Neural Network in the Clinical Diagnosis of Acute Pulmonary Embolism*

Sanjay Butil, M.D.; Jerald W. Henry, B.S.; Melvin Rubenstein, M.D., F.C.C.P.; and Paul D. Stein, M.D., F.C.C.P.

The purpose of this investigation was to test the hypothesis that computer-based pattern recognition can accurately assess the likelihood of acute pulmonary embolism (PE) based on readily obtainable clinical characteristics. Data were obtained from 1,213 patients who participated in the collaborative study of the Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED). Characteristics of the history, physical examination, electrocardiograph, chest radiograph, and arterial blood gases of patients with suspected acute PE were presented to a backpropagation neural network. The 1,213 patients were divided into training set A (n = 606) and test set B (n = 607). These groups were then reversed into training set B (n = 607) and test set A (n = 606). A receiver operating characteristic (ROC) curve was constructed from PIOPED clinical assessment, and from neural network clinical assessment in groups A and B. Areas under the respective ROC curves were 0.7450, 0.7477, and 0.7324. All differences were not significant. Areas under ROC curves for PIOPED clinical assessment combined with ventilation/perfusion (V/Q) scan results were compared with neural network clinical assessment combined with V/Q scan results in groups A and B. The respective ROC areas were 0.8324, 0.8203, 0.8496 (all differences not significant). These data show that neural networks were able to predict the clinical likelihood of PE with an accuracy comparable to experienced clinicians.

(CHEST 1993; 104:1685-89)

PE = pulmonary embolism; PIOPED = Prospective Investigation of Pulmonary Embolism Diagnosis; ROC = receiver operating characteristic; V/Q = ventilation/perfusion

The collaborative study of the Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED) showed that physicians with a special interest and experience in the diagnosis of acute pulmonary embolism (PE) were able to combine their prior clinical assessment with the radiologist's interpretation of the ventilation/perfusion (V/Q) lung scan to enhance the noninvasive diagnostic assessment. Other investigators have had a similar experience. Unfortunately, concordant diagnostic combinations were uncommon, occurring in only 25 percent of patients with clinically suspected PE. Either clinical uncertainty or uncertainty regarding the V/Q lung scan (intermediate probability V/Q scan) was present in 72 percent of patients. Nevertheless, when the clinical assessment and V/Q scan interpretation were concordant, the noninvasive diagnosis or exclusion of PE was strongly established.

The PIOPED and Canadian Investigators had a special interest and experience in the diagnosis of acute PE. Whether physicians who see acute PE only occasionally would have the same ability to make the clinical diagnosis of acute PE is uncertain. Some investigators have used an artificial intelligence paradigm known as a neural network to integrate multiple clinical variables to arrive at a correct diagnosis for other problems such as myocardial infarction and low back disorders. In the present investigation, we tested the hypothesis that computer-based pattern recognition, using an artificial neural network, with readily obtained objective clinical observations and historical data would perform as well as physicians with a special interest and experience in the diagnosis of acute PE.

METHODS

Data in this study were obtained from the national collaborative study of PIOPED. The eligible population consisted of patients aged 18 years or older in whom acute PE was of diagnostic concern. Symptoms suggestive of PE were required within 24 h of entry into the study. The PIOPED study consisted of two arms. In one arm, patients who consented to participate in the investigation were obligated to undergo pulmonary angiography if their V/Q scans were abnormal. The diagnosis in these patients was made by pulmonary angiography or autopsy. Pulmonary embolism was excluded by a normal pulmonary angiogram and/or outcome classification among patients who received no antithrombotic therapy.

There was a second arm in PIOPED that was not included in the original PIOPED report. This group included patients who by random sample were not selected for sensitivity and specificity analyses of the V/Q scans, and who, therefore, were not obligated by protocol to undergo angiography if their V/Q scans were abnormal. Many of these patients underwent diagnostic angiography at the request of their attending physicians. In these patients, the diagnosis of PE and the exclusion of PE was made by angiography. In total, there were 1,213 patients who had a clinical assessment and pulmonary angiography or clinical outcome assessment. Among these, 390 had PE and 833 did not have PE.

A history and physical examination were completed before the V/Q scans were obtained. Characteristics of the history and objective data from the physical examination were tabulated for each patient. A clinical assessment of the likelihood of acute PE was based on all available noninvasive data with the exception of the V/Q scan. The basis for arriving at the estimate of clinical likelihood was individual clinical judgment, and not any specific predetermined criteria.
Chest radiographs were obtained within 24 h of angiography in all patients. The partial pressure of oxygen in arterial blood (PaO₂), with the patient breathing room air, was measured within 24 h before the diagnostic pulmonary angiogram in patients with and without PE. Detailed methods have been reported in PIOPED. Pulmonary angiograms were completed within 24 h, and most were completed within 12 h after performance of the V/Q scan. Criteria for the diagnosis of PE were the identification of an embolus obstructing a vessel or the outline of an embolus (filling defect) within a vessel. Detailed methods for performing and interpreting the pulmonary angiograms, methods of performing V/Q scans, as well as the criteria for their interpretation have been described.

Artificial Neural Networks

Artificial neural network models were developed by presenting clinical data to the neural network in the form of numeric input variables. Fifty characteristics of the history, physical examination, chest radiograph, electrocardiogram, and arterial blood gases were presented to the computer algorithm (Table 1). These variables were coded as either binary variables (present or absent) or continuous values. The neural network used the V/Q assessment as reported in PIOPED for each scan probability. No computerized interpretation of V/Q scans was performed by the neural network.

Patients with a diagnosis of PE were assigned a score of 100, and those without PE were assigned a score of zero. The likelihood of PE was described as a percent probability score. During the training period, the neural network repetitively evaluated its error in each patient. The weights assigned to individual variables were altered during training based on a least mean squares analysis known as backpropagation. At the completion of the training period, the weights assigned were fixed within the neural network.

The neural network models were developed using specific software (Neuroshell, Ward Systems, Frederick, Md). This software used a feed-forward single hidden-layer backpropagation training algorithm with a sigmoidal transfer function. The population was divided into those with PE and those with no PE. Both groups were randomly divided in half. This gave group A with 606 patients (190 with PE and 416 with no PE) and group B with 607 patients (190 with PE and 417 with no PE). Group A was used as a training set and the resulting neural network was tested on group B. Subsequently, group B was used as a training set and the resulting neural network was tested on group A. The areas under receiver operating characteristics (ROC) curves of tested group A and tested group B were compared with ROC curves obtained by PIOPED clinical assessment.

The neural networks were evaluated for clinical assessment alone and for clinical assessment combined with V/Q scan results. The V/Q probability of PE was determined by V/Q scan readers in PIOPED. The V/Q scan diagnosis was not modified or reassessed by the neural network.

Sixty patients in the learning set of both group A and group B were set aside as an incremental evaluation group. These incremental evaluation groups were needed because of the tendency of backpropagation neural networks to “overlearn” their training sets. The error on the incremental evaluation group was used as the end point in learning. When the error would no longer decrease, learning was terminated.

Initially, learning was started with a learning rate of 1.0 and a momentum rate of 0.05, with 10 hidden nodes. The models tended to rapidly overlearn their training sets. We, therefore, began modeling again with learning and momentum rates both set at 0.01. In the remodeling, we used one hidden node in each model. When learning would no longer proceed after 50,000 learning events, we sequentially increased the number of hidden nodes by one until there were a total of 4 hidden nodes. The clinical neural network used 50 input nodes (50 input variables). The neural network that combined clinical characteristics with V/Q scans used 54 input nodes (54 variables).

When data were missing for a given patient, intermediate values for variables were substituted and presented to the neural network for both training and validation sets. For example, if a patient had calf tenderness, the variable was coded with the value 1. If the patient did not have calf tenderness, the variable was coded as 0. If the information regarding calf tenderness was not recorded, the
variable was assigned a value of 0.5.

Statistical Methods

Sensitivity was defined as the proportion of patients with PE in whom the diagnosis was made. Specificity was defined as the proportion of patients in whom the diagnosis of PE was correctly excluded among those who did not have PE.

For each of the diagnostic methods, a plot of 1 minus the specificity (false-positive rate) vs the sensitivity at each level of classification was created. This curve is known as a ROC curve. The area under the curve was calculated by summing the areas of the trapezoids between diagnostic cutoffs. Differences in the areas under the ROC curve were assessed for statistical significance by the method of Hanley and McNeil.2

The total sum of squared error between neural network estimate and actual outcome, and between PIOPED clinicians' estimate and actual outcome were calculated by the method described by Rumelhart and associates.5

A program (SAS PROC Logistic, SAS Institute, Cary, NC) was used to create a logistic model from combined PIOPED clinical assessment and V/Q scan interpretation for the estimation of statistical parameters. The logistic model used the following coefficients: intercept, 3.8303; high probability scan result, −4.2983; intermediate probability scan, −1.8927; low probability scan result, −1.1180; and clinical assessment, −0.0313.

The specificity of the models at a sensitivity of 90 percent was evaluated for both the PIOPED clinician assessment and the neural network. The corresponding false-positive, true-negative, and false-negative rates for both methods were calculated.

Results

The area under the ROC curve for clinical assessment alone, based on the PIOPED clinical assessment (n = 1,213) vs the neural network clinical assessment

Table 2—Nonparametric Estimates of Areas Under the Receiver Operating Characteristic Curves*

<table>
<thead>
<tr>
<th></th>
<th>Area ± Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIOPED clinical (n = 1,213)</td>
<td>0.7450 ± 0.015</td>
</tr>
<tr>
<td>PIOPED clinical set A (n = 606)</td>
<td>0.7346 ± 0.022</td>
</tr>
<tr>
<td>PIOPED clinical set B (n = 607)</td>
<td>0.7560 ± 0.021</td>
</tr>
<tr>
<td>Neural network clinical Training set A (n = 606)</td>
<td>0.8270 ± 0.019</td>
</tr>
<tr>
<td>Test set B (n = 607)</td>
<td>0.7324 ± 0.022</td>
</tr>
<tr>
<td>Training set B (n = 607)</td>
<td>0.8264 ± 0.018</td>
</tr>
<tr>
<td>Test set A (n = 606)</td>
<td>0.7477 ± 0.022</td>
</tr>
<tr>
<td>PIOPED clinical + V/Q scan (n = 1,213)</td>
<td>0.8324 ± 0.013</td>
</tr>
<tr>
<td>PIOPED clinical + V/Q scan set A (n = 606)</td>
<td>0.8138 ± 0.012</td>
</tr>
<tr>
<td>PIOPED clinical + V/Q scan set B (n = 607)</td>
<td>0.8515 ± 0.017</td>
</tr>
<tr>
<td>Neural network clinical + V/Q scan Training set A (n = 606)</td>
<td>0.9088 ± 0.013</td>
</tr>
<tr>
<td>Test set B (n = 607)</td>
<td>0.8496 ± 0.017</td>
</tr>
<tr>
<td>Training set B (n = 607)</td>
<td>0.9317 ± 0.011</td>
</tr>
<tr>
<td>Test set A (n = 606)</td>
<td>0.8203 ± 0.018</td>
</tr>
</tbody>
</table>

*V/Q = ventilation/perfusion scan. The following comparisons showed no significant difference: PIOPED clinical (n = 1,213) versus neural network group A or B; PIOPED clinical group A (n = 606) versus neural network group A; PIOPED clinical group B (n = 606) versus neural network group B; PIOPED clinical + V/Q (n = 1,213) versus neural network group A or B; PIOPED clinical + V/Q group A (n = 606) versus neural network group A; PIOPED clinical + V/Q group B (n = 607) versus neural network group B.

for group A (n = 606) was 0.7450 vs 0.7477 (NS). The PIOPED clinical assessment vs neural network group B (n = 607) was 0.7450 vs 0.7324 (NS) (Table 2).

The areas under the ROC curves for PIOPED clinical assessment combined with PIOPED V/Q scan results (n = 1,213) vs the neural network clinical assessment combined with the V/Q scan results for group A (n = 606) was 0.8324 vs 0.8203 (NS) and PIOPED vs group B (n = 607) was 0.8324 vs 0.8496 (NS).

The above comparisons were made with the entire patient population evaluated by PIOPED (n = 1,213) vs group A (n = 606) or group B (n = 607). Comparisons were also made of PIOPED assessments of group A and B vs neural network assessments of group A and B. The ROC areas did not differ significantly (Table 2). Comparisons were also made on the basis of the total sum of squared error. These also showed no significant difference (Table 3).

The percentage of patients with PE for deciles of predicted risk based on PIOPED clinical assessment and on neural network clinical assessment is shown in Figure 1. Values of neural network group A and group B were averaged. In general, there was a good correlation between PIOPED clinical assessment and that of the neural networks.

The percentage of patients with PE for deciles of predicted risk based on PIOPED clinical assessment plus V/Q and on neural network clinical assessment plus V/Q is shown in Figure 2. The neural network clinical assessment combined with the V/Q scan results and PIOPED clinical assessment combined with the V/Q scan results showed a comparable accuracy of predicted values.

Table 3 shows the specificity for each method of assessment based on an arbitrarily selected sensitivity of 90 percent. The specificity of the neural network model that combined clinical assessment with V/Q was higher than PIOPED clinical assessment combined with V/Q, 52 percent vs 46 percent (p<0.05).

Table 3—Performance of PIOPED and Neural Networks on Groups A and B*

<table>
<thead>
<tr>
<th>Assessment Method</th>
<th>TSS</th>
<th>Specificity at 90% Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIOPED clinical assessment (n = 1,213)</td>
<td>219.13</td>
<td>35%</td>
</tr>
<tr>
<td>Neural network clinical (n = 1,213)</td>
<td>219.01</td>
<td>35%</td>
</tr>
<tr>
<td>PIOPED clinical + V/Q (n = 1,213)</td>
<td>170.18</td>
<td>46%†</td>
</tr>
<tr>
<td>PIOPED clinical + V/Q Scan (n = 1,213)</td>
<td>179.23</td>
<td>52%</td>
</tr>
</tbody>
</table>

*V/Q = ventilation/perfusion; TSS = total sum of squared error.
†PIOPED clinical + V/Q versus neural network clinical + V/Q – p<0.05.
The specificity of PIOPED clinical assessment vs neural network clinical assessment did not differ significantly.

Table 4 shows the total sum of squared error for each neural network training group, as well as the number of learning events required to achieve optimal predictive values.

**DISCUSSION**

All of the patients in PIOPED were studied because of a clinical suspicion of pulmonary embolism. Primary physicians, therefore, being well aware of the signs and symptoms of PE, were suspicious if patients had well-known clinical manifestations. Because of this, in PIOPED, patients with signs and symptoms of PE
had similar clinical characteristics as those without PE. Because of these similarities, the neural network identified other important characteristics. Foremost among these were a high alveolar-arterial gradient, a low PaO₂ on room air, atrial flutter, marked right axis or left axis deviation, ST segment depression, T-wave inversion, and surgery within the last 3 months. These variables resulted in the greatest degradation of neural network model accuracy when removed from the input variables.

One technique by which these types of data are analyzed in the neural computational literature is to cross train and test the network. The neural network was first trained on half of the patients and tested on the second half of patients. The process was then reversed and the network was trained on the second half and tested on the first half. The results were then averaged. By exposing the network to all patients, the likelihood for optimization of performance was heightened.

The ROC curve is considered a useful method of evaluating tests that have more than two categories of classification and differing sensitivities and specificities at each classification level. The areas under the ROC curves did not differ significantly, indicating that the neural networks performed as well in achieving a clinical diagnosis as physicians knowledgeable and experienced in the diagnosis of acute PE.

It may be that the neural network would perform better if more clinical information were available to it. For example, a clinician may assess the suddenness of onset of shortness of breath, its severity, whether a patient is in acute distress, and whether the patient is using accessory muscles of respiration. The neural network had available only the presence or absence of shortness of breath and the respiratory rate. A limitation of this study is that the neural network retrospectively evaluated previously obtained prospective data.

Results of previously interpreted V/Q scans were presented to the neural network in our study. Neural network paradigms recently have been applied to the evaluation of V/Q lung scans, with excellent results. Neural network paradigms also have been reported that perform computerized interpretations of scintillation data on thallium scans of the heart.

Use of a neural network may enhance the noninvasive diagnosis of PE by giving physicians an objective estimate of the probability of PE. A neural network provides estimates of the probability of PE based on objective data readily obtained by any physician. A neural network is not subjected to observer bias in its assessment of the data presented to it. Neural network assessments are reproducible and consistent over time. Neural networks may assist clinicians in selecting patients for V/Q scans, as well as assist in further diagnostic evaluation once a V/Q scan result has been obtained.

REFERENCES
6. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143:29-36