clinical characteristics, including risks for DVT and anatomic extent of disease, to be appreciated. Finally, future studies should include larger numbers of patients to provide narrower confidence intervals for the reported sensitivities and specificities.

Technetium-99m-RBC venography has potential but unproven clinical utility for the diagnosis of DVT. Other nuclear medicine techniques for the diagnosis of DVT are also promising. Tests based on anti-fibrin (*16*) or anti-platelet (*17*) monoclonal antibodies may provide accurate imaging for fresh thrombi in any location. However, there is insufficient data on the accuracy and utility of these new tests. While contrast venography remains the gold standard of diagnosis for DVT, and IPG and ultrasonography have clinical utility for proximal DVT, there may still be an important role for nuclear medicine imaging in the diagnosis of DVT in certain clinical settings.

ACKNOWLEDGMENTS

The authors thank C. David Teates, MD for reviewing the manuscript, and Ms. Mary Beth Meachum-Whitehill and Ms. Deanna Brown for typing the manuscript.

REFERENCES

1. Coon WW, Willis PW, Keller JB. Venous thromboembolism and other venous disease in the Tecumseh Community Health Study. *Circulation* 1973;48:839–846.
2. Becker DM, Philbrick JT, Abbitt PL. Real-time ultrasonography for the diagnosis of lower extremity deep venous thrombosis: the wave of the future? *Arch Intern Med* 1989;149:1731–1734.
3. Lisbona R. Radionuclide blood-pool imaging in the diagnosis of deep-vein thrombosis of the leg. In: Freeman LM, Weissman HS, eds. *Nuclear medicine annual 1986.* New York: Raven Press; 1986:161–193.
4. Hayt DB, Binkert BL. An overview of noninvasive methods of deep vein thrombosis detection. *Clin Imaging* 1990;14:179–197.
5. Beswick W, Chmiel R, Booth R, Vellar I, Gilford E, Chesterman CN. Detection of deep venous thrombosis by scanning of ^{99m}Tc-labeled red-cell venous pool. *Br Med J* 1979;1:82–84.
6. Kempi V, van der Linden W. Diagnosis of deep vein thrombosis with in vivo ^{99m}Tc-labeled red blood cells. *Eur J Nucl Med* 1981;6:5–9.
7. Lisbona R, Stern J, Derbekyan V. ^{99m}Tc-red blood cell venography in deep vein thrombosis of the leg: a correlation with contrast venography. *Radiology* 1982;143:771–773.
8. Fogh J, Nielsen SL, Vitting K, et al. The diagnostic value of angioscintigraphy with ^{99m}Tc-labeled red blood cells for detection of deep vein thrombosis. *Nucl Med Commun* 1982;3:172–181.
9. Singer I, Royal HD, Uren RF, et al. Radionuclide plethysmography and Tc-99m-red blood cell venography in venous thrombosis: comparison with contrast venography. *Radiology* 1984;150:213–217.
10. Littlejohn GO, Brand CA, Ada A, Wong C. Popliteal cysts and deep venous thrombosis: Tc-99m red blood cell venography. *Radiology* 1985;155:237–240.
11. Philbrick JT, Becker DM. Calf deep venous thrombosis: a wolf in sheep's clothing? *Arch Intern Med* 1988;148:2131–2138.
12. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;299:926–930.
13. Lisbona R, Leger J, Stern J, Derbekyan V, Skinner B. Observations on Tc-99m-erythrocyte venography in normal subjects and in patients with deep vein thrombosis. *Clin Nucl Med* 1981;6:305–309.
14. Zorba J, Schier D, Posmituck G. Clinical value of blood pool radionuclide venography. *AJR* 1986;146:1051–1055.
15. Leclerc JR, Wolfson C, Arzoumanian A, Blake GP, Rosenthall L. Technetium-99m-red blood cell venography in patients with clinically suspected deep vein thrombosis: a prospective study. *J Nucl Med* 1988;29:1498–1506.
16. Alavi A, Palevsky HI, Gupta N, et al. Radiolabeled antifibrin antibody in the detection of venous thrombosis: preliminary results. *Radiology* 1990;175:79–85.
17. Peters AM, Lavender JP, Needham SG, et al. Imaging thrombus with radiolabeled monoclonal antibody to platelets. *Br Med J* 1986;293:1525–1527.

EDITORIAL

Diagnostic Accuracy and Deep Venous Thrombosis: A Biostatistician's Perspective

In this issue of the *Journal of Nuclear Medicine*, Pinson, Becker, Philbrick and Parekh (*1*) make another contribution to the already extensive literature concerning noninvasive alternatives to the use of contrast venography (CV) in the diagnosis of deep venous thrombosis (DVT). The direction of the literature is clear—noninvasive diagnostic methods continue to chip away slowly at the position of CV as the gold standard in DVT detection and characterization. However, there seems to be some disagreement over the stability of CV's standing. In 1988, Redman (*2*) concluded an editorial in *Radiology* by commenting:

> Clearly, CU (compression ultrasound) for diagnosis of acute DVT, either alone or in conjunction with Doppler or impedance plethysmography, meets the criteria for a screening test. CV can retain the title of "gold standard" while each radiologist traverses the learning curve for CU, but then CV should be positioned as a backup procedure for the times when results of less invasive procedures raise more questions than answers.

A year later, an anonymous editorial in the *Lancet* (*3*) offered a different viewpoint, saying:

> Efficient treatment of venous thrombosis demands accurate knowledge of the extent and appearance of the thrombus and, in particular, the limit of its proximal extension; this information may not be satisfactorily obtained with non-invasive investigations alone.

From my perspective as a biostatistician, the most significant contribu-

Received Aug. 20, 1991; accepted Aug. 27, 1991.

For reprints contact: Stephen J. Walsh, ScD, Department of Community Medicine and Health Care, Room AG-060, University of Connecticut Health Center, Farmington, CT 06030.

tion of the paper by Pinson and colleagues towards settling the debate over DVT diagnosis is neither its format as a critical review nor its focus on a radionuclide imaging technique that has, perhaps, not received as much consideration as it might merit. Rather, within the context of the DVT controversy, I consider the distinctive aspect of their paper to be its clear emphasis on the need for adherence to basic methodologic standards in the evaluation of competing diagnostic systems. This message is by no means new or obscure. It was initially voiced in 1978 in a seminal article by Ransohoff and Feinstein (*4*) that identified common problems with patient spectrum and bias in studies of diagnostic tests. The standards used in the Pinson paper are derived from that article by way of earlier critical reviews published by Philbrick et al. (*5*) and by Becker et al. (*6*). The Pinson paper provides evidence that the debate over DVT diagnosis is based, in part, on investigations that lack the level of scientific rigor encouraged by Ransohoff and Feinstein. More evidence to this effect is provided by the Becker article which applied similar standards in evaluating fifteen studies of real-time ultrasonography in DVT detection.

Resolution of the tension between CV and ultrasonography, plethysmography or scintigraphy cannot occur until studies of clearly recognizable validity provide consistent results. In the interim, the prevalence of DVT and the importance of its treatment will continue to promote extensive research. While the methodologic standards put forth by Pinson et al. provide a minimal basis for achieving validity, I would argue that, as expressed, they represent merely a checklist of guidelines. They lack a unifying focus and, by doing so, are prone to be applied too rigidly and without a full appreciation of the practical trade-offs that must be made in systematically evaluating diagnostic tests. Furthermore, the standards fail to reflect much of the development that has taken place over the last fifteen years concerning what needs to be measured in diagnostic trials, how those measurements should be obtained and, ultimately, how they should be combined across studies in order to reach higher levels of conclusiveness than can be achieved by most studies alone. The purpose of this editorial is to highlight major aspects of more recent methodologic developments and to offer a broad framework that characterizes the role they should play in diagnostic research.

Measurement lies at the heart of the scientific method and measuring "accuracy" is the ideal focus of an evaluation of diagnostic systems. The studies reviewed by Pinson all reflect an appreciation of the fact that accuracy cannot be adequately captured by a single measurement but requires, at very least, obtaining a pair of measurements—namely, sensitivity and specificity. However, the studies do not acknowledge the fact that even a pair of sensitivity and specificity estimates cannot fully represent the range of performance possible with any diagnostic system that relies on observer judgement for the determination of disease positivity. Reliance of a diagnostic test on observer interpretation means that the outcome from a single application of that test may be differentially classified by individual readers if they are using different criteria in reaching their judgements or if they have varying levels of skill in using the procedure. Even when the levels of skill are balanced, the potential operation of individual standards requires that accuracy be viewed in a broader sense as the *functional relationship* that exists between sensitivity and specificity as the threshold of disease positivity is varied from less to more lenient. This functional relationship is graphically represented by the receiver operating characteristic (ROC) curve. Table 1 in the Pinson article lists sensitivity and specificity pairs from the six studies under review. On the one hand, the various levels of observed "accuracy" may indicate quantitatively different diagnostic capabilities resulting either from how ^{99m}Tc venography was applied across the studies or from disparate reader skills. Alternatively, the variation may only be the by-product of different thresholds for pronouncing disease presence on the part of the participating radiologists. Without reference to ROC methodology, we simply cannot tell.

The ROC curve was first introduced to the medical setting in 1960 (*7*). Since then, efforts have been made to incorporate the concept into the wider field of medical decision making, to develop experimental designs appropriate to evaluating specific diagnostic tasks and to provide statistical methods for analyzing data from diagnostic trials. Two excellent reviews by Hanley (*8*) and Metz (*9*) include virtually complete bibliographies of the topic and can serve as concise introductions to the application of ROC methodology in diagnostic research.

The complement of knowing what to measure is knowing how to measure it. The question of how to collect data regarding diagnostic performance is really one of experimental design. One can claim that the standards put forth in the Pinson article should be applied in designing any study of diagnostic methods. However, I prefer to view them only as components that potentially contribute to satisfaction of the broader objectives of experimental design—broader objectives that include repeatability, validity and generalizability.

Repeatability of any study of a diagnostic test requires that the technique be formally standardized before the study begins and fully documented in any release of study results. This is, in essence, Standard 1 of Pinson et al. Yet, merely establishing a standard protocol may not be sufficient to ensure repeatability. In particular, a "quality control" process may need to be instituted, particularly in large or prolonged studies, to make sure that personnel who apply the technique do so uniformly and that no evolution in how the standard procedure is performed occurs over time.

A series of multi-institutional trials of imaging techniques for cancer staging have sought to maintain quality by establishing committees of co-investigators whose primary responsibility is the monitoring of protocol compliance throughout the period of data collection (*10*). The repetition of study results may also be difficult if non-standardized definitions of primary disease characteristics are employed by the individuals who interpret test results. Pinson's second standard alludes to this concern. One procedure for reducing this source of interobserver variability is to establish reference sets of images from selected cases before the study starts and to use those sets to reinforce in the readers a common sense of how to classify particular findings (*10*).

The goal of achieving validity in diagnostic trials can be expressed negatively as the intention to avoid study procedures that systematically yield inaccurate or biased measurements. Standards 6 and 7 of Pinson et al. warn against using study designs that allow test results to effect the application of the gold standard or that permit contamination of test results by the gold standard or vice versa. Begg has reviewed these and other biases that can occur in test evaluation studies (*11*). Among the other biases, two worth emphasizing include the difficulty of genuinely assessing accuracy in the absence of a definitive gold standard and the possibility that extraneous effects may be introduced when patients with uninterpretable or incomplete test results are eliminated from the study population.

Definition and description of the study population in a diagnostic trial are primary requirements for generalizability. These are the issues addressed by Standards 3, 4, and 5 in the Pinson article. Originally, these were the elements that, for Ransohoff and Feinstein, constituted patient "spectrum" or, in other terminology, "case mix." Patient spectrum effects both the levels of diagnostic performance observed as well as the applicability of study results to more routine clinical practice. However, the designers of diagnostic studies also need to be sensitive to the presence of a second population, specifically, the population of test readers (*9*). One would expect different levels of performance from academic radiologists as opposed to residents. The generalizability of study findings may well depend on the levels of expertise and experience possessed by the readers chosen to participate in the study.

I believe that any discussion of design issues for diagnostic trials needs to acknowledge the realities of clinical practice and the fact that compromises of the "perfect" design are inevitable. For example, in studying DVT, contrast venography is used as the reference test by which to establish the "true" diagnosis. Yet, CV is itself only another imaging method. It has been shown that variation exists among readers of CV, thereby implying that the "gold standard" is less than perfect (*12*). One way to avoid this issue is to analyze agreement between CV and its competitors instead of treating CV as a definition of truth. One commonly used measure of agreement between diagnostic tests is the kappa statistic (*13*). Agreement is a relative assessment of performance rather than the "absolute" or "pure" assessment of accuracy that we seek in the ideal. Recourse to the assessment of relative performance has also been suggested by Metz as a way of potentially tolerating biases when alternative diagnostic procedures are being compared, but only as long as there is some assurance that the biases are "balanced" across procedures (*14*).

The norm in research investigating diagnostic alternatives for DVT appears to be small, single institution studies. The median sample size for the six studies reviewed by Pinson et al. was 32. In the review of 15 studies of ultrasonography by Becker et al. the median sample size was 46 patients. Such small studies face difficulties achieving adequate precision in the estimation of even a single pair of sensitivity/specificity values. The possibility of valid sub-group analysis, e.g. proximal versus distal location, or of reliably estimating ROC curves is remote. Two options that potentially address this predicament deserve consideration from those actively involved in DVT diagnostic research. On the one hand, they should consider the possibility of organizing multi-institutional trials. This approach has recently been taken in studying imaging for cancer staging and has met with some success (*15*, *16*). A discussion of the methodologic questions faced by the cooperative group that organized these trials has been offered by Gatsonis and McNeil (*10*). Factors that favor the potential success of this approach for the DVT problem include the wide prevalence of appropriate study subjects and the ready availability of necessary equipment and trained personnel. On the other hand, it is clear that independent, small studies will continue to be performed and offered for publication. Recent years have seen the emergence of *meta-analysis* as a method of formally combining quantitative results of individual studies on the same topic in the expectation that aggregation will provide a greater level of precision or conclusiveness. A comprehensive review of the basic issues in its application has been published by Ellenberg (*17*). The specific adaptation of meta-analytic techniques to the field of diagnostic test evaluation is in its infancy. An initial contribution to this effort has been made by Kardoun and Kardoun (*18*), who have developed a method for combining the individual sensitivity/specificity assessments that often result from diagnostic trials to yield a common ROC curve for the particular diagnostic technique under study.

Even further developments in the application of meta-analysis for diagnostic test evaluation will not be able to overcome wide-spread neglect of basic methodologic issues by individual investigators. A critical review such as that presented by Pinson and colleagues reminds researchers of the need for sensitivity to these issues.

Unfortunately, there is no fixed experimental design or succinct list of guidelines that can be routinely applied to the study of DVT diagnostic systems with the assurance of valid and conclusive results. The complexities of the disease and of the available diagnostic technology will not permit such a simple approach. However, there is an extensive literature detailing methodologic developments specifically for diagnostic trials. My hope is that this editorial has identified the major themes in that literature and has provided basic bibliographic references to them.

ACKNOWLEDGMENTS

The author would like to thank Holger Hansen, MD, Dr. P.H., and Scott Wetstone, MD, for reviewing the manuscript.

Stephen J. Walsh
University of Connecticut Health Center
Farmington, Connecticut

REFERENCES

1. Pinson AG, Becker DM, Philbrick JT, Parekh JS. Technetium-99m-RBC venography in the diagnosis of deep venous thrombosis of the lower extremity: a systematic review of the literature. *J Nucl Med* 1991;32:2324–2328.
2. Redman HC. Deep venous thrombosis: is contrast venography still the diagnostic "gold standard"? *Radiology* 1988;168:277–278.
3. Anonymous. Diagnosis of deep-vein thrombosis. *Lancet* 1989;2:23–24.
4. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;299:926–930.
5. Philbrick JT, Horwitz RI, Feinstein AR. Methodologic problems of exercise testing for coronary artery diaease: groups, analysis and bias. *Am J Cardiol* 1980;46:807–812.
6. Becker DM, Philbrick JT, Abbitt PL. Real-time ultrasonography for the diagnosis of lower extremity deep venous thrombosis. *Arch Intern Med* 1989;149:1731–1734.
7. Lusted LB. Logical analysis in roentgen diagnosis. *Radiology* 1960;74:178–193.
8. Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. *CRC Crit Rev Diagn Imaging* 1989;29:307–335.
9. Metz CE. ROC methodology in radiologic imaging. *Invest Radiol* 1986;21:720–733.
10. Gatsonis C, McNeil BJ. Collaborative evaluations of diagnostic tests: experience of the Radiology Diagnostic Oncology Group. *Radiology* 1990;175:571–575.
11. Begg CB. Biases in the assessment of diagnostic tests. *Stat Med* 1987;6:411–423.
12. McLachlan MS, Thomson JG, Taylor DW, Kelly ME, Sackett DL. Observer variation in the interpretation of lower limb venograms. *AJR* 1979;132:227–229.
13. Fleiss JL. *Statistical methods for rates and proportions*, 2nd edition. New York: Wiley; 1981.
14. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest Radiol* 1989;24:234–245.
15. Rifkin MD, Zerhouni EA, Gatsonis CA, et al. Comparison of magnetic resonance imaging and ultrasonography in staging early prostate cancer. *N Engl J Med* 1990;323:621–626.
16. Webb WR, Gatsonis C, Zerhouni EA, et al. CT and MR imaging in staging non-small cell bronchogenic carcinoma: report of the Radiologic Diagnostic Oncology Group. *Radiology* 1991;178:705–713.
17. Ellenberg SS. Meta-analysis: the quantitative approach to research review. *Semin Oncol* 1988;15:472–481.
18. Kardoun JW, Kardoun OJ. Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation. *Meth Inform Med* 1990;29:12–22.