Diagnostic Test Comparisons in Patients with Deep Venous Thrombosis

Bruce R. Line, Teresa L. Peters and James Keenan
Department of Radiology, Albany Medical Center, Albany, New York

New diagnostic modalities are often judged relative to accepted standard procedures. These comparisons are influenced by the accuracy of the standard test and the prevalence of disease in the study population. We evaluated the importance of these factors in the assessment of antifibrin scintigraphy when used to detect deep venous thrombosis. **Methods:** Scintigraphy is compared to contrast venography in two populations of patients with different disease prevalence. We calculate the sensitivity and specificity by limb site (calf, knee, thigh) and the overall diagnosis for each modality. The sensitivity and specificity results obtained using venography as a gold standard are compared to those obtained using a maximum likelihood statistical procedure that does not require comparison to a standard test. **Results:** A significant variation in the apparent sensitivity, specificity and accuracy is found for antifibrin scintigraphy as related to limb site, disease prevalence and use of a gold standard. The value of antifibrin scintigraphy sensitivity (84.7%) and specificity (75.8%) predicted by the maximum likelihood analysis are substantially higher than those obtained from the estimates based on the use of venography as a gold standard for both high and low disease prevalence populations. The sensitivities and specificities of antifibrin scintigraphy (84.7% and 75.8%, respectively) and venography (71.7% and 80.7%, respectively) are comparable for the combined study group of 268 patients. **Conclusion:** To obtain unbiased evaluations of a new diagnostic modality, it is essential to take into account the errors of the standard reference test and disease prevalence in the study population. The results of our analysis suggest that it may not be appropriate to use contrast venography as a gold standard in the assessment of new diagnostic imaging procedures for DVT.

Key Words: antifibrin scintigraphy; deep venous thrombosis; diagnostic test comparisons

J Nucl Med 1997; 38:89–92

Thromboemboli either causes or contributes to 200,000 deaths a year in the U.S. (*1*). The vast majority of thromboemboli arise in the deep iliofemoral veins (*2*). Unfortunately, the clinical diagnosis of these disorders remains problematic and difficult. Clinical evidence of thrombosis occurs in only half of the patients with confirmed deep vein thrombosis, whereas only a third of patients with symptoms compatible with deep venous thrombosis (DVT) actually have the disease. Common diagnostic tests for DVT include contrast venography, Doppler ultrasound and impedance plethysmography. Each of these modalities has unique advantages and disadvantages; none has perfect accuracy. In the diagnosis of DVT, contrast venography is considered the most accurate means of detecting extremity thrombotic disease. It is the gold standard used to assess new procedures.

When a new test is developed, its accuracy must be determined by comparing its diagnostic predictions to the true condition of the patient. This comparison is usually reported as the sensitivity and specificity of the test. Sensitivity is defined as the probability of a positive test given the presence of disease, whereas specificity is the probability of a negative test given the absence of disease. Unfortunately, the presence or absence of disease is difficult to determine. Indeed, this uncertainty may not be resolved even with a battery of diagnostic procedures and with prolonged follow-up.

When a standard test is used to provide reference information about the disease state, estimates of the sensitivity and specificity for an investigational procedure are biased downward by the error in the standard test (*3,4*). For instance, when the false-positive rate of a standard test is incorrectly assumed to be zero, the false-negative rate (1-sensitivity) of the new test is overestimated. Likewise, the new test false-positive rate (1-specificity) is overestimated when the standard test is incorrectly assumed to have no false-negatives. Thus, such bias limits the accuracy of estimates of the sensitivity and specificity of a new test.

Using venography as a gold standard, i.e., one assumed to have no error, we have estimated the sensitivity and specificity of antifibrin scintigraphy in several populations of patients suspected of having DVT. The estimates are found to vary by limb site within a given population and vary between populations for each limb site. These results are contrasted with estimates of sensitivity and specificity produced by a maximum likelihood procedure that does not require comparison to an "error free" standard test. To obtain unbiased evaluations, it is essential to consider the errors of the standard test and disease prevalence. The results of our analysis suggest that contrast venography should not be assumed to be a gold standard in the assessment of new diagnostic imaging procedures for DVT.

MATERIALS AND METHODS

Patients

A large multicenter trial was performed in the United States and Europe to prospectively evaluate the diagnostic performance of ^{99m}Tc-antifibrin (^{99m}Tc-T2G1s Fab′) in patients with suspected acute DVT (*5*). In the Phase 1 and 2 portions of the trial, over 400 patients received 0.5 mg T2G1s labeled with 15–20 mCi ^{99m}Tc. Images were acquired immediately after injection to establish the blood pool distribution of the antibody and then again at 90 min and 4–6 hr postinjection. Contrast venography was performed within 24–36 hr of the antifibrin image acquisition.

Two patient populations from the Phase 2 trials were evaluated in this study. Multicenter trial inclusion and exclusion criteria provided study populations with clinically different disease prevalence. The high prevalence group included 145 patients who had clinical signs and symptoms of DVT. The low prevalence group was comprised of 123 patients at risk for developing DVT by virtue of having undergone orthopedic, abdominal, retroperitoneal, neurologic, gynecologic or urologic surgery.

Study Interpretation and Comparisons

All venogram data on these patients were interpreted by two consultant radiologists whose readings were accepted when concordant and otherwise were adjudicated by a third independent

Received Jan. 5, 1996; revision accepted Apr. 30, 1996.
For correspondence or reprints contact: Bruce R. Line, MD, Professor of Radiology, Nuclear Medicine, A-72, Albany Medical Center, Albany, NY 12208.

TABLE 1
2 × 2 Table for Comparison between Standard and Trial Test

	Positive standard Test (A+)	Negative standard Test (A−)	Sums
Positive trial Test (B+)	a	b	a + b
Negative trial Test (B−)	c	d	c + d
	a + c	b + d	n

reading. Venograms were interpreted by calf, knee or thigh limb site and were read as positive, negative or indeterminate for the presence of thrombus. Antifibrin studies were interpreted as either positive or negative by consensus of two nuclear medicine physicians. Antifibrin study interpretations of the anterior thigh, posterior knee and posterior calf were used for the purposes of comparison with the venography interpretations. Positive antifibrin uptake at a thrombus site was defined as a localization in the deep venous system where activity increased with time relative to a control site in contralateral or adjacent vascular regions. Limb site readings were included in the analysis when both venography and scintigraphy were available at that site and were either positive or negative for the presence of clot. Other venogram site readings such as indeterminate and suboptimal were excluded from analysis. The limb site was classified as positive if either of the two limbs was read as positive for thrombus. Similarly, the total leg interpretation was classified as positive if any site in the calf, knee or thigh was read as positive for thrombus.

Statistical Methods

The comparisons of antifibrin interpretations (trial test) with the venogram interpretations (standard test) were performed using the 2 × 2 arrangement shown in Table 1. After the notation of Gart and Buck (*3*) A, B and D are used to refer to the standard test, the trial test and disease state, respectively. For example, A+ indicates venogram positive, B− means antifibrin-negative and D− refers to disease-negative. The sensitivity of the venogram gold standard is denoted Sp = P(A+ D+) or the probability of a positive venogram given the presence of disease. The specificity of the standard test is indicated by Sn = P(A− D−). Similarly, Sp′ = P(B+ D+) and Sn′ = P(B− D−), respectively, refer to the sensitivity and specificity of the trial test. When the disease state (D+, D−) is not known directly, the standard test result (A+, A−) is often used to classify the individual. Under these circumstances, the terms co-positivity and co-negativity are used, respectively, in place of sensitivity and specificity. (*3*) Co-positivity is defined as Cp = P(B+ A+) and co-negativity is Cn = P(B− A−). Co-positivity and co-negativity are equal to the sensitivity and specificity of the antifibrin test when the venogram is an accurate reflection of the disease state of the patient, i.e., where P(B+ A+) ≅ P(B+ D+) and P(B− A−) ≅ P(B− D−).

The distribution of the scan and venogram test interpretations are shown in the left-hand columns of Table 2. Test results are presented according to the 2 × 2 table cell counts (Table 1, cells a = d). Co-positivity and co-negativity values are computed for both trial populations, for each limb site and for the leg as a whole using the definitions Cp = a/(a + c) and Cn = d/(b + d).

Hence, co-positivity and co-negativity can be expressed in terms of the disease prevalence and the sensitivity and specificity of both tests (*3*). Where n is the population size, and Pr = P(D+) is used to denote the prevalence of disease, the expected values (E) or probabilities for the four cells in the 2 × 2 table (Table 1, cells a-d) are:

$$E(a)/n = P(A+B+) = S_pS_p'Pr + (1 - S_n)(1 - S_n')(1 - Pr)$$

$$E(b)/n = P(A-B+) = (1 - S_p)S_p'Pr + S_n(1 - S_n')(1 - Pr)$$

$$E(c)/n = P(A+B-) = S_p(1 - S_p')Pr + (1 - S_n)S_n'(1 - Pr)$$

$$E(d)/n = P(A-B-) = (1 - S_p)(1 - S_p')Pr + S_nS_n'(1 - Pr)$$

and

$$C_p = \frac{P(A+B+)}{P(A+B+) + P(A+B-)} = \frac{(1 - S_n)(1 - S_n') + Pr[S_pS_p' - (1 - S_n)(1 - S_n')]}{(1 - S_n) + Pr(S_p + S_n - 1)} \quad \text{Eq. 1}$$

$$C_n = \frac{P(A-B-)}{P(A-B-) + P(A-B+)} = \frac{S_nS_n' + Pr[(1 - S_p)(1 - S_p') - S_nS_n']}{S_n - Pr(S_p + S_n - 1)}. \quad \text{Eq. 2}$$

The bias introduced into estimates of trial test sensitivity and specificity by the error in the standard test may be characterized by an adaptation of Youden's index for rating diagnostic tests (*8*). This bias measure is denoted by j = Cp + Cn − 1, where $-1 \leq j \leq +1$, and j equals zero whenever there is no introduction of bias (*3*). Equations 1 and 2 indicate that when Pr = 0, then Cp = 1 − Sn′, Cn = Sn′ and j = 0. This implies that at zero prevalence there will be no apparent association between the tests. Similarly, when Pr = 1 then Cp = Sp′, C_n = 1 − Sp′ and j = 0, again denoting no bias. Gart and Buck empirically verified that if Sp, Sp′, Sn, Sn′ ≥ ½, then Cp continuously increases as a function of Pr varying from a minimum of 1 − Sn′ at Pr = 0 to a maximum of Sp′ at Pr = 1. On the other hand, Cn is a monotonic decreasing function of Pr varying

TABLE 2
Co-positivity and Co-negativity Estimates for Scintigraphy Using Venography as the Standard Test for Trials with Low- and High-Disease Prevalence

Low-prevalence population	(a) A+ B+	(c) A+ B−	(b) A− B+	(d) A− B−	No.	Co-positivity	Co-negativity
Calf	6	14	12	80	112	30.0%	87.0%
Knee	1	3	4	97	105	25.0%	96.0%
Thigh	1	4	16	99	120	20.0%	86.1%
Leg	6	18	24	75	123	25.0%	75.8%
High-prevalence population							
	(a) A+ B+	(c) A+ B−	(b) A− B+	(d) A− B−	No.	Co-positivity	Co-negativity
Calf	33	13	28	49	123	71.7%	63.6%
Knee	17	27	22	73	139	38.6%	76.8%
Thigh	22	17	12	84	135	56.4%	87.5%
Leg	41	19	31	54	145	68.3%	63.5%

TABLE 3
Sensitivity, Specificity and Prevalence Parameter Estimates for Venography and Scintigraphy

	Sensitivity (%)		Specificity (%)		Population prevalence (%)	
Site	Venogram	Scan	Venogram	Scan	Low	High
Calf	58.3%	85.4%	85.7%	90.1%	8.2%	52.6%
Knee	45.0%	39.2%	98.4%	97.1%	5.2%	69.3%
Thigh	100.0%	60.9%	96.4%	86.1%	0.5%	23.9%
Leg	71.7%	84.7%	80.7%	75.8%	0.4%	42.0%

from a maximum of Sn′ at Pr = 0 to a minimum of 1 − Sp′ at Pr = 1 (*3*).

The maximum likelihood method described by Hui and Walter is used to estimate the disease prevalence as well as the sensitivity and specificity for both the standard and trial tests (*6*). The maximum likelihood analysis was applied as a macro function that utilizes the PROC MATRIX and other functions in the statistical analysis system (SAS) (*7*). The procedure depends on conditional independence of the tests and uses the outcome of both tests in each of two patient populations with different disease prevalence. As part of the trial design, the scans and venograms were read independently. Furthermore, given the separate physical basis for scintigraphy and venography, it is reasonable to assume conditional independence of antifibrin scintigraphy and contrast venography, i.e., the outcome of one of the tests does not in itself predispose the outcome of the other test. This assumption does not imply that the tests may not agree frequently. If both tests are good they will agree often, whereas if either or both are poor they may only agree occasionally (*3*).

RESULTS

Co-positivity (sensitivity) and co-negativity (specificity) were calculated for antifibrin scintigraphy using contrast venography as the standard test for the high and low prevalence populations (Table 2). For most limb sites, the data show higher co-positivity in the high compared with the low prevalence group and higher relative co-negativity in the low prevalence group. If venography is accepted as the gold standard, the data suggest a large variation in scintigraphic sensitivity and specificity between trial groups.

It is clear that the values for scan sensitivity and specificity calculated without reference to a standard (Table 3) are substantially higher than those obtained from the estimates based on the use of venography as a gold standard (Table 2). Table 3 contains the maximum likelihood parameter estimates of population disease prevalence and the sensitivity and specificity for both antifibrin scintigraphy and contrast venography. The results represent the best estimate of sensitivity and specificity for the two tests across both populations. The data suggest that antifibrin scintigraphy may be a better test than contrast venography at some sites (i.e., in the calf). For this series of patients, the whole leg data suggest that antifibrin and contrast venography have comparable accuracy for the detection of DVT.

The relationship of disease prevalence to co-positivity and co-negativity values for each of the limb sites and overall limb are shown in Figure 1. These curves are generated from Equations 1 and 2 using sensitivity and specificity estimates derived from the maximum likelihood procedure. The closed circles represent the location of the co-positivity and co-negativity values computed using the 2 × 2 table data for the high prevalence study, where the prevalence is predicted by the maximum likelihood procedure. The open circles are the corresponding values for the low prevalence trial.

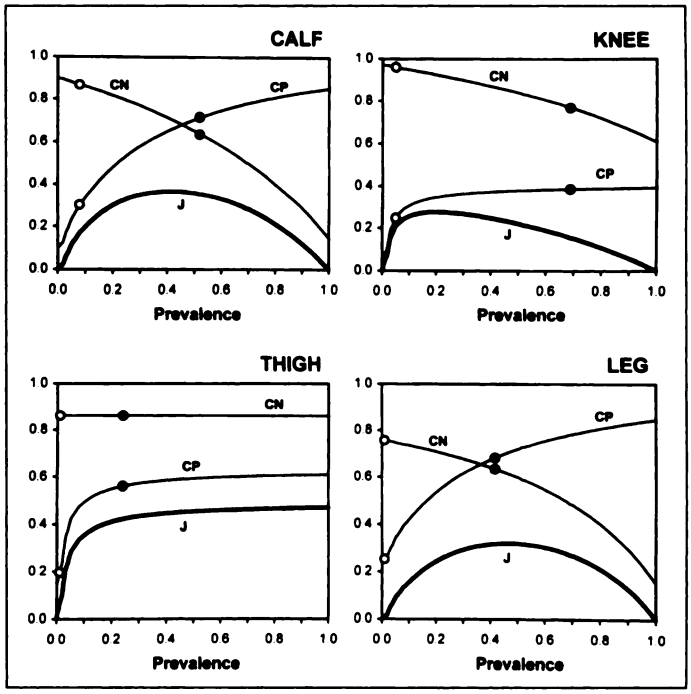

FIGURE 1. Co-positivity (C_p), co-negativity (C_n) and co-interaction (J) curves as a function of disease prevalence for the calf, knee, thigh and leg. Curves were generated from Equations 1, 2 using maximum likelihood estimates of sensitivity and specificity for venography and scintigraphy studies. Co-positivity and co-negativity values are shown for the high (closed circles) and low (open circles) prevalence population. In the thigh, the value of J at Pr = 1 is undefined because contrast venography sensitivity is estimated at 100%.

DISCUSSION

Vascular thrombosis is a life-threatening disease that has an estimated incidence of 2.5 million cases per year in the United States (*9*). Nearly 20% of all hospitalized patients develop DVT or pulmonary embolization (PE). Contrast venography is considered the most accurate of the tests for deep venous thrombotic disease and can evaluate the entire lower extremity. Unfortunately, contrast venography is an invasive procedure that requires cannulation of the dorsal foot veins. With iodinated contrast agents there is also a low but definite risk of anaphylactoid reaction and a risk of nephrotoxicity that is greatest in patients with borderline renal function, diabetes and cardiac failure.

In search of a highly specific and sensitive imaging technology, several groups have studied the use of radiolabeled monoclonal antibodies that bind to either fibrin or to platelet deposits (*10*). The greatest clinical experience to date has been achieved with T2G1s antifibrin, an antibody that is directed against an epitope expressed only on fibrin in newly formed thrombi (*11*). Antifibrin scintigraphy is technically uncomplicated and can assess the entire lower extremity including the calf and pelvic veins. Nephrotoxicity and anaphalactoid reactions have not been associated with its use.

When a new diagnostic test is developed, its error rates must be determined and weighed against its cost. Error rates can be estimated directly if the test can be applied to some individuals whose true disease states are known, but this is usually difficult or not feasible. In such cases, the new test is customarily evaluated against a standard test by applying both studies to a population at risk for the disease. For the purpose of this comparison, it is often assumed that the standard test accurately reflects the disease state of the patient. The standard test is thus

treated as if it were a gold standard, i.e., one with 100% sensitivity and specificity.

The problems associated with an imperfect gold standard are well-recognized (*4,12,13*). Greenberg and Jekel (*4*) found that the error rates of the standard test, if not taken into account, can lead to biased estimates of the error rates of the new test. Specifically, if the nonzero false-positive (or negative) rate of the standard test is assumed to be zero then the false-negative (or positive) rate of the new test will be overestimated. Therefore, the frequently made assumption of zero error rates for contrast venography may lead to high estimates of the error rates for antifibrin scintigraphy and then to unjustified pessimism about its clinical utility. For example, Table 2 shows reduced sensitivity and specificity for antifibrin when compared against a venography gold standard in both the high and low prevalence populations.

The bias in the error estimates for antifibrin scintigraphy can be avoided if an exact and independent assessment of the error rates of contrast venography is available (*4*). Alternatively, where venography error rates cannot be determined exactly, then it is possible to obtain accurate estimates of the false-positive rate of scintigraphy in subpopulations with low disease prevalence, and conversely, the false-negative rate can be obtained in subpopulations with high DVT prevalence. For these two circumstances, the bias introduced by the standard test is small (*4*).

Gart and Buck pointed out that the indices of agreement between tests may vary greatly in different populations. They described a probabilistic model that suggests that this variation in test co-positivity and co-negativity may be due to the difference in disease prevalence among the populations being compared (*3*). The model also indicates that the degree of bias introduced by error in the standard test is a minimum at the extremes of the prevalence range. The marked variability of scan co-positivity and co-negativity as a function of disease prevalence is shown in Figure 1. The maximum likelihood estimates (Table 3) can be used with Equations 1 and 2 to predict the sensitivity (co-positivity) and specificity (co-negativity) that would be found for antifibrin when using contrast venography as a gold standard (Table 2). As is evident from Equations 1 and 2, an unbiased estimate of the sensitivity of the trial test is expressed by the co-positivity ($Cp = Sp'$) where $Pr = 1$, and specificity is given by the co-negativity ($Cn = Sn'$) at $Pr = 0$. The extent to which the gold standard biases the estimates of the trial test sensitivity and specificity is reflected by $j = Cp + Cn - 1$, an analog of Youden's index (*3,8*). Figure 1 shows j as a function of disease prevalence. With the exception of the thigh, the bias introduced is zero at $Pr = 0$ and 1 and positive elsewhere.

Hui and Walter (*6*) proposed a maximum likelihood method to estimate the sensitivity and specificity of a diagnostic test without reference to a gold standard. The method utilizes maximum likelihood estimation and iterative convergence to determine all error rates and variances of these estimates (*7*). It uses diagnostic test results from two populations with different disease prevalence and estimates the error rates of both tests and the prevalence of both populations. In addition to requiring different prevalence rates, this methodology assumes the error rates of the two diagnostic tests are independent, i.e., the bases of the tests are not related (as in the case of a scan and a radiograph). The iterative solution to the maximum likelihood estimate is best obtained when the prevalence of the two populations are widely dissimilar (*7*).

The presence of a disease in an individual often cannot be determined with certainty. Such is the case with DVT, a condition where the cost of misdiagnosis can include the loss of life. The evaluation of a new diagnostic imaging test is a complex and protracted process, often described by a few critical indices; values that are used to judge clinical utility and the cost/benefit ratio of the new technology. Suboptimal sensitivity and specificity in early clinical trials may significantly reduce investigator enthusiasm and research funding support. Our initial evaluation of antifibrin scintigraphy in the low prevalence patient population caused concern that scan sensitivity (Table 2, co-positivity 20%–30%) was too limited for the procedure to be of clinical value. The subsequent maximum likelihood analysis, however, suggests that antifibrin scintigraphy compares well with contrast venography.

CONCLUSION

The use of contrast venography as a gold standard causes significant underestimation of scan sensitivity and specificity and potential pessimism that is unjustified. Thus, the maximum likelihood method provides a means to avoid the bias that the use of a gold standard is likely to entail. The maximum likelihood procedure also provides an important benefit to the execution of research trials. Diagnostic procedures with high accuracy are obviously desirable but are frequently too expensive or hazardous to be used on a large scale. Therefore, when investigating disease in large population groups, a less sophisticated screening test with greater error rates may be preferred. The use of a maximum likelihood analysis of a gold standard-less comparison between two populations of patients with different disease prevalence should make it possible to get accurate estimates of the trial test without choosing a highly accurate and potentially costly standard test or requiring the assumption that the standard test is error-free (*6*).

ACKNOWLEDGMENTS

We thank the investigators of the multicenter trial and Centocor Inc. who provided access to the data. The authors wish to thank Roberta Lukasiewicz, Paul Neuman, Bruno Caridi and John Lo-Piccolo who helped organize and process the statistical results. We also thank Eric Weinberg for reviewing this manuscript, James Ashton for his guidance and Cathy Maahs-Fladung of SAS Inc. for her help in implementing and executing the maximum likelihood procedure.

REFERENCES

1. Dalen JE, Alpert JS. Natural history of pulmonary embolism. *Prog Cardiovasc Dis* 1975;17:259–270.
2. Morrell MT, Dunnill MS. The post-mortem incidence of pulmonary embolism in a hospital population. *Br J Surg* 1968;55:347–352.
3. Gart JJ, Buck AA. Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. *Am J Epidem* 1966;83:593–602.
4. Greenberg RA, Jekel JF. Some problems in the determination of the false-positive and false-negative rates of tuberculin tests. *Am Rev Respir Dis* 1969;100:645–650.
5. Schaible T, Dewoody K, Weisman H, Line B, Keenan A, Alavi A. Accurate diagnosis of acute deep venous thrombosis with technetium-99m-antifibrin scintigraphy: final phase 3 trial results. *J Nucl Med* 1992;33:848.
6. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. *Biometrics* 1980;36:167–171.
7. Ashton JJ, Moeschberger ML. A SAS macro for estimating the error rates of two diagnostic tests, neither being a gold standard. *SAS Users Group International Conference Proceedings* 1989:995–996.
8. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950;3:32–35.
9. Sherry S. The problem of thromboembolic disease. *Semin Nucl Med* 1977;7:205–211.
10. Knight LC. Imaging thrombi with radiolabeled antifibrin monoclonal antibodies. *Nucl Med Commun* 1988;9:823–829.
11. Kudryk B, Rohoza A, Ahadi M, Chin J, Wiebe ME. Specificity of a monoclonal antibody for the NH2-terminal region of fibrin. *Mol Immunol* 1983;20:1191–1200.
12. Henkelman RM, Kay I, Bronskill MJ. Receiver operator characteristic (ROC) analysis without truth. *Medical Decision Making* 1990;10:24–29.
13. Phelps CE, Hutson A. Estimating diagnostic test accuracy using a fuzzy gold standard. *Medical Decision Making* 1995;15:44–57.