EDITORIAL

Referral Bias and the Efficacy of Radionuclide Stress Tests: Problems and Solutions

The assessment of test technology is a complex science. In this issue of *The Journal of Nuclear Medicine,* Allman et al. evaluate the ability of adenosine thallium scintigraphy to assess the extent and location of coronary disease (*1*). They begin their discussion by noting that while this test had high sensitivity for coronary disease detection in their study, it "appears to be of lower diagnostic accuracy in characterizing [the] extent and distribution of individual coronary artery stenosis" (*1*). Yet overall diagnostic accuracy was also low in their study: the high sensitivity of 93% was accompanied by a very low specificity of only 25%. The investigators attribute this very low specificity to "an extreme case of post-test referral bias." Notably, they state that "this bias affects only performance for disease detection." They assert that the comparisons between thallium scintigraphy and the extent of angiographic disease are independent of referral bias, so that post-test referral bias "is not present in the respective curves for disease extent and localization." In fact, "post-test" and other referral bias can affect all of the common clinical end-points for which noninvasive testing is employed, including evaluation of disease extent. Since referral bias is rampant in studies involving the assessment of imaging technology, this editorial will review current pitfalls and potential solutions to this problem.

POST-TEST REFERRAL BIAS AND DIAGNOSTIC TEST ACCURACY

"Post-test referral bias" (*2*) becomes operative whenever noninvasive tests are used to select the angiographic population upon which subsequent assessments of test efficacy are based. The effects of "positive" post-test referral bias—the preferential referral of positive noninvasive test responders to angiography—can be profound. For example, in initial validation studies, exercise radionuclide ventriculography had very high sensitivity and specificity. Based on such studies, it became a rational common practice to preferentially select positive test responders for angiography and negative test responders away from angiography. Taken to the extreme, the post-test referral of *only* positive noninvasive test responders to coronary angiography would result in an "apparent" (but artifactual) increase in test sensitivity of 100% but an "apparent" (but artifactual) specificity of 0% (due to the exclusion of all negative test responders) in the resultant angiographic population. In reality, the "apparent" specificity of exercise radionuclide ventriculography fell dramatically over a 5-yr period in our laboratory (Fig. 1), and a similar temporal fall was noted in the specificity for planar exercise thallium scintigraphy (*3*). If unrecognized, this type of referral bias can lead to a confusing discordance among published data. For instance, if one divides the angiographic normals in Figure 1 into those who were studied during the "validation" phase for exercise radionuclide ventriculography (the first 2 yr) and the "clinical" phase (the next 3 yr), markedly different responses are noted (Fig. 2). Which set of responses is correct? Clearly, one will derive a very different conclusion regarding the definition of a "normal" ejection fraction response, dependent on which group is selected. This issue is discussed further elsewhere (*4*).

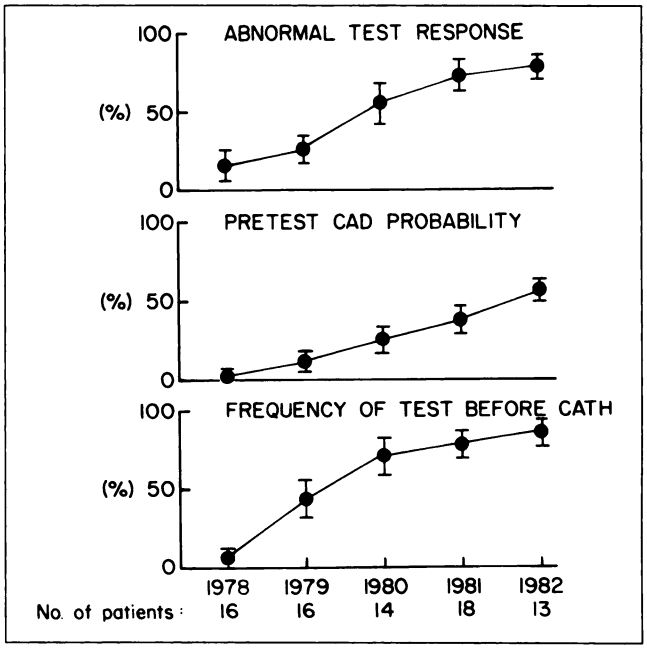

FIGURE 1. Illustration of how pre-test and post-test referral bias caused an "apparent" but artifactual declining specificity of exercise radionuclide ventriculography after its introduction into clinical practice. The top panel illustrates the frequency of false-positive exercise RNV responses (vertical axis) in angiographic normals over a temporal 5 yr period in our laboratory. The middle panel illustrates the mean pre-test probability of CAD in these angiographic normals, exclusive of the results of exercise RNV. Pre-test CAD probability rose progressively over this period. The bottom panel indicates the percent of patients in whom exercise RNV (TEST) preceded the performance of coronary angiography (CATH). Early, in the validation phase, the performance of exercise RNV before coronary angiography was very rare (6%). As clinical acceptance of the test proceeded, this sequence reversed completely. (Reprinted with permission from Reference 2).

Received Aug. 19, 1992; revision accepted Aug. 24, 1992.
For reprints contact: Alan Rozanski, MD, Division of Cardiology, St. Luke's-Roosevelt Hospital Center, 114th Street at Amsterdam Avenue, New York, NY 10025.

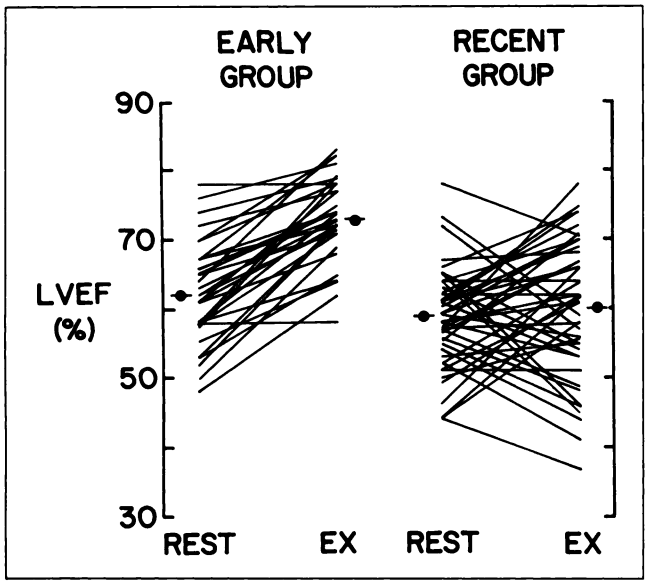

FIGURE 2. The angiographic normals that comprise Figure 1 are divided into an "Early Group" (1978–79) who were recruited during the validation phase of the test in our laboratory, and the subsequent angiographic normals ("Recent Group") who were garnered once exercise RNV was used in clinical practice. Shown are the rest and peak exercise (EX) values for left ventricular ejection fraction (LVEF) for each angiographic normal. In the Early Group, there was a significant increase in mean LVEF with exercise, with a uniform increase in most patients. In the Recent Group, the results are strikingly different. Many of the patients demonstrated severe falls in LVEF with exercise, and mean LVEF was unchanged from rest to exercise. Even though both groups of patients had the same normal angiographic findings, referral bias markedly distorted the characteristic LVEF responses in the Recent Group.

THE EFFECT OF POST-TEST REFERRAL BIAS ON OTHER CLINICAL END-POINTS

Post-test referral bias can also affect other indices of test efficacy. With respect to prognostic testing, a secondary bias becomes operative: the patients who are referred to angiography, preferentially selected on the basis of positive noninvasive stress tests, are also more likely to undergo early revascularization than patients not referred to angiography. Since early revascularization causes reduction in cardiac events in ischemic patients (*5*), the "natural" cardiac event rate in patients with a positive noninvasive test result are thus reduced, without the same impact on the negative test responders (who are preferentially referred away from angiography). The potential impact of this referral bias is best understood if it is also taken to its extreme: if *all* positive test responders would be referred for revascularization following angiography, any observed cardiac events would be paradoxically concentrated in the remaining patients with negative test responses!

Similarly, post-test referral could affect the type of analysis undertaken by Allman et al.: comparison of radionuclide test results versus the angiographic measurement of coronary disease extent and severity. Again, if we consider the extreme case of referring only positive test responders to cardiac catheterization, then those patients with single-vessel coronary disease and with multivessel disease would have positive test results. In reality, patients with single-vessel disease are more likely to have a negative test response than patients with multivessel coronary disease (*6–7*), but this observation would be completely obscured in the face of such "positive" post-test referral bias.

Indeed, the data by Allman et al., in which an "extreme" form of referral post-test referral bias is noted by the authors, supports this last point. Bias is initially suspected because their results are based on only a small sample of the patients referred for adenosine testing (76/1100 = 6.9%). The preferential referral of "positive" test responders is supported by the very high false-positive rate of 75%. The distorted impact that this referral bias has on estimating disease severity is indicated by Figure 3 of their study, which compares the results of angiography to the thallium severity score. In this analysis, there was no patient with a thallium severity score of less than four. The lack of patients with normal thallium scores, while a predictable consequence of post-test referral bias, is notable. Positive post-test referral bias will magnify mean thallium scores as well as reduce the potential range of thallium scores that might serve to distinguish patients with single- and multivessel coronary disease.

OTHER REFERRAL BIASES

Other biases can also affect the assessment of test efficacy (*2,4,8*). An important one, often overlooked, is "spectrum bias" (*8*) or what we have termed "pre-test referral bias"—the examination of test efficacy in only limited (unrepresentative) samples of the representative patient population (*4*). In its extreme, this bias consists of studies which primarily compare "the sickest of the sick" to the "wellest of the well" (*2*). We also noted this bias to be operative in our exploration of the "declining specificity" of exercise radionuclide ventriculography (*2*). In our initial experience, when the observed specificity of exercise radionuclide ventriculography was high, the pre-test probability of coronary disease among our angiographic normals was only 3% (Fig. 1). Later, as the observed specificity of exercise radionuclide ventriculography fell progressively, the probability of coronary disease *prior* to radionuclide testing in our angiographic normals rose progressively to a mean of 56%. This difference was no accident. The initial validation was based on recruitment of selected patients from the catheterization lab after performance of coronary angiography. These patients had low CAD probability because "healthy" angiographic normals were generally selected (i.e., they were preferentially selected on the basis of lack of symptoms or normal exercise ECG responses). But, in clinical practice, a certain percentage of angiographic normals will have angina and/or positive exercise ECG responses. Thus, it was natural for pre-test CAD probability to increase as the testing process switched

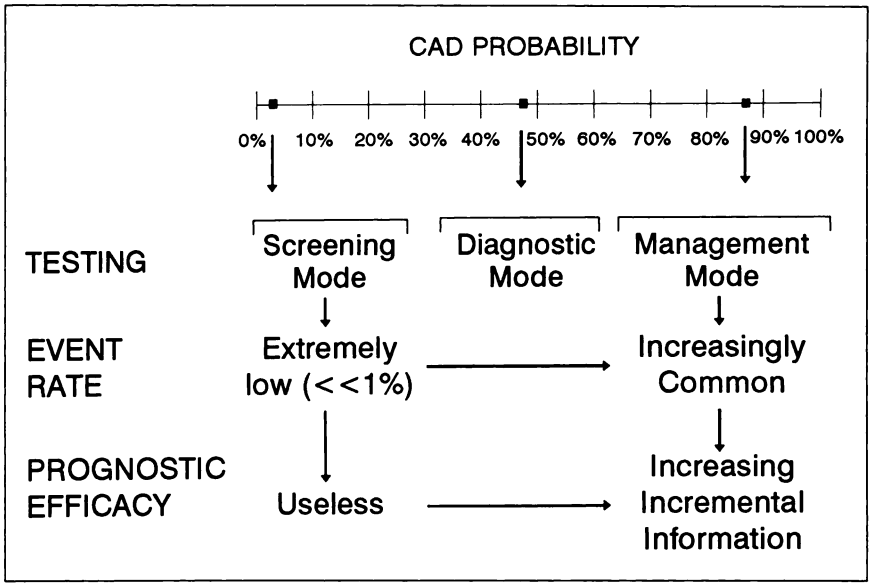

FIGURE 3. Schema for the clinical application of noninvasive stress testing and its impact on the assessment of prognostic test efficacy. When a noninvasive stress test is applied to individuals with a very low pre-test likelihood of disease, characterized by relatively young age and lack of anginal symptoms, the test is being used as a "screening" test. The application of noninvasive stress testing in patients with an intermediate likelihood of CAD represents the use of this test in its "diagnostic" mode. When patients have a high pre-test likelihood of CAD, noninvasive stress testing is not needed for diagnostic purposes, but it may be used for prognostic purposes ("management" mode). Bayesian analysis predicts—and clinical experience has demonstrated—that when noninvasive stress tests are applied in the screening mode, positive stress tests are uncommon and most commonly false-positive. The opposite is true in the management mode. Thus, testing for "diagnosis" is of limited value in the screening mode. Similarly, prognostic efficacy is also limited in the screening mode, because the baseline cardiac event rate is extremely low in "screening" populations. Even in the screening subgroup with positive noninvasive test results, the short-term cardiac event rate remains low—not significantly different from that noted in the subgroup with negative noninvasive test responses. As testing switches from the screening to diagnostic to prognostic mode, prognostic efficacy increases incrementally.

from the "validation" phase to the "clinical" phase in our laboratory.

An example of pre-test referral bias is the erroneous use of post-MI patients in the calculation of test sensitivity (*2*). *Pre*-test referral bias is most obvious, however, in terms of the current ongoing attempts by investigators to avoid the effects of *post*-test referral bias on the measurement of test specificity in their studies. To avoid the reliance on angiographic "normals", investigators have often turned to alternative "normal" populations, such as volunteer individuals or uncatheterized patients with low CAD likelihood (*4*). These groups have useful purposes—such as the establishment of normal test limits (*4*)—but their use as a reference standard for specificity is an erroneous practice (*4,9*).

Besides its effect on sensitivity and specificity, pre-test referral bias can also affect the assessment of prognostic efficacy. To understand this, consider the schema in Figure 3. Noninvasive stress testing can vary in its indications: from screening, to diagnosis, to its use in management decisions. Generally, these different modes of testing are applied in very different subgroups of individuals. Consider a hypothetical extreme: when a noninvasive stress test is applied—in its screening mode—to a selected group of asymptomatic individuals (e.g., if it were applied to asymptomatic airline pilots), one would expect a very low frequency of positive test responses. Many of these would be false-positive responses. But even in those with true-positive test responses, the one-year cardiac event rate would be extremely low (since these individuals were identified at a generally "latent" phase of their disease). If conclusions were based only on this subgroup or any other subgroup of extremely low risk individuals, *any* noninvasive stress test would necessarily appear to have relatively poor prognostic efficacy (i.e., because the cardiac event rate would appear low in both the positive and negative subgroups).

POTENTIAL SOLUTIONS

Referral bias has become ubiquitous, due to the acceptance of radionuclide stress testing as a decision guide for management decisions in cardiology. This ubiquity, however, represents a challenge to investigators seeking to validate new technology. There is ongoing need for such evaluation due to continued evolution of technology such as the current introduction of new isotopes (e.g., ^{99m}Tc-teboroxine, ^{99m}Tc-sestamibi), for example, or the introduction of new forms of pharmacologic stress (e.g., dobutamine). The determination of the "true" accuracy of current prognostic and new diagnostic tests in the face of ongoing

TABLE 1
Potential Methods for Reducing the Impact of Pre-test and Post-test Referral Biases

1. Evaluate the utility of debiasing algorithms
2. Consider prospective clinical trials *post*-angiography
3. Provide a full description of the patient population in studies
4. Tabulate individual patient results
5. Develop a date base registry
6. Adopt formal guidelines for reporting test data

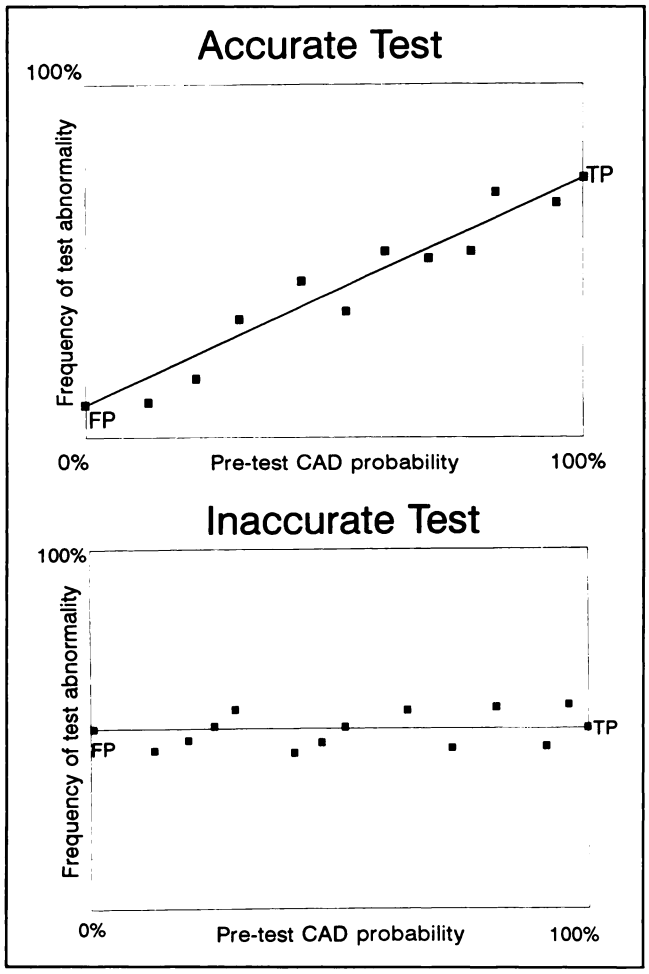

FIGURE 4. Schematic illustration of a probabilistic model for calculating sensitivity and specificity of a noninvasive test in "unbiased" fashion, without reference to coronary angiography. According to this proposal, the Bayesian probability of coronary disease would be calculated in EVERY diagnostic patient referred for stress testing. For each decile of CAD probability, the frequency of test abnormality would be determined. The figure illustrates a plot of frequency of test abnormality versus disease probability for two hypothetical sets of data. At the top are the results for an "accurate" test. The y-intercept at 0% CAD probability represents the false-positive (FP) rate of the test. It represents the number of times the test is abnormal in the absence of disease. The y-intercept at 100% CAD probability represents the true-positive (TP) rate, or sensitivity, of the test. For the "accurate" test, the calculated false-positive rate is low, and the true-positive rate is high. If a test was completely inaccurate, without discriminatory ability (bottom), the frequency of test abnormality would be the same regardless of the pre-test probability of disease, and a straight line would thus be obtained.

referral bias represents a daunting problem. Table 1 lists potential steps which may help alleviate this problem. These are discussed below.

APPLICATION OF DEBIASING TECHNIQUES

Algorithms have recently been developed for "debiasing" data that is influenced by post-test referral bias. One potential algorithm depends on calculation of the frequency of positive test responders as a function of CAD probability in the total diagnostic population (*10*). The obtained results are *independent* of cardiac catheterization. The approach is illustrated in Figure 4. The y-intercept of this plot at 0% CAD probability represents the false-positive rate (i.e., 1 − specificity) and the y-intercept at 100% CAD probability represents the true-positive rate (sensitivity). Alternatively, Begg and Greenes have developed an algorithm for "debiasing" biased catheterization data (*11*). Their algorithm depends on knowledge of the frequency of positive and negative test responses in both the catheterized population *and* the uncatheterized population. In the absence of bias, the frequency of positive test responders in the catheterized and uncatheterized populations would be identical. In the face of post-test referral bias, the frequency of positive test responders in the catheterized population exceeds that in the uncatheterized population. Diamond had performed a series of computer simulations which support the debiasing method proposed by Begg and Greenes (*12–13*). Clinical validation of these debiasing proposals, however, now need to be performed, so that their potential usefulness can be established.

Allman et al. used ROC curve analysis to evaluate the accuracy of thallium scintigraphy in their study (*1*). Along these lines, Diamond has suggested that such ROC curves are best analyzed by reporting the *area* under the ROC curve (*12*). This measurement has two potential advantages: (*1*) it provides a means of quantitatively comparing data from one study to another; and (*2*) using computer simulations, Diamond has suggested that the area under the ROC curve is resistant to post-test referral bias (*13*).

CLINICAL TRIALS FOR ASSESSING DIAGNOSTIC EFFICACY

Another potential approach to dealing with referral bias would be the development of "diagnostic" clinical trials, not dissimilar to some validation studies performed in the 1970s when radionuclide stress testing was still considered an "experimental" procedure. A random sample of angiographically diseased and normal patients would be selected for prospective radionuclide stress testing *after* the performance of coronary angiography. Here, the radionuclide stress test is no longer used to define the angiographic population to which the radionuclide stress test is to be compared. Still, this proposal is not so straightforward. Even without radionuclide stress testing as the criterion for referral, catheterized populations are biased. Thus, steps would still be required to insure a broad-spectrum of

catheterized patients. It would be interesting to compare the results obtained in such a trial versus those obtained by applying our probabilistic model (*10*), employed independent of angiography, and/or the Begg-Greenes debiasing algorithm (*11*), applied to conventional clinical data. There is sufficient merit for justifying such clinical trials to funding agencies.

FULL CHARACTERIZATION OF THE PATIENT POPULATION

As aforementioned, the indications for performing testing can vary widely from laboratory to laboratory. For the same indication, the subgroup of analyzed patients can also vary markedly from study to study. If investigators would fully characterize their overall patient population and the subgroup chosen for analysis, readers could better determine whether an investigative study refers to the patients typically seen in their own practice. Useful information would include a description of the indications for testing, the type and frequency of anginal symptoms, the hemodynamic and exercise ECG results, frequency of post-MI patients, medication usage, and the relative temporal sequence of performing noninvasive testing versus catheterization. Particular summary variables would be helpful. For instance, in diagnostic studies, the "normal" and diseased patient subgroups could be characterized by a single useful variable: the pre-test probability of coronary disease, using validated approaches, such as Bayesian analysis (*14*). If the pre-test probability of CAD in the angiographically normal population was very high, for example, the reader would know that a specific type of "pre-test referral bias" was operative in that particular study (a skew toward very sick normals). If pre-test CAD probability was very low among the angiographic normals, a different pre-test referral bias would be operative (preferential selection bias of "healthy" normals). Readers should be cautious about extrapolating these results to their laboratories if the pre-test probability in the investigative and reader's laboratories varied markedly.

TABULATION OF INDIVIDUAL PATIENT RESULTS

It would be useful if journal editors would permit investigators to present a tabulation of individual patient results as an appendix to their manuscripts. Today, journal articles are often markedly shorter than the size of articles published 20 to 30 yr ago. Previously, such tabulations were more common. Importantly, such tabulation would permit other, future investigators to build upon the cumulative experience of prior studies in a meta-analytic fashion. The usefulness of meta-analytic approaches has been re-enforced by a recent meta-analysis regarding the efficacy of clinical thrombolytic trials (*15*). In fact, many advances in our understanding of noninvasive stress testing, including evaluation of test algorithms or criteria (*16,17*), and Bayesian probability analysis (*14*), have come from the pooling of literature data. Unfortunately, in the current radionuclide stress literature, the presented data is most often too concise for meta-analytic purposes.

THE DEVELOPMENT OF A NATIONAL DATABASE REGISTRY

The development of some form of national data base registry would permit the pooling of large groups of relevant subsets of patients for meta-analysis and other purposes. Today, many laboratories already employ computer-generated reports which automatically furnish these laboratories with a computerized data base. If a group of these laboratories could agree on the encoding of selected variables, to be pooled by one center, the concept of a national data-base registry would no longer be so far-fetched.

ADOPTION OF STANDARDS FOR REPORTING TEST RESULTS

Just as standards have been adopted for the conduct of test protocols, it is now quite evident that the very process of testing is complex and that many biases may affect test results. It would be helpful if journal reviewers and editors were made to adopt a series of minimum guidelines for reviewing submitted publications. These would vary according to the type of study. For instance, in studies involving diagnostic tests, there would be at least minimal standards for characterizing the patient population, such as the following:

1. Post-MI patients must be automatically excluded for calculating sensitivity.
2. Volunteer and low CAD likelihood groups must be excluded for calculating specificity.
3. Investigators must specify how they handled "equivocal" test responses (this ubiquitous response is often ignored in reported studies).

Committees could be appointed through the relevant governing bodies to develop such guidelines.

CONCLUSIONS

The study by Allman et al. employs state-of-the art technology to address an important clinical question: given the high coronary flow rates induced by adenosine infusion and the fall off of myocardial thallium extraction at higher flow rates, might adenosine thallium scintigraphy underestimate the extent of coronary disease? The points raised in this editorial examine potential biases in their study without answering the validity of their conclusions. The results of this study may or may not prove robust to these biases. Further prospective studies are required.

This critique could be applied to virtually any radionuclide stress article published today. The current authors have simply followed existing standards, re-enforced by publisher decisions regarding manuscript length. While neither the focus of the current investigative report or this

editorial, it must be emphasized that radionuclide stress testing remains a test of physiology, not coronary anatomy. Radionuclide stress testing is best viewed as providing a "physiologic interpretation" of the coronary angiogram. By this approach, a nonischemic scintigram in the presence of a documented coronary stenosis represents a "false-negative" diagnostic response, but study after study has confirmed that these "false-negative responses" are not erroneous as prognostic predictors. A negative radionuclide stress test (under adequate stress) is a strong predictor of low cardiac risk among patients with suspected and confirmed coronary disease (*5,18–22*). This association has formed the basis for the application of radionuclide stress testing after coronary angiography, as a guide to patient management (*23*). Ultimately, it is the patient's prognosis—not the angiogram—to which radionuclide stress testing, angiography and clinical factors must all be compared.

Alan Rozanski
St. Luke's-Roosevelt Hospital Center
New York, New York

REFERENCES

1. Allman KC, Berry J, Sucharski LA, et al. Determination of extent and localization of coronary artery disease in patients without prior myocardial infarction by ^{201}Tl tomography in combination with pharmacologic stress. *J Nucl Med* 1992;33:2067–2073.
2. Rozanski A, Diamond G, Berman DS, Forrester JS, Morris D, Swan HJC. The declining specificity of exercise radionuclide ventriculography. *N Engl J Med* 1983;309:518–522.
3. Maddahi J, Rozanski A, Becerra A, et al. Patients with a calculative very low likelihood of coronary artery disease: an alternative population of cardiac normals. *Circulation* 1982;66:II-62.
4. Rozanski A, Diamond GA, Forrester JS, Berman D, Morris D, Swan HJC. Comparison of alternative referent standards for cardiac normality. Implications for diagnostic testing. *Ann Intern Med* 1984;101:164–171.
5. Jones RH, Floyd RD, Austin EH, et al. The role of radionuclide angiocardiography in the preoperative prediction of pain relief and prolonged survival following coronary artery bypass grafting. *Ann Surg* 1983;197:743–754.
6. Morris DD, Rozanski A, Berman DS, Diamond GA, Swan HJC. Noninvasive prediction of the angiographic extent of coronary artery disease following myocardial infarcttion: comparison of clinical, exercise electrocardiographic and ventriculographic parameters. *Circulation* 1984;70:192–201.
7. Hlatky MA, Prior DB, Harrell FE Jr, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariate analysis. *Am J Med* 1984;77:64–71.
8. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;299:926–30.
9. Diamond GA. An improbable criterion of normality. *Circulation* 1982;66:618.
10. Diamond GA, Rozanski A, Forrester JS, et al. A model for assessing the sensitivity and specificity of tests subject to selection bias. Application to exercise radionuclide ventriculography for diagnosis of coronary artery disease. *J Chron Dis* 1986;39:343–355.
11. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 1983;39:207–215.
12. Diamond GA. ROC steady: a receiver operating characteristic curve that is invariant relative to selection bias. *Med Decis Making* 1987;7:238–43.
13. Diamond GA. Affirmative actions: can the discriminant accuracy of a test be determined in the face of selection bias? *Med Decis Making* 1991.11:48–56.
14. Diamond GA, Forrester JS, Hirsch M, et al. Application of conditional probability analysis to the clinical diagnosis of coronary artery disease. *J Clin Invest* 1980;65:1210–1221.
15. Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N Engl J Med* 1992;327:248–254.
16. Rozanski A, Diamond GA, Jones R, et al. A format for integrating the interpretation of exercise ejection fraction and wall motion. With special reference to equivocal responses. *J Am Coll Cardiol* 1985;5:238–248.
17. Rozanski A, Diamond GA, Forrester JS, et al. Should the intent of testing influence its interpretation? *J Am Coll Cardiol* 1986;7:17–24.
18. Bonow RO, Kent KM, Rosing DR, et al. Exercise-induced ischemia in mildly symptomatic patients with coronary-artery disease and preserved left ventricular function. *N Engl J Med* 1985;312:389–394.
19. Pamelia FX, Gibson RS, Watson DD, et al. Prognosis with chest pain and normal thallium-201 exercise scintigrams. *Am J Cardiol* 1985;55:920–926.
20. Wahl JM, Hakki A, Iskandrian AS. Prognostic implications of normal exericse thallium-201 images. *Arch Intern Med* 1985;145:253–256.
21. Wackers FJT, Russo DJ, Russo D, et al. Prognostic significance of normal quantitative planar thallium-201 stress scintigraphy in patients with chest pain. *J Am Coll Cardiol* 1985;6:27–30.
22. Ladenheim ML, Pollack BH, Rozanski A, et al. Extent and severity of myocardial hypoperfusion as orthogonal indices of prognosis in patients with suspected coronary artery disease. *J Am Coll Cardiol* 1986;7:464–471.
23. Rozanski A, Berman DS. The efficacy of cardiovascular nuclear medicine studies. *Semin Nucl Med* 1987;27:104–120.