

---

# Outcome Analysis of Patients with Acute Pancreatitis by Using an Artificial Neural Network<sup>1</sup>

Mary T. Keogan, MD, Joseph Y. Lo, PhD, Kelly S. Freed, MD, Vasillios Raptopoulos, MD  
Simon Blake, MD, Ihab R. Kamel, MD, K. Weisinger, Max P. Rosen, MD, Rendon C. Nelson, MD

---

**Rationale and Objectives.** The authors performed this study to evaluate the ability of an artificial neural network (ANN) that uses radiologic and laboratory data to predict the outcome in patients with acute pancreatitis.

**Materials and Methods.** An ANN was constructed with data from 92 patients with acute pancreatitis who underwent computed tomography (CT). Input nodes included clinical, laboratory, and CT data. The ANN was trained and tested by using a round-robin technique, and the performance of the ANN was compared with that of linear discriminant analysis and Ranson and Balthazar grading systems by using receiver operating characteristic analysis. The length of hospital stay was used as an outcome measure.

**Results.** Hospital stay ranged from 0 to 45 days, with a mean of 8.4 days. The hospital stay was shorter than the mean for 62 patients and longer than the mean for 30. The 23 input features were reduced by using stepwise linear discriminant analysis, and an ANN was developed with the six most statistically significant parameters (blood pressure, extent of inflammation, fluid aspiration, serum creatinine level, serum calcium level, and the presence of concurrent severe illness). With these features, the ANN successfully predicted whether the patient would exceed the mean length of stay ( $A_z = 0.83 \pm 0.05$ ). Although the  $A_z$  performance of the ANN was statistically significantly better than that of the Ranson ( $A_z = 0.68 \pm 0.06$ ,  $P < .02$ ) and Balthazar ( $A_z = 0.62 \pm 0.06$ ,  $P < .003$ ) grades, it was not significantly better than that of linear discriminant analysis ( $A_z = 0.82 \pm 0.05$ ,  $P = .53$ ).

**Conclusion.** An ANN may be useful for predicting outcome in patients with acute pancreatitis.

**Key Words.** Computers, diagnostic aid; computers, ANN; pancreatitis.

© AUR, 2002

---

An episode of acute pancreatitis may run a variable and unpredictable course (1,2). Prediction of the likely severity of an episode in an individual patient is difficult but has implications for the institution of therapy and for pa-

tient information. For this reason, several predictive systems have been devised. Some of these systems are based on clinical criteria (eg, the Ranson and the acute physiology and chronic health evaluation [APACHE] scoring systems [2–4]), whereas others, such as the Balthazar system (5,6), are based on radiologic features at computed tomography (CT) (Fig 1). To our knowledge, no system has been devised in which both clinical and radiologic indicators are combined to predict severity or outcome. In this study, this task was investigated by using two multivariate decision model approaches: an artificial neural network (ANN) and linear discriminant analysis.

ANNs are computer models composed of parallel, nonlinear computational elements arranged in layers with a

---

**Acad Radiol 2002; 9:410–419**

<sup>1</sup> From the Department of Radiology, Beth Israel Deaconess Medical Center and Harvard Medical School, 330 Brookline Ave, Boston MA 02216 (M.T.K., V.R., S.B., K.W., M.P.R.); the Department of Radiology, Duke University Medical Center, Durham, NC (J.Y.L., K.S.F., R.C.N.); and the Russell H. Morgan Department of Radiology, Johns Hopkins University, Baltimore, Md (I.R.K.). Received December 10, 2001; revision requested December 19; revision received and accepted December 20. Supported by a grant from the Society of Gastrointestinal Radiologists. **Address correspondence to M.T.K.**

© AUR, 2002



**Figure 1.** Axial contrast material–enhanced CT scan obtained in a 51-year-old man with clinically mild pancreatitis. Note the mildly swollen pancreatic head and the mild soft-tissue stranding in the mesentery anterior to the pancreatic head (arrow). The Ranson score was 2, and the Balthazar grade was C.

structure that mimics the human brain. Computer elements (representing neurons) are nonlinear and are organized in highly interconnected layers that mimic biologic neural networks (7). The ANN can be trained with data from cases that have a known outcome. The network can evaluate the input data, recognize any pattern that may be present, and apply this knowledge to the evaluation of unknown cases. In the analysis of large data sets, networks have the advantage of relative insensitivity to noise while having the ability to discover patterns that may not be apparent to human observers (8). Because ANNs are well suited to pattern recognition, they may have a role in diagnostic radiology. Many potential applications of ANNs in diagnostic radiology have been tested, including the diagnosis of pulmonary embolism on ventilation-perfusion scans and the differentiation of benign from malignant breast lesions (9–11). Other applications include the diagnosis of diffuse lung disease and coronary artery disease with myocardial single photon-emission CT (12,13).

From a patient care point of view, identification of patients who are likely to have severe disease is very important so that treatment can be instituted as early as possible. We performed this study to investigate the ability of an ANN to predict severe illness (as predicted by length of hospital stay) with maximum sensitivity by combining both clinical and/or laboratory data and imaging input data. In addition, we compared the performance of the ANN to that of the Ranson and Balthazar grading

systems, which have previously been proposed for predicting the severity of an episode of acute pancreatitis.

For comparison, linear discriminant analysis was also performed (14). Like the ANN, both models combine multiple input findings to predict an outcome. For linear discriminant analysis, a simpler, linear decision boundary is used in  $N$ -dimensional feature space. For the ANN, conversely, an arbitrary, nonlinear decision boundary is used. Although the ANN's power has the potential to separate the two classes better, it can also overfit the data, leading to unrealistically complicated models that may not generalize well to new cases (15,16). Given the relatively small size of the current data set, it was important to assess what differences, if any, existed in the performances of linear and nonlinear models.

## MATERIALS AND METHODS

### Case Selection

The computerized hospital information system at Beth Israel Deaconess Medical Center (Boston, Mass) was searched to identify patients who had had a discharge diagnosis of pancreatitis during a 20-month period (October 1997 to June 1999). Ninety-two patients (52 men, 40 women; age range, 19–92 years) with a documented episode of acute pancreatitis who underwent abdominal and pelvic CT within 24 hours of admission were identified. Eight patients in the study group had been admitted to the intensive care unit, 12 had undergone percutaneous aspiration or drain placement for a fluid collection, three had undergone surgery, and three died.

The severity of pancreatitis was assessed by using the length of the hospital stay measured in days. The mean hospital stay was 8.4 days (range, 0–45 days; median, 5 days). With use of 8.4 days (mean stay) as the cutoff threshold, most patients ( $n = 62$ ) had a length of stay below the threshold, whereas 30 had an above average length of stay. Patients with hospital stays longer than the mean were considered to represent a positive case, whereas patients with hospital stays shorter than the mean were considered to represent a negative case. The choice of the mean as the threshold was completely arbitrary, but it was intuitively meaningful to distinguish between patients with above versus below average severity. Moreover, the greater number of negative cases is consistent with a diagnostic problem where the cost of a false-negative mistake outweighs that of a false-positive one. Because the tendency is to overestimate disease, there would be more negative cases, and it is advantageous to train

**Table 1**  
**Coding of Clinical Input Features**

Physiologic Variable	Above Normal Scores				Normal Score	Below Normal Scores			
	1.0	0.75	0.50	0.25		0.0	-0.25	-0.50	-0.75
Rectal temperature (°C)	>107	104–106	101–103	99–100	98	95–97	94–92	91–89	<88–86
Serum creatinine level (mg/dL)	≥3.5	2–3.4	1.5–1.9	...	0.6–1.4	...	<0.6	...	...
White blood cell count ( $\times 10^3/\mu\text{L}$ ) <sup>*</sup>	≥40	...	20–39.9	15–19.9	3–14.9	...	1–2.9	...	<1
Calcium level (mg/dL)	>10.2	...	...	...	8.7–10.2	...	5.6–4.1	...	<2.3
Amylase level (U/L)	>3,000	1,001–3,000	500–1,000	...	0–100	...	...	...	...
Glucose level (mg/dL) <sup>†</sup>	>501	...	400–500	...	70–105	...	50–60	...	<49
Lactate dehydrogenase level (U/L)	...	...	332–369	...	115–275	...	28–56	...	...
Base nitrogen deficit	...	...	2–4	...	...	...	...	...	...
Urea level (mg/dL) <sup>‡</sup>	>60	...	40–60	...	6–20	...	2–4	...	<2
Age (y)	>75	65–74	55–64	45–54	<44	...	...	...	...
Presence of severe chronic illness	Yes	...	...	...	No	...	...	...	...

Note.—Coding and table format are based on the APACHE 11 scheme (3).

<sup>\*</sup>Conversion factor for SI unit is  $10^6$ .

<sup>†</sup>Conversion factor for SI unit is 0.05551.

<sup>‡</sup>Conversion factor for SI unit is 0.357.

the models on the basis of a mixture that would be representative of such typical case samples.

Medical records were reviewed to determine patient age and the presence of severe chronic illness and/or immunosuppression. These data were used to calculate a Ranson score. The findings at physical examination performed at admission were noted, including temperature, mean arterial pressure, heart rate, and respiratory rate. Laboratory data noted at admission included the following: white blood cell count, serum sodium level, potassium level, creatinine level, hematocrit, bicarbonate level, calcium level, aspartate aminotransferase level, lactate dehydrogenase level, blood urea nitrogen level, base deficit, partial pressure of oxygen in arterial blood, fluid loss, lipase level, amylase level, and serum protein level. In this retrospective study, all data points listed above were not available for each patient; therefore, three demographic and/or physical data points (age, presence of other illness, temperature) and eight laboratory data points (white blood cell count, serum amylase level, lactate dehydrogenase level, calcium level, creatinine level, glucose level, blood urea nitrogen level, and base deficit), which were available for each patient, were used in the final assessment.

For entry into the ANN, these quantitative data were assigned to a five-point scale (0.0, 0.25, 0.5, 0.75, 1.0), where a value of 0.0 was assigned to a result within the normal range and higher or lower values were indicative of progressively greater deviation (positive or negative,

respectively) from the normal (on the basis of the APACHE II coding system) (3). Severe chronic illness was assigned a score of 0.0 (absent) or 1.0 (present). For age scoring, an age of 44 years or younger was considered normal, and incremental increases higher than 44 years were scored as progressively more abnormal. The encoding system is shown in Table 1 (based on the APACHE II coding system) (3).

CT scans were obtained by using a helical scanner (Hi-Speed; GE Medical Systems, Milwaukee, Wis) after intravenous injection of 150 mL of nonionic contrast material with a mechanical power injector (Medrad, Pittsburgh, Pa) at 3 mL/sec. Eight patients did not receive intravenous contrast material owing to the presence of contraindications. Oral contrast material was administered according to patient tolerance to a maximum of 1,200 mL. Section thickness varied between 3 and 10 mm; 7-mm-thick sections (pitch of 1.5) were used in most patients ( $n = 71$ ).

The admission CT scan for each patient was reviewed by an abdominal radiologist (M.T.K.) with 10 years of experience in reading abdominal CT studies. The reader was blinded to the patient's laboratory data and the eventual outcome. The radiologist used a checklist entry form consisting of 12 radiologic findings (Table 2). These features were used to assess the severity of inflammation affecting the pancreas (gland size, enhancement, vascular complications), the abdomen (presence and size of inflammation, fluid collections, pseudocysts), and the thorax

**Table 2**  
Coding of Radiology Input Features

Node Feature and Finding	Value
Pancreas size	
Normal	0.00
Increased, focal	0.25
Increased, diffuse	0.50
Pancreatic enhancement	
Normal	0.00
Inhomogeneous	0.25
Necrosis less than 30%	0.50
Necrosis greater than 30–50%	0.75
Necrosis greater than 50%	1.00
Pancreatic inflammation	
None	0.00
Haziness	0.25
Fine stranding	0.50
Coarse strands	0.75
Inflammatory mass	1.00
Inflammation extent	
None	0.00
Anterior pararenal	0.50
Mesentery	0.50
Posterior pararenal	0.75
Lesser sac	0.50
Perirenal	0.75
Flank and/or pelvis	0.50
Retroperitoneal fluid	
None	0.00
Anterior pararenal	0.50
Posterior pararenal	0.75
Lesser sac	0.50
Perirenal	0.75

*(continues)***Table 2 (continued)**  
Coding of Radiology Input Features

Node Feature and Finding	Value
Peritoneal fluid	
None	0.00
Any	1.00
Pseudocyst	
None	0.00
1	0.50
2–4	0.75
>4	1.00
Pseudocyst size	
>6 cm	0.00
<6 cm	1.00
Pseudocyst location	
None	0.00
Anterior pararenal	0.50
Posterior pararenal	0.75
Lesser sac	0.50
Perirenal	0.75
Flank and/or pelvis	0.50
Abscess	
Yes	0.75
No	0.00
Indeterminate	0.50
Vascular	
None	0.00
Pseudoaneurysm	0.75
Splenic vein thrombosis	0.25
Portal vein thrombosis	0.50
Thoracic extension	
None	0.00
Right or left	1.00

(pleural effusions). These features were ranked for entry into the ANN on a five-point scale between 0.0 and 1.0 (with 0.0 indicating a normal result and 1.0 indicating the most abnormal result). A Balthazar grade (grades A through E) (5) was calculated on the basis of the CT findings. Aspiration of abdominal or pelvic fluid detected on the initial CT scan was noted.

The 23 possible findings were then reduced to six with stepwise linear discriminant analysis by using SAS/STAT software (SAS Institute, Cary, NC) (user's guide, SAS Institute, 1998). This step was deemed necessary to minimize the effects of overtraining because of the relatively small number of cases in this data set. These six findings were (in order of decreasing importance) fluid aspiration, extent of inflammation, serum creatinine level, presence of concurrent severe illness, blood pressure, and serum calcium level.

The ANN was trained and tested with data from all 92 patients by using a round-robin method, and its performance was compared with that of linear discriminant analysis. Receiver operating characteristic (ROC) curves were generated to evaluate both predictive models. The performance of the ANN was compared with that of both the Ranson and Balthazar grading systems.

Length of hospital stay was used as a marker of disease severity. Patients were classified as representing a positive case (ie, having a longer hospital stay than the mean) or a negative case (ie, having a shorter hospital stay than the mean).

### Construction of the ANN

With use of the six findings described earlier, a three-layer back-propagation ANN (Fig 2) was constructed to predict the occurrence of a positive case. Network param-

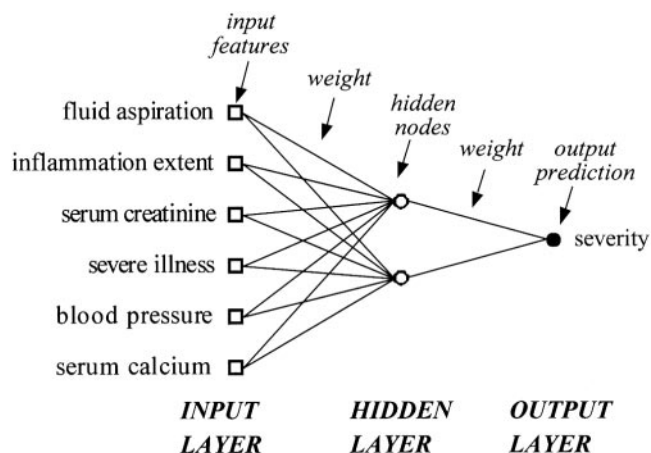
eters such as the number of training iterations, training rate and momentum constants, and number of hidden nodes were all optimized empirically to provide the best ROC area index ( $A_z$ ). The network program was custom written in the C programming language and run on commercially available workstations (UltraSPARC 80; Sun Microsystems, Mountain View, Calif). The network was constructed with three layers: an input layer with six nodes, a hidden layer with two nodes, and an output layer with one node. The six findings described earlier (ie, fluid aspiration, extent of inflammation, serum creatinine level, presence of concurrent severe illness, blood pressure, and serum calcium level) were the six input nodes, and the positive or negative patient outcome was the output node.

### Training and Testing

The network was trained and tested with all 92 cases by using a round-robin technique, as noted earlier, in which one case was left out and the network was trained on the remaining 91. Once the outcome of a test case was determined, the test case was added back to the pool of training cases and another case was removed to serve as the test case. In this way, the network was trained and tested with all 92 cases in the data base. The network was presented with the training cases 100 times (ie, 100 iterations). Each time, a new network was trained starting from random weight values, and the 92 separate testing results were pooled together in the end for analysis. In this way, the network was trained and tested on all 92 cases in the database while still maintaining independence between training and testing subsets.

The ANN was provided with the case inputs paired with the outcome (ie, length of hospital stay greater [positive] or less [negative] than the mean stay) of each training case, and network weights were updated after each case. Output of the ANN ranged continuously from 0 to 1, with greater output values corresponding to a higher likelihood of being a positive case. The ANN modified the weights of the node connections to minimize the mean squared error between the known, correct outcome and the network output. Training was halted when the ROC area index,  $A_z$ , over the testing cases was maximized.

Histograms of network output values over the testing cases for the ANN and linear discriminant analysis were generated. The specificity was calculated for both predictive models at several fixed, nearly perfect levels of sensitivity. ROC curve fitting and statistical comparisons were performed with CLABROC software (Charles Metz, Uni-



**Figure 2.** Schematic of the ANN. Diagram depicts a simplified version of the three-layer back-propagation feed-forward ANN used in this study. Inputs to the network consist of patients' clinical, laboratory, and radiologic data. Each hidden node performs a nonlinear weighted sum of all the input values; these outputs are similarly combined at the output layer. A single output node generates a single output value, which represents the network's prediction of whether disease is severe.

versity of Chicago, Ill) (17). In particular, CLABROC's area test was used to perform a univariate z-score test of the difference between the areas under the two ROC curves. The null hypothesis is that the data sets arose from binormal ROC curves with equal areas beneath them.

## RESULTS

The severity of pancreatitis varied widely among the 92 patients in this study, as evidenced by the distribution of the Ranson and Balthazar scores. The average Ranson score ranged from 0 to 9, with scores of 1–4 (suggesting less severe disease) occurring most frequently (68 patients [74%]). The distribution of patients according to Ranson score is shown in Table 3. The distribution of patients according to the Balthazar grade is shown in Table 4. In contrast to the Ranson score, the increased frequency of Balthazar grades C, D, and E is suggestive of more severe disease in most patients (79 patients [86%]).

With use of six features (ie, fluid aspiration, extent of inflammation, serum creatinine level, presence of concurrent severe illness, blood pressure, and serum calcium level), the ANN was able to predict the length of stay by using mean stay as the threshold ( $A_z = 0.83 \pm 0.05$ ) (Fig 3). ROC curves for the performances of the trained ANN and linear discriminant analysis are depicted in Figure 4.

**Table 3**  
Distribution of Ranson Scores

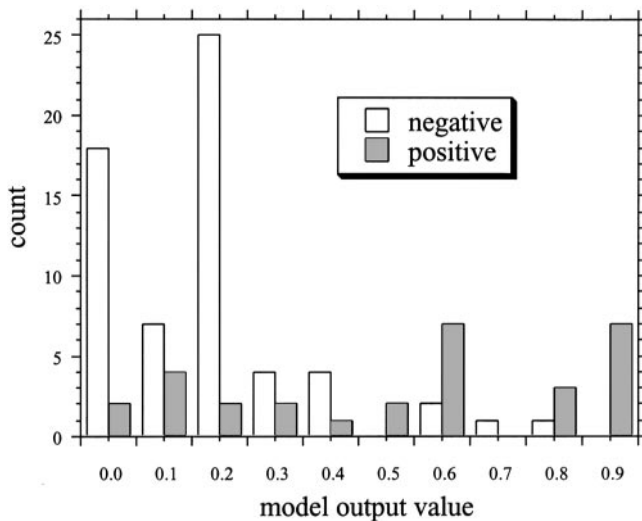
Patient Groups	Ranson Score									
	0	1	2	3	4	5	6	7	8	9
All	5	12	26	15	15	6	6	4	0	3
Negative cases	5	10	18	9	11	4	4	1	0	0
Positive cases	0	2	8	6	4	2	2	3	0	3

Note.—Data are given as numbers of patients.

**Table 4**  
Distribution of Balthazar Grades

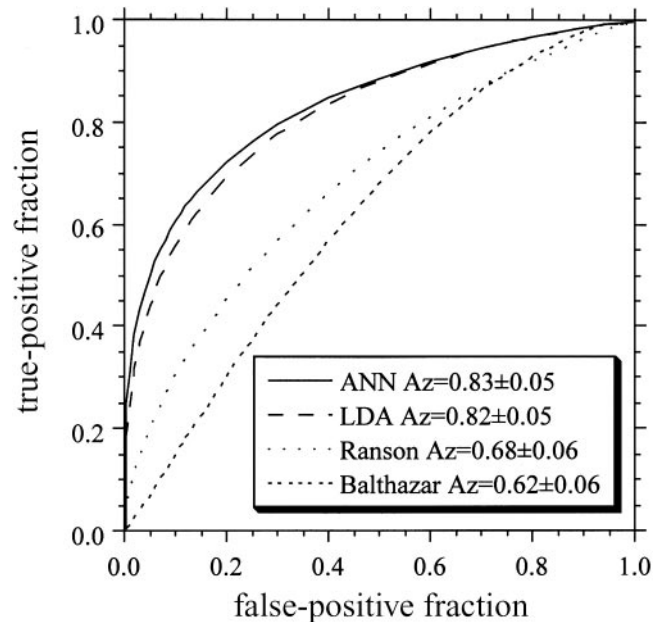
Patient Groups	Grade				
	A	B	C	D	E
All	11	2	39	38	2
Negative cases	10	1	27	23	1
Positive cases	1	1	12	15	1

Note.—Data are given in numbers of patients.



**Figure 3.** Histogram of the ANN output. With a threshold of 0.09, the model achieves a sensitivity of 100% and a specificity of 29%. Count represents the number of positive and negative cases according to score.

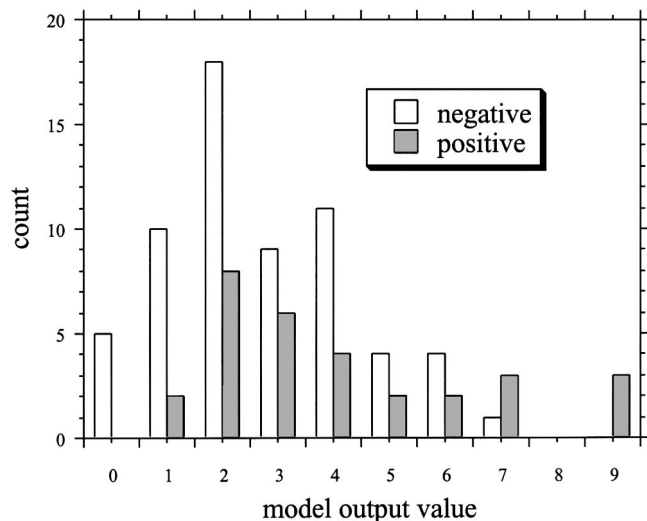
$A_z$  measures the performance of the ANN or linear discriminant analysis over the entire range of sensitivities and specificities. The mean  $A_z$  values were  $0.83 \pm 0.05$  for the ANN and  $0.82 \pm 0.05$  for linear discriminant analysis. The difference in  $A_z$  between the two models was not statistically significant ( $P = .53$ ). The performance of the ANN was also compared with that of both the Ranson score ( $A_z = 0.68 \pm 0.06$ ) and Balthazar grade



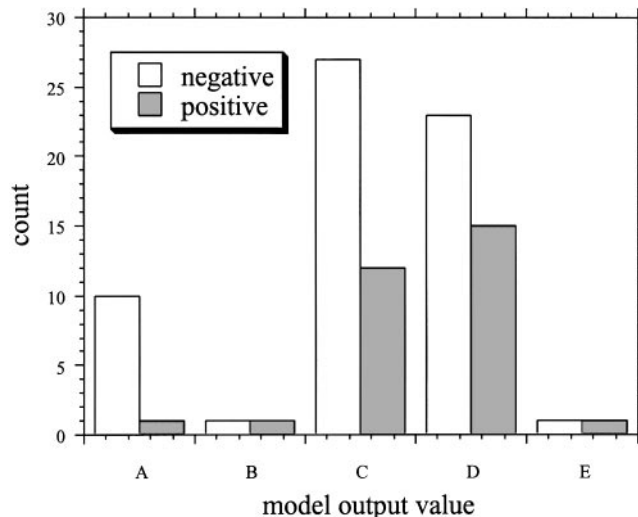
**Figure 4.** Graph shows ROC curves for ANN, linear discriminant analysis (LDA), and Ranson and Balthazar systems. Although the performance of the ANN and linear discriminant analysis is comparable for the full range of thresholds, the ANN performed slightly better than linear discriminant analysis at lower thresholds. Both predictive models performed better than the Ranson and Balthazar systems.

( $A_z = 0.62 \pm 0.06$ ). The ANN performed significantly better than both ( $P < .02$  and  $P < .004$ , respectively).

Histograms of the output values of the trained ANN, the Ranson score, and the Balthazar grade are shown in Figures 3, 5, and 6. The histograms indicate the distribution of output predictions, or decision variables, according to those models. The physician may select an arbitrary threshold over these outputs, each resulting in a different number of true- or false-positive findings and true- or false-negative findings. In the clinical setting of acute pancreatitis, it is important to maximize the sensitivity for severe disease as measured with ROC curve analysis or



**Figure 5.** Histogram of outputs according to Ranson score (range, 0–9). The count represents the number of positive and negative cases according to score. With a threshold of 0.00, the sensitivity is 100% and the specificity is only 8%.



**Figure 6.** Histogram of outputs according to Balthazar grade (range, A–E). The count represents the number of positive and negative cases according to grade. For the threshold of grade A negative, the sensitivity is 97% and the specificity is 16%.

histogram output to institute therapy or diagnostic procedures for sicker patients.

The histograms for the Ranson scores and Balthazar grades show poor discrimination between positive and negative outcomes. The histogram for the ANN shows better performance, with improved discrimination between positive and negative cases. At a sensitivity of 100%, the ANN had a specificity of 29% and a positive predictive value of 43% (Table 5). The specificity and positive predictive value were not improved substantially when the threshold was raised, at the cost of missing several more severe cases. In comparison, at the same 100% sensitivity, linear discriminant analysis had a specificity of 21% and a positive predictive value of 41%. At 100% sensitivity, by definition the negative predictive value is also 100% for both models.

Tables 5–7 show possible performances of the models as the threshold over the decision variable is raised in the smallest increments possible with respect to changes in sensitivity for that model. The performances across the three models are consistent with that shown by the ROC curves, with the ANN as the best, followed by the Ranson grade and then the Balthazar grade. At 100% sensitivity, the ANN would have identified 29% of the less severe cases, whereas the Ranson score would have identified only 8%. The Balthazar grade would not have allowed 100% sensitivity at all owing to a complete overlap in the scores all the way down to the lowest category. At 93% sensitivity (missing only two of the 30 severe cases),

**Table 5**  
ANN Performance at Several Thresholds

Sensitivity (%) <sup>*</sup>	Specificity (%) <sup>†</sup>	Positive Predictive Value (%)	Negative Predictive Value (%)	Threshold for ANN Outputs (%)
100 (0/30)	29 (18/62)	41 (30/74)	100 (18/18)	0.088
97 (1/30)	29 (18/62)	41 (29/73)	95 (18/19)	0.092
93 (2/30)	32 (20/62)	40 (28/70)	91 (20/22)	0.186

<sup>\*</sup>Numbers in parentheses are the missed positive outcomes.  
<sup>†</sup>Numbers in parentheses are the spared negative outcomes.

32% of the less severe cases would have been identified by the ANN, 24% by the Ranson score, and 18% by the Balthazar grade. The categorization of the Ranson and Balthazar systems further makes it more difficult to achieve particular desired operating points, as shown by the rapid decrease in sensitivity as the threshold is raised successively.

## DISCUSSION

The clinical course of acute pancreatitis may vary from a mild self-limiting disease to a life-threatening illness (1,2). Patients whose condition progresses to severe disease may require intensive therapy, which may include surgery, and such treatment may require admission to a specialized center. Determination of the likely severity of

**Table 6**  
Ranson Score Performance at Several Thresholds

Sensitivity (%) <sup>*</sup>	Specificity (%) <sup>†</sup>	Positive Predictive Value (%)	Negative Predictive Value (%)	Threshold for Ranson Score
100 (0/30)	8 (5/62)	34 (30/87)	100 (5/5)	0 is negative
93 (2/30)	24 (15/62)	37 (28/75)	88 (15/17)	1 is negative
67 (10/30)	53 (33/62)	41 (20/49)	77 (33/43)	2 is negative

<sup>\*</sup>Numbers in parentheses are the missed positive outcomes.

<sup>†</sup>Numbers in parentheses are the spared negative outcomes.

**Table 7**  
Balthazar Grade Performance at Several Thresholds

Sensitivity (%) <sup>*</sup>	Specificity (%) <sup>†</sup>	Positive Predictive Value (%)	Negative Predictive Value (%)	Threshold for Balthazar Grade
97 (1/30)	16 (10/62)	36 (29/81)	91 (10/11)	A is negative
93 (2/30)	18 (11/62)	35 (28/79)	85 (11/13)	B is negative
47 (14/30)	61 (38/62)	40 (16/40)	73 (38/52)	C is negative

<sup>\*</sup>Numbers in parentheses are the missed positive outcomes.

<sup>†</sup>Numbers in parentheses are the spared negative outcomes.

disease is, therefore, crucial for patients with acute pancreatitis. Because the clinical course in this condition is so variable, such a determination is typically very difficult. Previous clinical approaches to this problem have required the collection of a large amount of clinical and laboratory data to derive prognostic scores such as the APACHE (3) or Ranson scores (2). These systems reflect the systemic effects of acute pancreatitis. The prognostic system derived by Balthazar (5), which is based on findings at CT, requires fewer data points and is based on the local manifestations at CT of both the pancreas itself and the surrounding intraabdominal structures. It seems intuitive that a prognostic system that could combine information relating to both the systemic and local effects of pancreatitis would be optimal. Such a system, however, would require the collection and storage of a large number of data points, which might limit its practical use. Recent modifications to the scoring systems have explored approaches in which fewer data points are used (18). Agarwal and Pitchumoni (18) proposed a simplified score based on four criteria—the decrease in hematocrit, serum calcium level, partial pressure of oxygen, and fluid sequestrations—as a simple gauge, with results comparable to Ranson's early objective signs.

The ANN approach offers two potential advantages. First, the ability of the ANN to assign variable weights to the input parameters to predict outcomes (by comparison with known outcomes) during the training process means that the most important of the many clinical and radiologic parameters can be identified. Therefore, many input data that have little relevance to the outcome can potentially be discarded, thus simplifying the data collection process. Second, the ANN offers a computer-aided model that can easily combine the relevant clinical and radiologic inputs in an accessible way.

Several ANNs were evaluated in this study; however, an ANN composed of only six features performed statistically better than both the Ranson or Balthazar systems. Therefore, this ANN offers a way to identify serious cases on the basis of just six parameters, which are easy to obtain in routine practice. The six parameters (blood pressure, extent of inflammation, fluid aspiration, serum creatinine level, serum calcium level, and concurrent severe illness) are weighted toward clinical parameters with just one radiologic parameter—extent of inflammation—included. Options for the extent of inflammation on the original data sheet were none (graded as 0.0), anterior pararenal, mesenteric, lesser sac, or flank and/or pelvis (all graded equivalently as 0.50) and posterior pararenal or perirenal (graded higher as 0.75). Fluid located in the posterior pararenal or perirenal space was graded higher because this is considered an uncommon finding (19). It is interesting that, although pancreatic glandular necrosis has long been considered a very important radiologic prognostic sign (5,20,21), this input did not improve the performance of the ANN. This is likely because substantial necrosis (ie, >30%) will be seen in only a small number of consecutive patients with acute pancreatitis.

It should be noted that although only six features were required for the ANN, this was still a complex model involving 17 weight parameters and seven nonlinear thresholding functions. It would be difficult to extract diagnostic criteria from this model that radiologists would be able to use without reference to the model. This non-transparency is an unfortunate characteristic of this type of model. Fortunately, the model would be very straightforward to implement with any personal computer or personal digital assistant and would enable the evaluation of any new patient in a fraction of a second.

Determination of the outcome measure to be evaluated is crucial when constructing and evaluating an ANN. An ANN can be used to predict an outcome as either positive or negative and has been used effectively to determine

whether a breast lesion at mammography is benign or malignant (11). In the clinical situation of acute pancreatitis, choosing a single indicator of severe versus nonsevere disease is difficult because the spectrum of disease is very wide. When constructing the ANN, we considered several possible outcome measures that might help differentiate negative (not severe) from positive (severe) cases. Mortality from the disease is a poor outcome because relatively few patients will die even when severe disease is present. Similarly, surgery is relatively rarely performed if one considers all patients with acute pancreatitis. A potentially useful indicator of severe disease is total cost of the hospital admission; however, this indicator may be influenced by local economic factors that may fluctuate, thus limiting the usefulness of this approach. A further outcome indicator might be admission to an intensive care unit; however, this was an infrequent occurrence in our patient group and is very dependent on individual physician practice and bed availability. Length of hospital stay was finally chosen as the outcome parameter. Patients staying less than the mean number of days for the entire group were considered to have less severe disease, whereas patients staying longer than the mean were considered to have severe disease. One potential limitation might have been that a very sick patient might die very quickly after admission, particularly if the patient had been transferred from another institution where most of the hospital stay had occurred; however, this situation did not occur during our study.

A potential limitation of this study is that training and testing were performed with a single data set. With use of the round-robin approach, the potential negative effect of this technique is minimized as each patient's data set is removed from the "training pool" while that data set is being "tested." Ideally, our results should be validated with a larger study with distinct "training" and "test" data.

The ANN that we constructed was able to differentiate between patients with positive and patients with negative outcomes on the basis of the combination of a small number of clinical, laboratory, and radiologic parameters. In the setting of acute pancreatitis, it is important to maximize sensitivity so that patients with the potential for a serious outcome are identified early so that appropriate therapy may be instituted as soon as possible. Specificity must also be acceptably high so that patients who are not likely to progress to severe disease are not inappropriately subjected to unnecessary interventions or an excess of imaging procedures. The ANN in this study achieved a

sensitivity of 100% (all severe cases detected) with a corresponding specificity of 29%. Although specificity can certainly be improved (up to 53%), this improvement will inevitably be at the expense of a reduction in sensitivity (reduced to 67%) (Table 6).

It was found for this data set that the ANN performed only slightly better than the linear discriminant analysis, and this difference was not statistically significant. In general, when two models are not significantly different it is preferable to use the simpler one, in this case linear discriminant analysis. This tends to reduce the bias caused by overtraining, especially with a relatively small data set such as the one used herein. The focus of this study has been on the ANN results, however, because of several reasons. First, the number of input features (and, thus, the complexity of the overall model) was greatly reduced by using a simple linear discriminant analysis. Second, the ANN outperformed the linear discriminant analysis slightly, and although the difference was not significant over these cases, it may become more significant with a larger data set. Third, the ANN seemed to have a greater advantage over the linear discriminant analysis in the high sensitivity portion of the ROC curve, improving specificity from 21% to 29% for a given 100% sensitivity. Unfortunately, this improvement cannot be appreciated on the fitted ROC curves. It should be emphasized, however, that all of these observations depend on the data and that the data set in this study was relatively limited in size.

As radiology departments are increasingly moving toward computerized picture archiving and communication systems, many are becoming more closely linked to hospital data information systems. In this environment, the merging of radiologic and clinical and/or laboratory data for a given patient has become simplified. Therefore, an ANN, particularly one that is effective on the basis of a small number of input data, could be easily applied to aid clinicians in identifying patients who may have a severe episode of pancreatitis. Such information may be valuable as new treatment options are developed for this condition.

#### REFERENCES

1. Bradley EL. A clinically based classification system for acute pancreatitis. *Arch Surg* 1993; 128:586-589.
2. Ranson JHC, Rifkind KM, Roses DF, Fink SD, Eng K, Spencer FC. Prognostic signs and the role of operative management in acute pancreatitis. *Surg Gynecol Obstet* 1974; 139:69-81.
3. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med* 1985; 12:818-827.

4. Larvin M, McMahon MJ. APACHE-II score for assessment and monitoring of acute pancreatitis. *Lancet* 1989; 2:201-205.
5. Balthazar EJ, Ranson JHC, Naidich DP, Meigibow AJ, Caccavale R, Cooper MM. Acute pancreatitis: prognostic value of CT. *Radiology* 1985; 156:767-772.
6. Balthazar EJ, Robinson DL, Megibow AJ, Ranson JHC. Acute pancreatitis: value of CT in establishing prognosis. *Radiology* 1990; 174:331-336.
7. Tourassi GD, Markey MK, Lo JY, Floyd CE Jr. A neural network approach to breast cancer diagnosis as a constraint satisfaction problem. *Med Phys* 2001; 28:804-811.
8. Freed KS, Lo JY, Baker JA, et al. Predictive model for the diagnosis of intraabdominal abscess. *Acad Radiol* 1998; 5:473-479.
9. Tourassi GD, Floyd CE, Sostman HD, Coleman RE. Acute pulmonary embolism: artificial neural network approach for diagnosis. *Radiology* 1993; 189:555-558.
10. Chan HP, Sahiner B, Petrick N, et al. Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network. *Phys Med Biol* 1997; 42: 549-567.
11. Baker JA, Kornguth PJ, Lo JY, Williford ME, Floyd CE. Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon. *Radiology* 1995; 196:817-822.
12. Fujita H, Katafuchi T, Uehara T, Nishimura T. Application of artificial neural network to computer-aided diagnosis of coronary artery disease in myocardial SPECT bull's-eye images. *J Nucl Med* 1992; 33: 272-276.
13. Ashizawa K, Ishida T, MacMahon H, Vyborny CJ, Katsuragawa S, Doi K. Artificial neural networks in chest radiography: application to the differential diagnosis of interstitial lung disease. *Acad Radiol* 1999; 6:2-9.
14. Sharma S. *Applied multivariate techniques*. New York, NY: Wiley, 1996.
15. Schwarzer G, Vach W, Schumacher M. On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Stat Med* 2000; 19:541-561.
16. Sargent DJ. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. *Cancer* 2001; 91: 1636-1642.
17. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest Radiol* 1989; 24:234-245.
18. Agarwal N, Pitchumoni CS. Simplified prognostic criteria in acute pancreatitis. *Pancreas* 1986; 1:69-73.
19. Raptopoulos V, Kleinman PK, Marks S, Snyder M, Silverman PM. Renal fascial pathway: posterior extension of pancreatic effusions within the anterior pararenal space. *Radiology* 1986; 158:367-374.
20. Bradley EL, Allen K. A prospective longitudinal study of observation versus surgical intervention in the management of necrotizing pancreatitis. *Am J Surg* 1991; 161:19-25.
21. Berger HG, Krautzberger W, Bittner R, Block S, Buchler M. Results of surgical treatment of necrotizing pancreatitis. *World J Surg* 1985; 9:972-979.