Editorial Comment

The Academic Life Cycle of a Noninvasive Test

David B. Pryor, MD

The article of Lachterman et al1 in this issue of Circulation examines the limitation of the ST/heart rate index compared with standard ST criteria for the interpretation of the exercise electrocardiogram, and the findings reported disagree with previously published reports.2,3 The study raises two important issues: Why the disagreement? and what are the implications for future research and reporting?

It is tempting to dismiss the disagreement as a consequence of poor-quality research on the part of one of the groups advancing or negating the concept of the value of the heart rate–adjusted index. Such an interpretation should be avoided. Proponents of both views have published extensively in the field and are well-recognized investigators. Furthermore, as reviewed in the article by Lachterman et al, many different groups have considered the problem with varying results. To understand the disagreement, therefore, requires a broader perspective.

See p 44

It is helpful in understanding the disagreement to consider the academic life cycle of an “imperfect” noninvasive test such as the exercise treadmill test. Most published work regarding the test can be divided into one of six categories: 1) the original description; 2) the limitations of the original approach; 3) new advances—modifications, refinements, or applications of the test; 4) further limitations of the new advances; 5) review articles; or 6) studies that use the exercise test as an outcome measure (e.g., for drug evaluations). Within this framework, the work of Lachterman et al1 would fall into the fourth category, whereas that of Elamin et al2 and Klingfield et al3 would fall into the third category.

The problem is pervasive. The disagreement over the value of the ST/heart rate index has been duplicated for most of the “advances” described in exercise testing. Original promising reports of additional parameters, provocative stimuli, methods of testing, incorporation of radionuclide procedures, computerization of results, etc., have nearly all been followed by reports citing limitations when the “new” techniques have been applied to subsequent populations. This is not to say that important refinements have not occurred, indeed the methods of testing have clearly improved, but rather to note that conclusions based on initial promising reports inevitably become tempered with subsequent evaluation.

The magnitude of the problem is easily recognized by considering the vast array of literature reported concerning exercise testing. An Index Medicus search for 1970–1988 limited to English articles in humans with “treadmill” in the title or “exercise exertion or test” in the title with “treadmill” in the abstract produces 2,245 citations. If the search is limited to articles with “treadmill” in the title, 559 citations still result. With the classification scheme described above, 165 articles describe refinements or new applications of the technique, and 52 describe limitations for patients with coronary artery disease (based on title review only). Thus, the disagreement of the findings of Lachterman et al1 with those previously reported for the value of the ST/heart rate index can only be understood in the broader context of the methodology used for the evaluation of noninvasive tests.

The methodology used to evaluate noninvasive tests is unfortunately more difficult to understand than most clinicians would desire. Ideally, the clinician would read the results of the published study, learn how to make the appropriate measurement, and then use the measurement in the new patient being evaluated to determine the likelihood of an important outcome (e.g., the likelihood of coronary artery disease or anatomically severe coronary artery disease). In this context, the clinician has used the results of the published study to form a clinical prediction rule that is then applied to a patient. In addition to understanding how to make the prediction, the clinician should also consider two other relevant questions: How helpful will the measure-
ment be? and how likely is it to apply to my patient? The methodology for evaluating a noninvasive test can be considered as a matrix of three related issues that address three fundamental questions to the clinician.4 The corresponding issues to the three clinical questions—how do I make a prediction? how helpful will it be? and how likely is it to apply to my patient?—are the predictive method, the measurement of prediction quality, and the generalizability of the findings.

In the simplest case, a single finding on the exercise test, for example, an ST/heart rate index of greater than 0.024 mV/beats/min, corresponds to a probability of 100% for significant coronary artery disease. Although appealing in its simplicity, such an approach is rarely sufficient with an imperfect test. Should the prediction rule be stratified for the presence of previous infarction, the presence or type of chest pain, the age or gender of the patient? As more factors are considered, the clinical prediction rule becomes more complicated requiring the use of tables, nomograms, or equations to apply. Although not always necessary, it is inevitable that mounting pressures to develop more cost-effective approaches to the evaluation of patients with coronary artery disease will force clinicians to improve their understanding and facility in using alternative methodologies for the development of clinical prediction rules.

Three separate aspects of quality of a prediction complicate its assessment. Lachterman et al1 examine one such measure—discrimination. Discrimination is the ability to separate patients with an outcome from those without an outcome. Traditional measures of discrimination reported in the literature are sensitivity (among patients with the outcome, the proportion with a positive test) and specificity (among patients without the outcome, the proportion with a negative test). The clinician wishing to apply the results does not know whether the patient has the outcome or not—that is the reason for performing the test. The physician must convert the sensitivity and specificity to predictive accuracy of a positive or negative test (among all patients with a positive or negative test, what proportion do or do not have disease?). Simple formulations of Bayes’ rule have often been used to accomplish this.

The simple reporting of sensitivity and specificity is inadequate, particularly for test result measures that are continuous. By varying the cutoff point of the measure used to denote a positive or negative test, sensitivity can be increased at the expense of specificity or vice versa. For that reason, it has become more common for investigators to report ROC (receiver operating characteristic) curves that present the continuous spectrum of sensitivity and specificity as the cutoff point of the test is varied. The area under the ROC curve is then compared between the two measures to determine whether one test measure is significantly better than another (as was done by Lachterman and colleagues1).

Although a detailed discussion of the methodology for assessing the quality of a prediction4 is beyond the scope of this editorial comment, two fundamental issues should be recognized by the clinician. The first is that measures of sensitivity and specificity will vary depending on the characteristics of the population or the testing methodology.3 The second is that while discrimination is important, the two other components of predictive quality may need to be considered. Reliability or accuracy refers to how close a given predictive estimate corresponds to the true estimate, and precision refers to the confidence limits of a predictive estimate. Particularly, where the clinician makes a decision based on a probability rule (i.e., catheterize the patient when the probability of disease is greater than 85%), unreliable or imprecise estimates may lead to incorrect decisions.

The methodological counterpart to the clinician’s question (i.e., how likely will the prediction rule apply to my patient?) is the generalizability of the estimate. Generalizability or validation strategies can be subdivided into internal or external techniques. Internal techniques (e.g., jackknife or bootstrap procedures) use the same population on which the prediction rule is developed to estimate its generalizability. In general, such approaches underestimate bias. External techniques apply the prediction rule to an independent population. The rigor of the validation strategy can be considered from a perspective of how different the population sample used for the validation study is from that in which the clinical prediction rule was developed. This may range from a randomly divided population in which one subset is used to develop the prediction rule (the training sample) and the other to validate it (the test sample) to a strategy where the clinical prediction rule is developed in a population and then prospectively applied to a new population of patients at a different institution with different referral characteristics.

In attempting to explain the differences in their results, Lachterman et al1 have considered six forms of bias or technical differences that may have been partly responsible for the disagreement with results from earlier reports. Philbrick et al6 describe seven methodological standards for the evaluation of exercise testing studies. A more complete description of 35 potential sources of bias that may apply to observational studies has been described by Sackett.7 Given the difficulty in recognizing the sources of bias that may influence the generalizability of the test results, the validation strategy must be critically examined by the clinician to have confidence that the finding will apply to a particular patient.

With a broader perspective in mind, it is not surprising that the work of Lachterman et al1 disagrees with previous studies. Recognized and unrecognized sources of bias are likely present in both the initial optimistic reports of Elamin et al2 and the pessimistic reports of Lachterman et al1. Given the wealth of studies that report the importance of the ST segment displacement, its magnitude, and the timing of this
change, we must conclude that exercise test results span a continuum from strongly positive to strongly negative tests. The appropriate application of a test result to an individual patient must consider the characteristics of the patient, many characteristics from the test, and their value in the individual clinical situation.

Both the use of the exercise test by a clinician and the publication of results by investigators will be improved by recognizing the best approaches for developing clinical prediction rules and the assessment of their quality and generalizability. Refinements in exercise testing techniques that solve all clinical problems in evaluating patients with coronary artery disease should be greeted with appropriate skepticism by practicing clinicians. When I was in medical school, my father, a practicing internist, discussing the problem of staying current with the medical literature, commented that half of everything he knew changed every 5 years, he just did not know which half it would be (personal communication, 1974). Perhaps this figure can be improved on as our understanding of the methodology of clinical research improves.

(Circulation 1990;82:302–304)

Acknowledgments

The author appreciates the expert assistance of Susan Feingloss for the literature search and Mary Scharenbroich for clerical support.

References

The academic life cycle of a noninvasive test.
D B Pryor

Circulation. 1990;82:302-304
doi: 10.1161/01.CIR.82.1.302

Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 1990 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on
the World Wide Web at:
http://circ.ahajournals.org/content/82/1/302.citation