On the uniformity of cardiopulmonary exercise testing in chronic heart failure

Jonathan Myers, PhD Palo Alto, Calif

See related article on page 466.

Peak oxygen uptake during exercise (peak VO₂) has evolved to become a vital component of the risk paradigm in patients with chronic heart failure (CHF), particularly in the context of considering patients for transplantation listing. Over the last 10 years, dozens of studies have been published demonstrating that peak VO₂ is a strong univariate or multivariate predictor of outcomes in CHF. Although “outcomes” have differed in these studies, in many instances peak VO₂ has outperformed clinical, invasive hemodynamic and other exercise data in predicting risk.¹ Such information is extremely important today given the large and widening gap between available donor hearts and potential transplant recipients.

Although peak VO₂ is usually assumed to be an objective and unqualified number, any value ascribed to a given patient carries some inherent qualifications. Few practitioners appreciate the variety of factors that influence this number. Aside from the obvious physiologic factors that determine peak VO₂, the measurement is also influenced by psychologic and pragmatic factors. Psychologic factors that influence peak VO₂ include coaching by testing personnel, feedback, and variation in the patient’s ability to tolerate discomfort.²-⁴ Pragmatic factors that influence peak VO₂ include the exercise protocol, the exercise mode, whether VO₂ is expressed as an absolute value or as a percentage of an age-predicted standard, and how the data are sampled (eg, 10-second, 30-second, 60-second, or various breath-averaging methods). Because a great deal of functional and prognostic data have been generated over several decades of its use, however, exercise on the cycle ergometer yields peak VO₂ values significantly lower than those on the treadmill, with differences ranging from 5% to 25%.⁶,¹⁴-¹⁷ Thus the exercise mode alone could potentially cause a patient to fall into or out of the conventional cutoff level for transplant listing. In the study of Strzelczyk et al peak VO₂ on the cycle ergometer was 15% lower than that on the modified Naughton treadmill protocol, findings similar to those from our laboratory⁶ and others.¹⁴,¹⁷ For particular patients with CHF these differences in peak VO₂ were as much as 5 to 7 mL/kg/min.

Strzelczyk et al found that the Bruce and modified Naughton protocols resulted in similar values for peak VO₂ and VO₂ at the ventilatory threshold. This is important clinically in that, from the standpoint of measuring the functional limits of the patient, it suggests that the widely used Bruce test is an acceptable protocol for patients with severe CHF. The advantages of the Bruce protocol include its familiarity, time efficiency, and the fact that a great deal of functional and prognostic data have been generated over several decades of its use. Thus many published normative values have been derived from it. For example, the most robust databases on the use of the exercise test for assessing prognosis, such as those from the Coronary Artery Surgery Study (CASS)¹⁸ and the Duke Treadmill Score¹⁹ were generated from patients who underwent exercise testing with the Bruce protocol. A recent survey of exercise laboratories within the Veterans Administration indicated that 82% of laboratories used the Bruce or modified Bruce protocols for routine testing.²⁰ These and other data have documented that the Bruce is by far the predominant protocol clinically, probably for reasons of tradition, convenience, familiarity, or their combination.

However, exercise testing guidelines (eg, American Heart Association, American College of Sports Medicine, and European²¹-²³) published over the past 10 years have been consistent in their recommendation that the exercise protocol should be individualized for the patient being tested and the purpose of the test. It has also been recommended that protocols with more gradual increments (such as Naughton or ramp) are more appropriate for patients with cardiovascular disease. This recommendation has evolved from studies demonstrating that work increments that are relatively large or rapid result in (1) a tendency to overestimate exercise
capacity. (2) a less reliable test for studying the effects of therapy, (3) decreased sensitivity for detecting coronary disease, and (4) less uniform and reproducible hemodynamic and gas exchange responses to exercise.6,17,21-27 Intuitively, the standard Bruce test is also problematic in that the initial stage of the protocol (5 metabolic equivalents) is too demanding for many patients with cardiovascular disease, particularly those with CHF. The predicted oxygen cost of this initial stage (17.5 mL/kg/min) is higher than the measured VO2 many patients with CHF achieve at maximal exercise. The mean measured value in the study of Strzelczyk et al was 17.7 mL/kg/min, suggesting that for many patients a wide discrepancy exists between physiologic work (measured VO2) and the external work rate. In effect, the protocol would increase the oxygen deficit, meaning that a greater proportion of the energy requirements is produced from oxygen-independent (anaerobic) glycolysis. In practical terms, adenosine triphosphate production during this type of exercise is considerably less efficient and causes greater metabolic acidosis and therefore hyperventilation, all having the potential to reduce exercise capacity.

It has also been demonstrated that protocols individualized to last approximately 10 minutes yield the highest values for peak VO2.17 This, along with the fact that a protocol that is individualized tends to have a closer relationship between the external work rate performed and measured VO2 throughout exercise, is the underlying basis by which the guidelines recommend individualizing the exercise protocol to last between 8 and 12 minutes.21-23 Exercise duration can also influence the limiting symptom in patients with CHF. Lipkin et al28 observed that 23 of 25 patients were limited by fatigue during a relatively slow protocol, whereas these same patients were all limited by shortness of breath when tested with a rapidly incremented protocol. Most likely for the reasons cited above, the rapid protocol also yielded greater evidence of hyperventilation during exercise. The latter has the potential to cause an earlier occurrence of the anaerobic threshold and thus a reduction in exercise capacity.

An important question raised by the study by Strzelczyk et al is whether the reservations about rapidly incremented protocols contained in the exercise testing guidelines21-25 and experimental data6,17,24-28 are outweighed by the time saved (3.5 minutes), familiarity, and portability of the Bruce protocol. Despite the potential physiologic consequences of rapid, more demanding protocols discussed above, there were no differences in peak VO2 percent predicted VO2, or VO2 at the anaerobic threshold between the Bruce and Naughton protocols in the study of Strzelczyk et al. Although this question cannot be completely resolved by these 15 patients, the article brings to light the lack of standardization in methods for testing patients with severe CHF. Because the exercise test plays a critical role as a major "gate-keeper" to transplant listing, future editions of the transplant recipient guidelines should address this issue.

**Cardiopulmonary exercise testing and data sampling**

While on the topic of subtle but potentially important differences in peak VO2 attributable to methods, the wide variation in data sampling warrants comment. This is a problem that has evolved with technologic advances in on-line acquisition of gas exchange data. The current computerized systems permit the user to acquire, sample, and express the data in innumerable ways, and there is a great deal of variation in the methods by which laboratories express gas exchange data. Although this is rarely reported in the methods among studies addressing prognosis and cardiopulmonary exercise testing in CHF, it has at least as potent an effect as the protocol on the value used for "peak VO2." Several groups have shown that a given VO2 value can vary in the order of 20% depending on the sampling interval chosen.29-31 Although a number of studies assessing prognosis in CHF have reported their data using 30-second intervals, a wide range of intervals have also been used, potentially having a significant effect on the VO2 value used for stratifying a patient’s risk. In addition, it is rarely specified whether the sample reported was a “rolling average” (running recursive sum). The latter is important because it affects the precision of VO2 relative to the test end point. In practice, however, it is common to take the value rounded to the nearest minute or half minute. It has been recommended that the data be averaged for about 30 seconds but reported in rolling averages every 10 seconds.32 This smooths the data yet permits adequate resolution for choosing test end points, the ventilatory threshold, and other relevant analysis points. Standardization would go a long way toward improving the portability of findings from one group of investigators to another. Regardless of the specific sampling interval chosen, studies should always report the specific sampling interval used.

**Is there an optimal cut point for peak VO2?**

A peak VO2 of 14 mL/kg/min has become widely recognized as a cut point for categorizing patients into groups who are likely to survive and those who are not likely to survive over a given follow-up period. This cut point is based on observations that patients who achieve peak VO2 values ≥14 mL/kg/min have a survival rate similar to that of patients who receive transplants (approximately 90% survival at 1 year).1,33-35 This implies that transplantation can be safely deferred in these patients. However, as pointed out by Strzelczyk et al,5 the cut point has often been chosen arbitrarily in the studies associating peak VO2 with survival in CHF.
Although the use of a cut point such as 14 mL/kg/min is inherently attractive to clinicians, it has several limitations. First, it is often difficult to separate a given patient’s true cardiopulmonary limits from motivation or deconditioning, particularly in a population of patients with severe heart failure. Second, although stratifying patients above and below 14 mL/kg/min has demonstrated marked differences in survival, so too have values above versus below 10 mL/kg/min,33-35,38 11 mL/kg/min,39,40 12 mL/kg/min,40,41 13 mL/kg/min,42 15 mL/kg/min,43 and a variety of exercise tolerance values predicted from exercise workload or time. Third, dichotomizing the data in this way tends to oversimplify the issue by forcing patients into 1 of 2 categories, compromising the ability to carefully assess those in the most clinically relevant range of interest, in this case 10 to 15 mL/kg/min. Moreover, not all studies have found 14 mL/kg/min to be particularly sensitive for separating survivors from non-survivors. Two recent studies have suggested that peak VO2 has limited prognostic value within the range of 10 to 18 mL/kg/min44 and 12 to 17 mL/kg/min.45 In these studies peak VO2 was strongly associated with mortality only among patients achieving values below the lower limit and strongly predicted survival among those achieving above the higher limit. In reviewing the literature, Strzelczyk et al have made the important point that there has been wide variation in both the exercise mode and protocol used in these studies.

It was recently demonstrated that 14 mL/kg/min was not superior to other cut points ranging between 10 and 17 mL/kg/min.40 An improvement in survival of approximately 20% was observed for patients achieving a high versus a low peak VO2 irrespective of the cut point used within this range. Although this knowledge (20% gain or reduction in risk) contributes greatly to the estimation of risk in patients considered for transplant listing and comes from a noninvasive and relatively inexpensive procedure, it does not indicate a particular advantage for one cut point over another. An “optimal” cut point is likely to change depending on the severity of disease in the population. This, along with the knowledge that there are a number of other clinical, exercise, and hemodynamic variables that stratify risk, suggests that peak VO2 might be better considered as a continuous variable in a multivariate model. A number of such models have been developed that have been shown to powerful stratify risk in patients considered for transplantation.36,37,46

Summary

The continued high mortality rate and widening gap between patients listed for transplantation and available donor hearts have magnified the need for reliable prognostic markers in CHF. Currently a great deal of faith is placed on peak VO2 for stratifying risk in patients with CHF, and this faith has substantial justification the literature.1 Without question, peak VO2 is one of the most important markers of risk in patients with CHF. Yet comparatively large differences in peak VO2 can occur as a result of differences in the exercise mode (eg, 2.4 mL/kg/min difference in the study of Strzelczyk et al,5 4.0 mL/kg/min in our CHF patients5 for Bruce treadmill versus cycle ergometer protocols). Similarly, the method of sampling the data has been inconsistent in the literature, and differences in the order of 20% have been observed attributable to differences in the method of sampling and summarizing the data.29-31 The effects of exercise mode and method of reporting the data on peak VO2 suggest the need for uniformity when exercise testing patients considered for transplant listing. Finally, clinicians should resist the tendency to rely strictly on a single cut point in peak VO2 to list patients for transplantation. Numerous variables have been shown in various studies to predict outcomes in patients with CHF; over the last 20 years more than 150 clinical, hemodynamic, or exercise variables have been identified as predictors of mortality. Obviously, many are interrelated, and it is not possible to measure them all in a given study. This underscores the fact that exercise capacity should not be the sole determinant in any risk-stratification model or in transplant candidate selection. Assessment of risk in patients with heart failure remains largely an art form; an appropriate risk paradigm should consider a variety of prognostic markers.

References
