Do Diagnostic Algorithms Always Produce a Uniform Lung Scan Interpretation?

James A. Scott and Edwin L. Palmer

Division of Nuclear Medicine and Department of Radiology, Massachusetts General Hospital, Boston, Massachusetts

Several algorithms have been devised to assist in the interpretation of ventilation-perfusion (V/Q) scans performed to diagnose pulmonary embolism. The degree to which adherence to a single algorithm facilitates diagnostic homogeneity among different readers, however, has been little investigated. We evaluated the individual variability in V/Q lung scan interpretation in a large, academic nuclear medicine division to determine the degree of interpretive heterogeneity among a group of physicians all using the same image interpretation algorithm. Ventilation-perfusion scan interpretive patterns and the diagnostic accuracy of individual physicians were evaluated using quantitative parameters to establish group norms and to detect variations from these norms. The performance of each reader was tracked over a 4 yr period. There was a significant variation in V/Q interpretive patterns and diagnostic accuracy between readers despite the attempted use of a uniform diagnostic algorithm. Subgroups of interpretive styles could be defined based on the percentage of intermediate (including both indeterminate and intermediate categories) scans read. Although there was significant variation in diagnostic accuracy among readers, there was no obvious correlation between accuracy and reading style except that the most nonstandard diagnostic patterns were associated with the most variable diagnostic accuracy. These data show a measurable variation in interpretive patterns and accuracy among multiple readers of V/Q scans despite attempted group adherence to an established diagnostic algorithm.

J Nucl Med 1993; 34:661–665

Quality assurance (QA) issues have become an increasingly important component of medical practice (*1*). The initial organized QA efforts were motivated by the Joint Commission on Accreditation of Healthcare Organizations (JCAHO) regulations. Early topics of QA in diagnostic imaging included technical quality, appropriateness of examinations and the delivery of timely service (*2*). As QA methods have matured, there has been an increasing shift from these and similar topics to the direct evaluation of physician performance, often a more demanding task (*3*). The most common methods employed to evaluate physicians' work are double reading and random review of imaging studies (*4*). To date, relatively little attention has been paid to the variability in scan interpretative patterns among physicians using a standard diagnostic algorithm such as the PIOPED (*5*) criteria for ventilation-perfusion (V/Q) scan interpretation. Such data are of interest to quality assurance efforts both in terms of physician training and the process of continuous quality improvement.

Within the field of nuclear medicine, the diagnostic accuracy of many physiologically-based observations may be especially difficult to confirm. Although V/Q images show physiologic data, they are used to infer morphologic information about the presence or absence of emboli. Pulmonary arteriography provides a definitive "proof" of this and is widely employed as a "gold standard." Although imperfect, carefully performed pulmonary arteriography provides a reliable standard against which V/Q scans can be judged.

We hypothesized that, despite attempted conformity to uniform interpretive criteria, there were significant individual variations in their clinical application. Here, V/Q imaging provides a convenient model since a limited number of diagnostic categories exist together with a gold standard which has a binary result (angiography). Thus, we examined the interpretive patterns among a group of six readers over a 4-yr interval during which a uniform diagnostic algorithm was applied by all readers.

MATERIALS AND METHODS

The V/Q scan data were acquired for the nuclear medicine division of a tertiary care hospital during a 4-yr period from February 1988 through April 1992. Included in this database were the reader interpreting the study, the interpretation, the presence or absence of angiographic follow-up and the results of such follow-up. Each reader classified V/Q scans into one of four categories; normal, low probability, intermediate or indeterminate probability and high probability of pulmonary embolism. Technically limited examinations were excluded from consideration, as were studies not performed for the clinical

Received Aug. 11, 1992; revision accepted Nov. 17, 1992.
For reprints or correspondence contact: James A. Scott, Dept. of Radiology, Massachusetts General Hospital, Boston, MA 02114.

TABLE 1
Interpretive Accuracy

	Reader					
	1	2	3	4	5	6
Accuracy	86%	81%	85%	82%	83%	67%*
NI/Low (+)	15%	26%	18%	25%	20%	25%
Int (+)	49%	44%	29%	38%	39%	19%
High (+)	88%	100%	92%	100%	88%	62%
Total scans	625	613	586	370	273	211
Total angiograms	72	87	89	57	41	33
%Angiograms	12%	14%	15%	15%	15%	16%

*$p < 0.05$

indication of suspected pulmonary embolism. A chart listing the PIOPED interpretive criteria was present on the wall adjacent to the reading viewboxes for reference.

There were six individuals who regularly interpreted studies during this period. Two of these readers were involved, on a rotating basis, each day in the clinical interpretation of V/Q scans, among other studies. The staffing assignment was made according to a fixed schedule during the first 2.5 yr of the study, after which the schedule became more variable owing to factors unrelated to this study. While the two daily readers do not normally interpret clinical studies jointly, they may consult with one another on difficult or unusual cases.

Each reader's accuracy was defined as the number of normal and low probability scans showing no embolus at angiography plus the number of high probability scans showing embolism at angiography divided by the total number of normal, low and high probability studies. Intermediate and indeterminate probability scans were not included in this calculation since the accuracy of this interpretation is not readily definable. Thus, we defined accuracy from a perspective of leading the clinician towards or away from the correct diagnostic direction.

Cluster analysis is a multivariate process for detecting groupings in data. It permits the grouping of observations from a multivariate distribution into clusters of similar points. The number of clusters is specified a priori and the obtained groupings are then tested for statistically significant differences. For instance, three groups (clusters) may be specified that differ significantly in terms of the values of one or more variables. No significant differences may be found, however, when the data set is subjected to a four-group analysis. This would indicate that the data set may be only meaningful when divided into three groups. Statistical analysis was performed using the F test followed by the t-test corrected for multiple comparisons (Bonferroni) for comparison between means, the chi-square test for comparison between rates and the K-means method for analysis of clustering (*6*).

Cluster analysis was performed as follows: Each reader's interpretations were sorted into the above four categories of pulmonary embolism probability. The percentage of total interpretations in each category was established and compared between readers. The numerical differences between percentages for each interpretive category were then summed over all categories to produce the reading similarity (RS) value, reflecting the cumulative absolute value of the percent difference in reading patterns between two interpretive styles. These differences might be between two readers (e.g., RS_{12} between Readers 1 and 2) or between a reader and a published algorithm (e.g., RS_{1P} between Reader 1 and the PIOPED distribution). Thus, if Reader 1 interpreted 5% normal, 50% low, 25% intermediate and 20% high, while Reader 2 interpreted 10% normal, 50% low, 30% intermediate and 10% high, the RS_{12} value would be (10–5) + (50–50) + (30–25) + (20–10) = 20. Additional parameters were established as follows:

N = number of normal scans read.
L = number of low probability scans read.
I = number of intermediate or indeterminate probability scans read.
H = number of high probability scans read.
1 and 2 subscripts = Reader 1 and Reader 2.
P subscript = PIOPED published data.
Reading Similarity (RS_{12}) between Readers 1 and 2: $|(\%N_1 - \%N_2)| + |(\%L_1 - \%L_2)| + |(\%I_1 - \%I_2)| + |(\%H_1 - \%H_2)|$.

The RS value was also used to compare each reader's distribution of interpretations to that obtained during the PIOPED study. Here it was assumed that the PIOPED investigators, working from and faithful to a consensus interpretive scheme, would act as an individual reader (R_P). The distribution of interpretations in the PIOPED study as determined by the percentage scans falling into normal, low, intermediate or indeterminate and high probabilities was compared to that of each reader. The RS value between Reader 1 and PIOPED (R_{1P}) was thus calculated as above for two individual readers except that the PIOPED interpretive distribution was used instead of that of the second reader. Over a sufficiently large number of cases, the degree of adherence to the PIOPED interpretive criteria should be reflected in these outcome values. Although this presumes similarity of the patient population to that in the PIOPED study, there are at least two reasons to infer that this is the case. First, this institution participated in the PIOPED study. Second, the average probability of pulmonary embolism per scan (estimated as the sum of the percent readings in each diagnostic category multiplied by the probability of embolism in that category as verified at angiography) was similar for this group of patients and the PIOPED study (31%). These figures are, however, subject to selection bias, particularly in our own data where no attempt was made to insure uniform angiographic follow-up.

TABLE 2
Distribution of Interpretations at Conclusion of Study

Reader	Normal	Low	Inter/Indet	High	$R_N R_P$
1	5.6%	57.9%	30.6%	5.9%	48.8
2	5.3%	55.9%	31.4%	7.3%	44.9
3	10.6%	52.7%	25.9%	10.8%	38.4
4	7.9%	50.6%	28.8%	12.7%	34.2
5	9.2%	60.0%	20.3%	11.1%	52.4
6	3.9%	58.9%	22.7%	14.5%	53.2
PIOPED	14.1%	33.5%	39.1%	13.3%	

RESULTS

Table 1 shows the accuracy of the various readers with subcategorization as to the incidence of pulmonary embolism in each diagnostic category. The number of scans interpreted during the study, the number of cases in which angiographic follow-up was obtained and the percentage of interpreted studies on which angiographic follow-up was pursued is shown for each reader. Although the number of readers (six) was too small to permit statistical validation of small differences, the Pearson correlation coefficients describing the relationship between accuracy and individual similarity to the PIOPED interpretive scheme (RS_{1P}, RS_{2P}, RS_{3P}, RS_{4P}, RS_{5P} and RS_{6P}) was 0.524; 0.674 between accuracy and the total number of scans interpreted; and 0.397 between accuracy and the percentage of total scans interpreted as intermediate. Of the two readers with interpretive patterns most dissimilar to PIOPED, one performed with an interpretive accuracy significantly less than the group mean (Reader 6, $p < 0.05$). The accuracy of the other (Reader 5) was similar to the group mean.

The range of positive angiograms performed in patients with intermediate scan interpretations varied between 19%–49% (Reader 6 and Reader 1, respectively). Across the group as a whole, 39% of intermediate readings showed embolism at angiography.

Table 2 shows the interpretive distribution for each reader at the conclusion of the study in April 1992. Here, the PIOPED classifications normal and near normal have been grouped together under normal probability. The RS values between each reader and PIOPED (RS_{1P}–RS_{6P}) provide a measure of the similarity of the individual reader's interpretive style to that obtained in the PIOPED study. The smaller the RS_{NP} value, the more closely that reader's interpretations were in accord with the distribution obtained in the PIOPED study.

Partitioned cluster analysis was performed to determine whether certain reading "styles" could be identified within the group. This analysis was based upon the RS between each reader and the other five. The results of this analysis are shown in Table 3. This analysis was performed at yearly intervals from April 1989 to April 1992 to determine the stability of such clustering. The interpretive category which most effectively classified the readers into the given group is indicated as the classifier. The most effective classifier was the percentage of intermediate diagnoses. In other words, the six staff members were sorted into different diagnostic reading styles based upon the number of nondefinitive readings produced as a percentage of the total scans interpreted. The actual interpretive percentages by group are shown at the bottom for April 1991 and April 1992 and are separated into either two or three groups. For the two group cluster analysis, Group A_2 consists of Readers 1, 2, 3 and 4; Group B_2 consists of Readers 5 and 6. Note that Group B_2 read fewer intermediate scans than did Group A_2. This grouping of reading patterns evolved and stabilized with time as shown in the data reflecting changes from April 1991 to April 1992. In an effort to further characterize this heterogeneity in reading style, the patterns were grouped into three different groupings as shown at the right of Table 3. The statistical success of this three-group cluster analysis indicated that interpretive styles among the group could be further refined past the classification produced by the two-group analysis. These three group sortings also evolved into a relatively stable configuration during the final 2 yr of the study. The three group cluster analysis produced Group C_3 (Readers 1 and 2), Group D_3 (Readers 3 and 4) and Group E_3 (Readers 5 and 6). The normalized Euclidian distance metric, an index of the degree of similarity within the groups, was 0.65 for Group C_3; 1.22 for Group D_3; and 1.71 for Group E_3. Thus, Readers 5 and 6, although classified into a single group, showed the least similarity in their interpretive patterns. This is a consequence of the cluster analysis' limitations: it is asked only to obtain the most effective separation of the data into a given number of groups. It does not guarantee a high degree of homogeneity within each group. Note that the three-way grouping partitioned Group A_2 into two subgroups with different interpretive styles (Groups C_3 and D_3), based upon the percentage of low and nondefinitive interpretations. Notably, Reader 1 and Reader 2 (comprising Group C_3) had the most similar interpretive patterns across the group as a whole ($p < 0.05$ by cluster analysis of RS values) and were the only two readers who consistently worked together in the clinic during the study. Cluster analysis using four or more groups failed to produce a

TABLE 3
Interpretive Style Groupings Among Readers

Date	Two group			Three group			
	A_2	B_2	Classifier	C_3	D_3	E_3	Classifier
April 1989	1,2 3,4 5	6	Low p = 0.027	1,2,4	3,5	6	Int p = 0.006
June 1990	3,4 5,6	1,2	Int p = 0.004 High p = 0.021	1,2	3,4,5	6	Int p = 0.007
April 1991	1,2 3,4	5,6	Int p = 0.024	1,2	3,4	5,6	Low p = 0.017 Int p = 0.001
April 1992	1,2 3,4	5.6	Int p = 0.018	1,2	3,4	5,6	Low p = 0.019 Int p = 0.020

	Group percentages								
	N1	Low	Int	High		N1	Low	Int	High
April 1991									
Group A_2	7%	53%	30%	10%	Group C_3	6%	55%	33%	7%
Group B_2	7%	58%	22%	14%	Group D_3	9%	52%	27%	12%
					Group E_3	7%	58%	22%	14%
April 1992									
Group A_2	7%	54%	29%	9%	Group C_3	5%	57%	31%	7%
Group B_2	7%	59%	22%	13%	Group D_3	9%	52%	27%	12%
					Group E_3	7%	59%	22%	13%

significant categorization, indicating that the data were incapable of further refinement.

DISCUSSION

Our purpose in this study was to determine the extent to which the application of a uniform set of diagnostic image criteria would lead to homogeneous interpretive patterns and diagnostic accuracies. Our results suggest that neither was the case not only because there was significant variability in interpretive accuracy among the readers but also three different interpretive patterns developed despite attempted adherence to a single diagnostic algorithm on the part of all readers.

We employed one of several algorithms that have been presented for the interpretation of radionuclide V/Q scans, most commonly including the Biello (*7*), McNeil (*8*) and PIOPED (*5*) systems. Each staff member included in this study subscribed to the interpretive algorithm defined by the PIOPED investigators. There is no specific test by which a physician can be certified as correctly adhering to a single interpretive style. No panels of unbiased reading style classification experts exist and even if they did, it is difficult to envision how a reliable test could be devised to distinguish interpretive styles. For this reason, we used the RS values over a large number of interpretations to attempt to define interpretive styles among the physicians studied. Needless to say, a physician's low RS value as compared to PIOPED alone does not guarantee that the physician uniformly implements the PIOPED criteria. Over a large number of scans, however, close adherence to a particular diagnostic algorithm should produce a distribution of scan classifications similar to those produced in large published studies using the same algorithm and amenable to statistical analysis. Similarly, when applied to a group of physicians ostensibly adhering to the PIOPED diagnostic criteria, it is likely that those with the smallest RS values with respect to the PIOPED study adhere closest to the PIOPED diagnostic criteria. A significant difference in nondefinitive interpretations between two readers suggests that the readers use different interpretive styles since there is no reason to believe that V/Q scan patients assort somehow differently between the various readers.

As a group, the distribution of interpretations was somewhat dissimilar to the findings of the PIOPED study in that a higher percentage of studies were interpreted as "low probability" (52%–59%) than in the PIOPED data (39%). This difference, however, becomes less striking if normal, near-normal and low probability scans are grouped together. Although generally consistent with other published reports, our data are not in accord with all studies of V/Q scans, such as the more controversial results of Hull et al. (*9*).

In our data, there was a significant variation between staff members both in the categorical distribution of interpretations and the accuracy of these interpretations.

There was a broad range of apparent interpretive patterns among the staff. The percentage of nondefinitive scans read ranged from 20% (Reader 5) to 31% (Reader 2) by the end of the study and the percentage of angiographically-proven pulmonary embolism in nondefinitive scans varied from 19% (Reader 6) to 49% (Reader 1). Despite these differences, the only significant variation in accuracy occurred in the reader whose interpretive style varied most from the PIOPED interpretive distribution. Clearly, it would be possible for a reader to increase accuracy by placing any study with the slightest diagnostic uncertainty into the intermediate category. This process, taken to extremes, would render the scan valueless. That this did not happen in our study is suggested by the poor correlation between accuracy and the percentage of studies interpreted as intermediate and by our finding that most readers classified fewer studies as intermediate than did PIOPED. In particular, the reader with the fewest intermediate classifications (Reader 5) maintained an acceptable accuracy figure of 83%. Nonsignificant associations were obtained between interpretive accuracy and both reading similarity to PIOPED (RS_{NP}) and the number of scans interpreted. These trends lend some support to the reasonable hypotheses that accurate scan interpretation increases with clinical activity and conformity to an established interpretive scheme. Our results suggest that the readers whose interpretive styles (based on RS values) were most similar also had homogenous accuracies.

The percentage of nondefinitive readings was best able to identify the different interpretive groups as shown in Table 3. At least two different interpretive patterns were readily apparent by the end of the study. It is possible that the variety of existing interpretive schemes may have contributed to the observed variability in interpretive patterns even in this presumably homogenous group. Thus, although a particular reader may consciously subscribe to the PIOPED criteria, awareness of other interpretive schemes may "contaminate" the application of this algorithm.

This type of quantitative analysis of image interpretation is only possible if a sufficient database permits reliable statistical analysis by virtue of an accessible "gold standard." Ventilation-perfusion imaging lends itself to this type of study because a binary gold standard exists (pulmonary angiography) together with a limited number of well-defined scan categories (facilitating statistical analysis), and a sufficiently large database exists to permit statistical analysis of the results. The relatively large number of physicians (six) regularly involved in the performance of these studies allowed us to characterize a spectrum of interpretive patterns.

Our data are subject to some of the limitations encountered by many previous studies of V/Q scanning. The frequency of pulmonary emboli in each diagnostic category is difficult to establish from our data because only selected patients underwent pulmonary angiography. In general, these patients include those in whom the V/Q scan findings are discordant with the clinical assessment, those in whom an indeterminate or intermediate scan requires further pursuit and those requiring inferior vena cava (IVC) filter placement because of a contraindication to heparinization. Accuracy in the high probability category exceeds that for low probability scans because the former group of patients often undergo arteriography at the time of IVC filter placement rather than because of significant discord between scan and clinical findings. This should not influence the comparative accuracies between the different readers, however, since there is no reason to assume that, for example, the scans of patients requiring IVC filter placement do not assort normally among the various readers.

CONCLUSIONS

In conclusion, our study has determined the following two points: (1) there is a spectrum of interpretive patterns for V/Q scanning among practicing nuclear physicians despite self-perceived adherence to uniform diagnostic criteria and (2) deviations from standard interpretive patterns may be more likely to be associated with diminished diagnostic accuracy.

REFERENCES

1. Athanasoulis CA, Thrall JH. Standard of radiology practice: an approach to development. *Radiology* 1989;173:613–614.
2. Friedman BI. Quality assurance and nuclear medicine: the challenge of change. *J Nucl Med* 1986;27:1366–1372.
3. Cascade PN. Quality improvement in diagnostic radiology. *AJR* 1989;154: 1117–1120.
4. Lamki LM, Haynie TP, Podoloff DA, Kim EE. Quality assurance in a nuclear medicine department. *Radiology* 1990;177:609–614.
5. PIOPED Investigators. Value of the ventilation/perfusion scan in acute pulmonary embolism. *JAMA* 1990;263:2753–2759.
6. Wilkinson L. *SYSTAT: the system for statistics.* Evanston, IL: SYSTAT Inc., 1990;18–47.
7. Biello DR, Mattar AG, McKnight RC, Siegel BA. Ventilation-perfusion studies in suspected pulmonary embolism. *AJR* 1979;133:1033–1037.
8. McNeil BJ. Ventilation-perfusion studies and the diagnosis of pulmonary embolism: concise communication. *J Nucl Med* 1980;21:319–323.
9. Hull RD, Raskob GE. Low probability lung scan findings: a need for change. *Ann Intern Med* 1991;114:142–143.