on exercise time (a highly variable measure), is provided by this approach.

Despite these potential advantages there have been few direct comparisons of the ramp protocol with others. Of these, three groups of researchers investigated cardiovascular patients whereas two studied healthy women and obese women. Two studies used a ramped Bruce protocol (ramping between the stages of the classic Bruce) whereas the others used individualised protocols. Taken overall, the findings from these studies suggest that a ramped protocol is preferred and better tolerated by patients, produces an oxygen uptake-to-work ratio closer to unity (in some patients but not those over 60 years), and may result in higher values for metabolic equivalents and exercise duration. Although this last finding was reported in only one study, such an observation has relevance for the predictive characteristics of the test. The higher workloads achieved by the same patients on the treadmill compared with the cycle ergometer are associated with improved exercise-test sensitivity for coronary artery disease. It makes intuitive sense, given two tests in the same patient, that the one capable of eliciting a higher oxygen uptake would represent a truer examination of cardiopulmonary function (rather than the limitations of local muscle fatigue).

The generalist is increasingly called upon to begin the work-up of the cardiac patient and, therefore, potential enhancement of non-invasive tests should be welcomed. As such, developments summarised in this paper give the test a new lease of life. The inertia that sees the majority of centres still using unequal stage and rapidly incremented exercise protocols should be overcome. Well-validated prediction equations and novel diagnostic markers which lend increased value for predictive-accuracy should be adopted. Finally, major advances in the use of the test for prognosis, with the discovery of novel risk factors and the addition of gas analysis, may yet shift the primary emphasis away from diagnosis. The exercise test is inexpensive, brief, requires minimum equipment and space, and can be done in most cases safely without the presence of a cardiologist. Par from fading away, the technique first proposed in 1932 as an aid to the diagnosis of angina is in the process of being reborn.

References

Figure 2: Schematic representation of the ramp treadmill test

After a 1 min warm-up at 2 km/h and zero gradient, the rate of change of speed and grade is individualised to yield a work rate corresponding to the estimated maximum exercise capacity (y axis) in about 10 min.


For personal use only. Not to be reproduced without permission of The Lancet.