Strategic Use of Individual and Combined Enzyme Indicators for Acute Pancreatitis Analyzed by Receiver-Operator Characteristics

Mario Werner,1 William M. Steinberg,2 and Carol Pauley1

The optimal strategy for the diagnosis of acute pancreatitis with enzyme assay results as indicators was evaluated in 67 emergency cases in whom this condition was suspected. We measured urine amylase expressed as activity concentration (U/L), timed excretion (U/h), and amylase/creatinine clearance ratio, and also serum amylase, elastase, lipase, and trypsinogen, at admission and repeatedly during hospitalization. The receiver-operator characteristic function was used to evaluate the diagnostic discrimination of each variable among initial findings and among the highest individual findings established retrospectively. We applied the same treatment to multiple univariate discrimination, using the six possible pairs and the four possible triplets of serum indicators. The results suggest that such urine assays should be abandoned, that all individual serum assays combine about 0.9 sensitivity with 0.9 specificity, that pairing of two assays does not clearly enhance discrimination, and that triplets of tests may degrade discrimination. The trade-off between sensitivity and specificity is a function not only of the chosen decision threshold but also of the sampling strategy (initial vs highest values) and of the interpretation rule (Boolean "and" vs "or" strategy).

Additional Keyphrases: amylase · lipase · trypsinogen · amylase/creatinine clearance ratio · multivariate discrimination · decision-threshold values · cutoff values

The armamentarium of enzyme indicators for acute pancreatitis continues to improve, both in quality and number. Not only have modifications substantially enhanced the reliability of such standbys as amylase and lipase assay (1), but newer markers such as trypsinogen and elastase assay find expanding clinical use (2–4). Clearly, such a choice of indicators imposes the question of what constitutes the optimal diagnostic strategy for acute pancreatitis. Unfortunately, conclusive answers are lacking, because the varied experimental conditions under which individual indicators have been assessed preclude comparisons of outcome.

More fundamentally, to establish the optimal diagnostic strategy, a valid trial must duplicate as closely as possible the real-life circumstances in which these indicators of pancreatitis are used (5). In this respect, most evaluations appear deficient in at least one of four ways: First, normal limits must not be based on the exclusive study of healthy subjects (6–8). Rather, decision thresholds should be derived from a mixed sample, representing the patient population in whom the tests are used. Second, diagnostic discrimination must not be evaluated based on peak values selected retrospectively from serial measurements. Rather, the initial findings confronting the physician forced to diagnose an acute condition should be used. Third, the prevalence of acute pancreatitis in the tested population may not have been considered (9). Fourth, results from more than one indicator may not have been combined to evaluate the odds of a correct diagnosis in the face of conflicting findings (3, 6–8).

Here, we compare the trade-offs imposed by actual clinical circumstances between sensitivity and specificity of amylase in urine and four enzymes in serum: amylase (EC 3.2.1.1), elastase (EC 3.4.21.11), lipase (EC 3.1.1.3), and trypsinogen (EC 3.4.21.4). We apply receiver-operator characteristic (ROC) functions to assess the effects of choosing different decision thresholds for individual indicators, and the effects of combining two or more indicators in evaluating a mixed emergency-room population in whom the diagnosis of acute pancreatitis was entertained. In basing diagnostic decisions on multiple indicators, we use multiple univariate rather than true multivariate discrimination because univariate cutoff values can be retained unchanged, as continues to be the clinician's approach, and because discrimination potentially can be optimized without the use of complex decision techniques.

Materials and Methods

Patients. We investigated 33 patients with acute pancreatitis and 34 patients in whom this condition was suspected but ruled out. All patients were hospitalized as emergencies at the 500-bed George Washington University Medical Center, which serves a mixed metropolitan population. The diagnostic classification was established after each patient's discharge by review of all clinical findings exclusive of the investigated enzyme indicators. The exception to this rule was that patients in whom both serum amylase and lipase exceeded twice the upper-normal limit were automatically classified as having acute pancreatitis according to previously established high specificity for such values (2). Other cases were confirmed by sonogram, computerized tomography, laparotomy, or aggregate clinical presentation and hospital course. Table 1 lists the patients by disease state and sex. Patients with calcific chronic pancreatitis were excluded.

Samples. Venous blood was collected into evacuated tubes (Corvac®; Corning Glass Works, Corning, NY 14830), and 2-h timed urine specimens were also obtained on the day of admission (day 1). Subsequent samples were collected once daily for various intervals (1 to 10 days). We measured amylase (Amylase DS; Beckman Instruments, Inc., Carlsbad, CA 92028, performed at 37 °C in the Technicon RA 1000, Technicon Instruments Corp., Tarrytown, NY 10591), elastase (radioimmunoassay; Abbott Laboratories, North Chicago, IL 60064), lipase (manual turbidimetric assay with colipase, performed at 37 °C; Boehringer Mannheim Diagnostics, Indianapolis, IN 46250), and trypsinogen (CISTRIPIK; Cambridge Diagnostics, Needham Heights, MA 01865). Amylase in urine and serum was

1 Division of Laboratory Medicine, Department of Pathology, and 2 Division of Gastroenterology, Department of Medicine, The George Washington University Medical Center, 901 Twenty-Third Street, N.W., Washington, DC 20037.

Received January 30, 1989; accepted March 10, 1989.
Table 1. Characteristics of Investigated Subjects

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Total no.</th>
<th>No. of males</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute pancreatitis</td>
<td>33</td>
<td>18</td>
</tr>
<tr>
<td>Alcoholism</td>
<td>19</td>
<td>13</td>
</tr>
<tr>
<td>Idiopathy</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Gallstone</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Endoscopic retrograde cholangiopancreatography</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Systemic lupus erythematosus</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>Controls</strong></td>
<td><strong>34</strong></td>
<td><strong>15</strong></td>
</tr>
<tr>
<td>Peptic ulcer disease</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Cholecystitis</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Small bowel obstruction</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Appendicitis</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Gastritis</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Various other</td>
<td>9</td>
<td>4</td>
</tr>
</tbody>
</table>

Multivariate "and" sensitivity = product of univariate true-positive rates
Multivariate "and" specificity = 1.00 - product of univariate false-positive rates
Multivariate "or" sensitivity = 1.00 - product of univariate false-negative rates
Multivariate "or" specificity = product of univariate true-negative rates

Results

Individual Discriminators

The diagnostic performance of each urine and serum enzyme indicator was studied by two ROC functions, one derived from initial findings obtained on the first day of hospitalization, the other from the highest findings retrospectively identified among the serial measurements of each individual (Figures 1 to 7). At any given threshold, using the highest individual findings favored sensitivity at the expense of specificity, whereas the converse was true for initial findings.

Urine amylase was evaluated in three ways: first, in terms of activity concentration; second, as activity excreted per hour (13); and third, as amylase/creatinine clearance ratio (14). Activity per volume identified acute pancreatitis with a success rate about halfway between perfect and random discrimination, whether initial or highest individual values were used (Figure 1). Normalizing urinary amylase to activity excreted per hour barely improved discrimination (Figure 2). The still more complex normalization demanded by the amylase/creatinine clearance ratio actually so worsened discrimination that it became almost random (Figure 3).

Assays of serum provided markedly better discrimination. For serum amylase, the highest individual findings yielded a sensitivity of 0.91 and a specificity of 0.94 at a threshold of 200 U/L (Figure 4). At the same threshold, initial findings yielded slightly poorer discrimination, because sensitivity was lost. There were further differences between highest and initial findings. Among highest individual findings, false-positive outliers widened the zone of data overlap between patient and control populations, compromising sensitivity at high decision thresholds. Thus, thresholds greater than 200 U/L traded only small gains in specificity for larger losses of sensitivity. Conversely, among initial findings, false-negative outliers had

assayed by the same method. Urine and serum creatinine were measured with alkaline picrate (Jaffé reaction) without use of Lloyd's reagent (10). During the study period, day-to-day coefficients of variation calculated for monthly control periods never exceeded 7% for serum creatinine (mean 10 mg/L), 3% for urine creatinine (mean 1.15 g/L), 3% for amylase (mean 165 U/L), 10% for elastase (mean 285 μg/L), 6% for lipase (mean 385 U/L), and 10% for trypsinogen (mean 25 μg/L). The reference interval in clinical use is <25 U/24 h for urine amylase, and 800-2000 mg/24 h for urine creatinine. The upper limits of the reference intervals in clinical use for serum are: 150 U/L for amylase, 338 μg/L for elastase, 208 U/L for lipase, 38 μg/L for trypsinogen, and 14 mg/L for creatinine.

Data analysis. As previously defined, "sensitivity" is the probability of a positive test among patients with acute pancreatitis, and "specificity" is the probability of a negative test among patients without acute pancreatitis (11):

Sensitivity = true positives/(true positives + false negatives)

Specificity = true negatives/(true negatives + false positives)

Sensitivities and specificities were plotted as functions of each other on ROC grids in which the upper right-hand corner represents perfect discrimination. To construct the ROC function (12), findings from patients with and without acute pancreatitis were first separated. Second, a number of decision thresholds were applied sequentially within each group to distinguish positives and negatives. Third, for each threshold the test sensitivity was derived from acute pancreatitis patients and the test specificity from patients without acute pancreatitis.

Multivariate positives were determined either by demanding that all findings be positive (Boolean intersection, the "and" strategy), or by accepting just one positive finding (Boolean union, the "or" strategy) (11). The former rule should reduce the effect of sporadic false positives occurring with a single discriminator, and so favor specificity at the expense of sensitivity. The latter rule should reduce the effect of sporadic false negatives occurring with a single discriminator, and so favor sensitivity at the expense of specificity. The theoretical discriminatory properties of multiple test combinations were calculated as follows:

![Image](https://via.placeholder.com/150)

![Image](https://via.placeholder.com/150)

Fig. 1. ROC functions in the diagnosis of acute pancreatitis for urine amylase, expressed as U/L (activity concentration), at admission (initial findings) and among the highest individual findings retrospectively identified.

In Figs. 1-7, threshold values (U/L, μg/L, or %) producing given combinations of sensitivity and specificity are indicated.
The opposite effect, compromising specificity at low decision thresholds. Thus, for initial findings, thresholds smaller than 200 U/L traded only small gains in sensitivity for larger losses of specificity.

The highest individual findings for each serum elastase (Figure 5), lipase (Figure 6), and trypsinogen (Figure 7) combined at some discriminatory threshold sensitivity of 0.91 or better with specificity of 0.87 or better. At those same thresholds, initial findings combined sensitivity of 0.82 or better with specificity of 0.93 or better. Just as for initial amylase findings, false-negative outliers caused asymmetry in the ROC curves of elastase and lipase, but this effect was seen about equally among initial and highest individual values.

Pairs of Discriminators

In an attempt to neutralize the effects of outliers occurring with single discriminators and to enhance diagnostic classification, pairs of discriminators were combined to distinguish positives and negatives with respect to acute pancreatitis. To implement the classification based on two variables, those round numbers that best balanced sensitivity and specificity of the highest individual findings were chosen as decision thresholds for each indicator. Using threshold values of 200 U/L for amylase, 200 U/L for lipase, 300 μg/L for elastase, and 50 μg/L for trypsinogen, we compared the ROC values of the seven individual indicators with those of the six indicator pairs possible with the four serum assays (Figure 8). Because of their poorer diagnostic performance, assays of urine were omitted from the multivariate analysis.

Pairs of both the highest individual findings and the initial findings produced the expected discrimination characteristics. For the highest values, the "and" option improved specificity to the maximum of 1.0 while reducing sensitivity (0.91 or less), whereas the "or" option improved...
Specifity

Highest

Specifity

INITIAL FINDINGS

HIGHEST FINDINGS

Fig. 8. ROC values in the diagnosis of acute pancreatitis for urine amylase, expressed in three ways (dotted circles); for serum amylase, elastase, lipase, and trypsinogen (closed circles); and for pairs of the serum indicators used according to the "and" strategy (dotted squares) as well as the "or" strategy (dotted diamonds) at (left) admission (initial findings) and (right) among the highest individual findings retrospectively identified. Note truncated scales defining the grid.

sensitivity (0.97 or better) while diminishing specificity (0.91 or less). As was true for individual values, pairs of initial values overall favored specificity at the expense of sensitivity. This effect enhanced the bias of the "and" option for specificity, decreasing sensitivity (0.85 or less) while increasing specificity to the maximum of 1.00. Conversely, this effect partly offset the bias of the "or" option for specificity. When compared with the discrimination obtained with highest findings, the balance of these opposing forces decreased sensitivity minimally (0.94 or better) while increasing specificity uniformly (0.87 or better).

In each of the four situations arising from applying the "and" or the "or" option to initial or to highest individual findings, different pairs of indicators provided the largest sums of sensitivity and specificity and no winner pair emerged. Still, in all circumstances the best pair included lipase. With the "and" strategy, lipase–trypsinogen performed best for initial values, while elastase–lipase performed best for highest values. With the "or" strategy, elastase–lipase performed best for initial values, while lipase–trypsinogen performed best for highest values. In all four situations amylase–trypsinogen performed worst.

Triplets of Discriminators

Four different indicator triplets could be formed with the serum assays (Table 2). For the highest individual findings, all "and" triplets, just as expected, had less sensitivity than obtained with "and" pairs, whereas all "or" triplets had less specificity. For initial findings, these reductions were less clear-cut. The discriminatory properties of "and" as well as "or" triplets roughly fell into the domains also occupied by component pairs. The combination lipase, elastase, and trypsinogen provided the greatest sum of sensitivity and specificity in all four conditions tested, but this triplet never outperformed the most successful pair of its component indicators.

Table 2 further compares the actually observed discriminatory performance of triplets with the theoretical values calculated from the combined performance of individual discriminators according to the formulas given at the end of the section on Data analysis. For both the "and" as well as the "or" strategy, observed and calculated specificity values matched closely. For sensitivity values, on the other hand, discrepancies appeared. These were small in the case of the "or" strategy, where observed sensitivities did not achieve the calculated values, implying the presence of more than expected pancreatitis patients with all three tests negative. Larger discrepancies occurred in the case of the "and" strategy, where observed sensitivities exceeded the calculated values, implying the presence of fewer than expected pancreatitis patients with but one or two positive tests. Consequently, more pancreatitis patients than expected had either all tests positive or negative, while fewer than expected combined conflicting positive and negative findings.

Discussion

Our trial was designed to assess test performance in acute pancreatitis realistically, in order to define a practical diagnostic strategy. Specifically, we derived decision thresholds from a patient population in whom the tests were clinically required, analyzed the test results available when the diagnosis must first be made, and evaluated the odds of a correct diagnosis when indicators are combined. However, in estimating outcome probabilities from our data base, our purpose was not to establish absolute measures of test capability, but only the relative measures obtained with different strategies, in order to define the best diagnostic approach among several. Thus, we do not propose that the probabilities we found be used to assess

<table>
<thead>
<tr>
<th>Discriminator</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>O</td>
<td>C</td>
</tr>
<tr>
<td>&quot;and&quot; strategy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amylase, elastase, lipase</td>
<td>0.75</td>
<td>0.68</td>
</tr>
<tr>
<td>Amylase, lipase, trypsinogen</td>
<td>0.76</td>
<td>0.68</td>
</tr>
<tr>
<td>Elastase, lipase, trypsinogen</td>
<td>0.81</td>
<td>0.75</td>
</tr>
<tr>
<td>Amylase, elastase, trypsinogen</td>
<td>0.75</td>
<td>0.68</td>
</tr>
<tr>
<td>&quot;or&quot; strategy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amylase, elastase, lipase</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>Amylase, lipase, trypsinogen</td>
<td>0.94</td>
<td>1.00</td>
</tr>
<tr>
<td>Elastase, lipase, trypsinogen</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>Amylase, elastase, trypsinogen</td>
<td>0.97</td>
<td>1.00</td>
</tr>
</tbody>
</table>

* See Methods for calculation of theoretical discriminatory properties.
the likelihood of a correct diagnosis, but it is unlikely that the reported sensitivities and specificities are overly optimistic. Any error in patient classification should have caused us to underdetermine these values.

ROC functions, the tool used to optimize discrimination, focus on the data interval where findings from controls and patients overlap (12). They provide a graphic representation of the impact of false-positive and false-negative outliers on the trade-off between sensitivity and specificity. The characteristics of this interchange distinguish different tests, say elastase from trypsinogen, and in some cases even different ways of using the same test (Figure 4). Clearly, such peculiarities would not be recognized if healthy subjects served as controls. ROC functions do not assess data distribution in the additional domains where only findings from controls or from patients are encountered, but it is still possible to infer the relative data densities in these zones from the sensitivity value at which specificity reaches unity, and vice versa.

We have not presented the effect of a varying prevalence of acute pancreatitis in the tested population. Should this variable differ from the "neutral" 50% we chose, the relevant posterior probabilities (i.e., "predictive values") can readily be calculated from Bayes Theorem if sensitivity and specificity are known (11). Nevertheless, some comments about test effectiveness (i.e., the sum of the population proportions of "true positives" and "true negatives") are in order (11). For highest individual values, effectiveness of single discriminators is independent of prevalence, because we chose decision thresholds to balance specificity and sensitivity. For univariate initial values, or when the "and" strategy is applied to multivariate values, effectiveness increases with decreasing prevalence. The "or" strategy applied to multivariate values increases the effectiveness of highest individual values with prevalence and renders effectiveness independent of prevalence for initial values.

Several practical conclusions emerge from our analysis. First, assay of amylase in urine provides inadequate diagnostic discrimination of acute pancreatitis, improved neither by repeated measurement nor by data manipulation that attempts to "normalize" findings. Use of all these tests for this purpose should be abandoned, with the possible exception of such unusual circumstances as when the pancreatic duct is anastomosed to the urinary bladder. Second, contrary to intuitive expectation, serial enzyme measurements repeated over time do not provide better discrimination of acute pancreatitis than values obtained at admission. However, initial values have different optimal decision thresholds than highest values in a series of assays. A single cutoff value initially favors specificity and subsequently sensitivity. Third, pairing two assays does not clearly enhance discrimination of acute pancreatitis. If the same thresholds are used as in the evaluation of individual variables, sensitivity is simply traded for specificity or vice versa, depending on whether the "and" or the "or" interpretation rule is applied. Fourth, using triplets of simultaneous tests has a high likelihood of degrading diagnostic discrimination of acute pancreatitis, depending on the decision rule applied. Fifth, when pairs of assays are repeated over time, two interacting influences shape diagnostic discrimination of acute pancreatitis: on the one hand, initial values favor specificity, whereas subsequent values favor sensitivity; on the other hand, the "or" rule favors sensitivity, the "and" rule specificity. Given these twin dilemmas, the rational choice of a decision threshold must take into consideration the circumstances under which the test will usually be used.

To test the broader validity of our conclusions, we compared our findings with multiple univariate discrimination with a similar study utilizing true multivariate discrimination based on discriminant analysis of enzyme values for the diagnosis of acute myocardial infarction (15). As is the case in acute pancreatitis, several indicators of acute myocardial infarction—such as results for creatine kinase, aspartate aminotransferase, and lactate dehydrogenase—provide about equal specificities (0.9 or better) and sensitivities (0.7 or better). Pairing of indicators enhances discrimination of infarction only marginally on the first day of hospitalization. Repeating a single or a pair of assays on the second day improves sensitivity somewhat, but either adding assays on subsequent days or adding a third discriminator offers no further advantage. Analogous strategic inferences thus appear pertinent to acute pancreatitis and acute myocardial infarction, as well as to multiple univariate and true multivariate strategies: namely, choice of the proper decision threshold improves the diagnostic odds more than proliferating testing. In acute pancreatitis, selection of the cutoff value is intrinsically facilitated, because serum discriminators combine good specificity with good sensitivity, whereas in acute myocardial infarction, the typical discriminator pairs good specificity with only mediocre sensitivity.

We thank Glenn P. Gwоздz, M.D., and James P. Henry, B.S., for clinical support.

References