The reliability of EUS for the diagnosis of chronic pancreatitis: interobserver agreement among experienced endosonographers

Michael B. Wallace, MD, MPH, Robert H. Hawes, MD, Valerie Durkalski, MPH, Amitabh Chak, MD, Shawn Mallory, MD, Marc F. Catalano, MD, Maurits J. Wiersema, MD, Manoop S. Bhutani, MD, Donato Ciaccia, MD, Michael L. Kochman, MD, Frank G. Gress, MD, Annette van Velse, LPN, Brenda J. Hoffman, MD

Charleston, South Carolina, Cleveland, Ohio, Minneapolis, Minnesota, Milwaukee, Wisconsin, Rochester, Minnesota, Gainesville, Florida, Indianapolis, Indiana, Philadelphia, Pennsylvania, and Mineola, New York

Background: Endoscopic ultrasound (EUS) is a minimally invasive, low risk method of diagnosis for chronic pancreatitis (CP). The degree to which endosonographers agree on the features and diagnosis of CP is unknown. For EUS to be considered an accurate test for CP, there must be good interobserver agreement.

Methods: Forty-five pancreatic EUS examinations were videotaped by 3 experienced endosonographers. Examinations from 33 patients with suspected CP based on typical symptoms, as well as 12 control patients without suspected CP, were included. Eleven experienced endosonographers (“experts”) who were blinded to clinical information independently evaluated all videotaped examinations for the presence of CP and the following 9 validated features of CP: echogenic foci, strands, lobularity, cysts, stones, duct dilatation, duct irregularity, hyperechoic duct margins, and visible side branches. The experts also ranked (most to least) which features they believed to be the most indicative of CP. Interobserver agreement was expressed as the kappa (κ) statistic.

Results: There was moderately good overall agreement for the final diagnosis of CP (κ = 0.45). Agreement was good for individual features of duct dilatation (κ = 0.6) and lobularity (κ = 0.51) but poor for the other 7 features (κ < 0.4). The expert panel had consensus or near consensus agreement (greater than 90%) on 206 of 450 (46%) individual EUS features including 22 of 45 diagnoses of CP. Agreement on the final diagnosis of CP was moderately good for those trained in third tier fellowships (κ = 0.42 ± 0.03) and those with more than 1100 lifetime pancreatic EUS examinations (κ = 0.46 ± 0.05). The presence of stones was regarded as the most predictive feature of CP by all endosonographers, followed by visible side branches, cysts, lobularity, irregular main pancreatic duct, hyperechoic foci, hyperechoic strands, main pancreatic duct dilatation, and main duct hyperechoic margins. The most common diagnostic criterion for the diagnosis of CP was the total number of features (median 4 or greater, range 3 or greater to 5 or greater).

Conclusions: EUS is a reliable method for the diagnosis of chronic pancreatitis with good interobserver agreement among experienced endosonographers. Agreement on the EUS diagnosis of chronic pancreatitis is comparable to other commonly used endoscopic procedures such as bleeding ulcer stigmata and computed tomography of the brain for stroke localization and better than the physical diagnosis of heart sounds. (Gastrointest Endosc 2001;53:294-9.)

EUS was developed during the early 1980s specifically for the purpose of imaging the pancreas. Changes in the pancreatic parenchyma and ducts can be recognized by transabdominal ultrasound (TUS) in patients with chronic pancreatitis (CP). US features of CP were incorporated into the standardized grading of CP established in Cambridge, England, known as the Cambridge classification.

When EUS was applied in the evaluation of CP several important observations were made. EUS had significant technical advantages over TUS: the lack of overlying bowel gas and the higher resolution images obtained with high frequency US transducers. These high quality images led to the detection of several new features of CP not previously
observed by TUS or CT. These include hyperechoic margins of the main pancreatic duct (MPD), lobularity of the parenchyma, small cystic changes in the parenchyma, and side branch duct ectasia. It was also found that EUS has significant advantages in terms of safety in reference to endoscopic retrograde pancreatography (ERP). By comparison, the risk of inducing acute pancreatitis by EUS is low. Only a single case of acute pancreatitis due to diagnostic EUS has been reported.9 Overall the rate of serious complications associated with diagnostic EUS is 1 per 2000 procedures, and for EUS with fine-needle aspiration, 1 per 100 procedures.10

It is difficult to determine the accuracy of EUS in the diagnosis of CP because of the lack of a widely accepted standard. Several studies have compared EUS with ERP,11-16 pancreatic function testing,12,14 and histology.17-21 These single center studies suggest that EUS findings correlate well with other common measures of CP, although each study used slightly different EUS criteria for the diagnosis. The International Working Group for Minimal Standard Terminology (MST) in Gastrointestinal Endosonography has recently developed MST for the description of CP changes detected by EUS.22

The aim of this study was to evaluate the degree of agreement among 11 endosonographers by using MST definitions for the diagnosis of CP by EUS and the influence of training and experience on interobserver agreement.

METHODS

Forty-five patients undergoing upper GI EUS had their examination videotaped by 1 of 3 experienced (more than 600 EUS procedures) endosonographers (M.B.W., B.J.H., R.H.H.). These included 33 patients undergoing EUS for abdominal pain of suspected pancreatic origin and 12 control patients undergoing EUS for other indications (esophageal cancer 4, submucosal tumor 4, common bile duct stricture 3, and gastric cancer 1). Each videotaped segment included complete inspection of the MPD and pancreatic parenchyma in the body and tail of the pancreas. Although the head of the pancreas was examined during the procedure, there is inconsistency in the interpretation of features that are seen only in the head. Tapes ranged from 45 seconds to 2 minutes in length. Caliper measures of the MPD diameter in the body of the pancreas were provided in all but 5 cases. No other annotations such as labeling, arrows, or other markers were used on the videotapes. The master videotape was recorded by using high-resolution recording equipment (Betacam SP; Sony Corp., Tokyo, Japan). Each of the 11 endoscopists, including the 3 who videotaped the examinations, was provided with a first generation copy made on VHS tape. The 3 original endosonographers who taped the examinations viewed the videotapes at least 3 months apart from the original examination. All endoscopists, including the original 3 who taped the procedure, were blinded to the clinical information and the results of the original diagnostic evaluation.

Each of the 11 endoscopists was provided with MST definitions of terms22 and instructions to score each examination for the presence or absence of 9 features of CP (hyperechoic foci, hyperechoic strands, cysts, lobularity, MPD duct dilatation, visible side branches, hyperechoic MPD margins, and irregular MPD contour) and was asked to make a final diagnosis of CP (yes/no). MPD dilatation was defined as a width greater than 2 mm in the body or greater than 1 mm in the tail according to standards set for normal persons by Wiersema et al.12 Endoscopists were allowed to view and review examinations as much as desired. The 3 endosonographers who videotaped the examinations viewed the taped examinations at least 3 months after the original procedure and were blinded to the identity of each patient.

Each endoscopist also completed a survey regarding their training, experience, and beliefs about EUS for the diagnosis of CP. Training variables included the endoscopist’s primary mentor and type of training (third tier, during general fellowship or postfellowship tutorial). Measures of experience included total lifetime and annual case volume for EUS overall, EUS of the pancreas, and EUS for CP. Beliefs about EUS included how strongly the endoscopist believed that EUS accurately detects CP (5 point scale from strongly believe to strongly don’t believe), what criteria they used to diagnose CP (i.e., the number of features), and the relative ranking of importance of each of the 9 features.

Each score sheet was completed by the endoscopist and consisted of 45 examinations and 10 columns (9 individual features and final diagnosis), each scored 0 (if absent) or 1 (if present). All statistical analyses were performed with statistical software (SAS V. 6.12; SAS Institute, Cary, N.C.). Interobserver agreement and qualitative description (e.g., good or moderately good) were determined by computing the kappa statistic for multiple observers of binary (0/1) data as described by Landis and Koch23 and Fleiss.24 A two-sided p value of 0.05 or less was considered significant.
RESULTS

Eleven endosonographers completed evaluation of videotaped EUS examinations of the pancreas from 45 patients. All 11 endosonographers were highly experienced with EUS and had performed a median of 175 (range 90 to 600) pancreatic EUS examinations per year and 1100 (range 600 to greater than 5000) lifetime cases. Eight of the 11 had completed third tier fellowships in EUS.

Examinations of the pancreas included 33 patients undergoing EUS for suspected pancreatic disease and 12 control patients undergoing EUS for nonpancreatic indications. There was a consensus or near consensus (greater than 90% of endoscopists agreeing) on the EUS diagnosis of CP in 22 of 45 (49%) of cases (14 diagnosed as CP and 6 as normal) and greater than 70% agreement on 33 of 45 (73%) cases. On 12 of 45 examinations there was no consensus regarding the diagnosis of CP (less than 70% agreement). The overall agreement for the diagnosis of CP was moderately good ($\kappa = 0.45 \pm 0.02$). The percentage of cases diagnosed as CP by EUS by each endosonographer is shown in Figure 1.

The degree of agreement on individual EUS features of CP was variable. Agreement was best for MPD dilatation and parenchymal lobularity, borderline for stones and hyperechoic duct margins, and poor for the remaining 5 features (Table 1). Examples of consensus normal, consensus abnormal, and nonconsensus cases are shown in Figure 2.

All 11 endosonographers reported using the total number of EUS features as their criterion for the diagnosis of CP, although the threshold varied. Three endosonographers used 3 or more features, four used 4 or more features, and one used 5 or more features to diagnose CP. Three endosonographers used variable thresholds (3 or more to 5 or more) depending on the age and the clinical history of the patient.

Some EUS features were considered more predictive of CP than others. The presence of stones was ranked number 1 by all 11 endosonographers. Other features are shown in Table 2. Five endosonographers strongly believed and 6 somewhat believed that EUS accurately detects CP.

The type of training and experience of the endoscopist did not significantly affect the degree of agreement. Agreement on the diagnosis of CP was similar among the 3 endosonographers who were self-trained ($\kappa = 0.55 \pm 0.09$) and moderately good among the 8 trained in third tier fellowships ($\kappa = 0.42 \pm 0.03, p = 0.21$). Agreement on diagnosis of CP was similar for those who performed more than 1100 total pancreatic EUS examinations ($\kappa = 0.46 \pm 0.05$) and those who performed less than 1100 pro-

Figure 2. Examples of cases diagnosed by (A) greater than 90% of endosonographers as normal, (B) no consensus on the diagnosis, and (C) greater than 90% of endosonographers as CP.
Table 1. Interobserver agreement as measured by the kappa statistic for the EUS diagnosis of CP and for individual EUS features of CP

<table>
<thead>
<tr>
<th>EUS feature</th>
<th>Kappa (± SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global diagnosis of CP</td>
<td>0.45 (0.02)</td>
</tr>
<tr>
<td>MPD dilatation</td>
<td>0.61 (0.02)</td>
</tr>
<tr>
<td>Lobularity</td>
<td>0.51 (0.02)</td>
</tr>
<tr>
<td>Stones</td>
<td>0.38 (0.02)</td>
</tr>
<tr>
<td>Hyperechoic MPD margins</td>
<td>0.36 (0.02)</td>
</tr>
<tr>
<td>Cysts</td>
<td>0.32 (0.02)</td>
</tr>
<tr>
<td>Hyperechoic strands</td>
<td>0.31 (0.02)</td>
</tr>
<tr>
<td>MPD irregularity</td>
<td>0.29 (0.02)</td>
</tr>
<tr>
<td>Hyperechoic foci</td>
<td>0.29 (0.02)</td>
</tr>
<tr>
<td>Visible side branches</td>
<td>0.18 (0.02)</td>
</tr>
</tbody>
</table>

Table 2. Ranking (1 = most, 9 = least) of importance of individual EUS features for predicting CP

<table>
<thead>
<tr>
<th>EUS feature</th>
<th>Rank (interquartile range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stone</td>
<td>1 (1-1)</td>
</tr>
<tr>
<td>Visible side branches</td>
<td>4 (3-5.5)</td>
</tr>
<tr>
<td>Lobularity</td>
<td>4 (2.5-7.5)</td>
</tr>
<tr>
<td>Cysts</td>
<td>5 (4-6)</td>
</tr>
<tr>
<td>MPD irregularity</td>
<td>6 (4.5-8)</td>
</tr>
<tr>
<td>Hyperechoic foci</td>
<td>6 (4-8)</td>
</tr>
<tr>
<td>Hyperechoic strands</td>
<td>6 (5-7)</td>
</tr>
<tr>
<td>Hyperechoic MPD margins</td>
<td>6 (4.5-8)</td>
</tr>
<tr>
<td>MPD dilatation</td>
<td>7 (4.5-8)</td>
</tr>
</tbody>
</table>

The number shown represents the median (interquartile range) rank by all 11 endosonographers.

cedures (κ = 0.46 ± 0.04, p = 0.87). Agreement on the diagnosis of CP was also similar for the 3 endosonographers who performed the original examinations (κ = 0.54 ± 0.09) compared with the 8 endosonographers who did not perform the original examinations (κ = 0.41 ± 0.03, p = 0.36).

DISCUSSION

This study demonstrates that there is moderately good interobserver agreement among experienced endosonographers in the EUS diagnosis of chronic pancreatitis based on review of short preselected videotaped examinations. Among these highly experienced endosonographers, the type of training or the volume of experience of the endosonographer did not significantly affect agreement. Only 2 of the 9 individual features had good levels of agreement (κ ≥ 0.40). There were minor variations in the criteria used to diagnose chronic pancreatitis, although all endosonographers diagnosed CP when 5 or more EUS features were present. Endosonographers may have also considered age, clinical history, and the relative importance of individual features when making a final diagnosis. Interobserver reliability for the EUS diagnosis of chronic pancreatitis is comparable to other commonly used endoscopic procedures such as bleeding ulcer stigmata (κ = 0.34 to 0.66)25 and radiologic procedures such as brain CT for stroke localization (κ = 0.56 to 0.62).26 Reliability is better for EUS than for the physical diagnosis of heart sounds (κ = 0.05 to 0.18).27

A diagnostic test must possess both accuracy (the degree to which the diagnosis compares to a standard) and reliability (the degree to which the test is reproducible in the same clinical situation). When a reference standard does not exist, this is often measured as the degree to which practitioners agree on a diagnosis. Although histology is often considered a reference standard in medicine, there are no agreed on criteria for histologic changes of chronic pancreatitis, and those that are described are often focally distributed. In the absence of a reference standard, a clinical diagnosis of chronic pancreatitis often requires multiple tests and clinical judgment.

Wiersma et al.12 compared the degree of agreement among 3 experienced endosonographers at a single center for interpretation of individual features of CP. The agreement was 88% for hyperechoic foci, 94% for focal reduced echogenicity, 94% for lobularity, 83% for hyperechoic duct margins, and 94% for duct irregularity. To further improve the reliability, the International Working Group for Minimal Standard Terminology (MST) in Gastrointestinal Endosonography has published a set of MST definitions for many of the EUS features of CP.22

The accuracy of EUS compared to other diagnostic tests has been extensively evaluated. These tests, such as histology, pancreatography, and pancreatic function testing, are not universally considered to be standards for the diagnosis of CP. Histologic studies support the concept that patients with isolated EUS do have mild chronic pancreatic inflammation that may be the cause of their symptoms. In a study of 486 patients evaluated by a single experienced pancreatic surgeon, Walsh et al.28 identified 43 who had characteristic symptoms of pancreatic disease but normal or equivocal ERCP, CT, or US. Those whose symptoms failed to respond (16 patients) to medical therapy (enzyme replacement, low fat diet, and at least 3 trials of bowel rest with total parenteral nutrition) underwent pancreatic resection. The histologic appearance of the resected pancreata was that of subtle but distinct evidence of minimal change chronic pancreatitis. These changes were “focally” distributed throughout the gland and included lymphocytic cell infiltrates, intralobular and periductal fibrosis, and focal ductal dilatation with inspissated protein plugs. Nine of the 16 patients had complete or significant improvement in pain after total pancreatectomy, whereas 5 did not respond and 1 died of unrelated causes.

Early studies by Lees5 compared EUS morpholo-
ogy to histology after pancreatic resection. In 6 of 7 patients with EUS changes, a diagnosis of chronic pancreatitis was confirmed by histology.

Multiple studies have compared EUS to ERP in patients with abdominal pain and suspected chronic pancreatitis.\textsuperscript{11-16} Three\textsuperscript{12,14,15} used standardized EUS and ERP (Cambridge classification) grading systems. Each of these studies evaluated the pancreas for the presence of 9 to 11 features and then considered EUS to be abnormal if the total number of features exceeded a threshold number (e.g., 3 or more, 4 or more, or 5 or more). With a threshold of 3 or more features, EUS and ERP agreed in 215 of 269 patients (80%). EUS was abnormal in 40 of 269 (15%) and ERP was normal and EUS was normal and ERP was abnormal in 14 of 269 (5%).

Two studies have compared EUS to pancreatic function testing for chronic pancreatitis.\textsuperscript{12,14} The combined results from 96 patients evaluated with analysis of secretin stimulated pure pancreatic juice or duodenal bicarbonate are reported. EUS and pancreas function testing agreed in 72 of 96 cases (75%). EUS was abnormal and ERP normal in 17 of 96 (18%), and EUS was normal and ERP abnormal in 7 of 96 (7%).

There are several limitations of our study. EUS diagnoses of CP were based on videotaped examinations that may not completely reproduce the features of CP compared with an actual examination. However, videotape examination is preferable to diagnoses based on still images. In addition, a clinical history was not provided to the endosonographers who rated the videotaped examinations. Providing the age of the patient may have improved the diagnostic agreement for EUS because several endosonographers took age into account and all of these required fewer features to make a diagnosis of CP for younger patients. Although no difference in the interobserver agreement was seen between third tier and non-third tier trained endosonographers, the study was not powered to exclude a type II error in this assessment.

How can the interobserver agreement for the EUS diagnosis of CP be improved? It is clear from this study that both clinical and EUS variables must be taken into account, and that equal weight is not given to each variable. Multivariable analysis and the development of a weighted scoring system may overcome this problem, but such an analysis requires a reference standard for comparison. More specific definitions of features can be completed in the absence of such a reference standard and would likely improve interobserver agreement.

There is also lack of agreement on the implication of visible side branches. The study by Wiersema et al.\textsuperscript{12} found that normal persons had visible side branches that were similar in diameter to patients with nonpancreatic abdominal pain and those with chronic pancreatitis. This study suggests that patients may normally have visible side branches of up to 1 mm in the head or body and up to 0.5 mm in the tail. Hyperechoic foci and strands also had poor agreement. This may be due to the subjective and relative nature of the term “hyperechoic.” One possible solution is to use image analysis, which would provide objective measures of the echogenicity of foci, and the surrounding parenchyma. Although cysts and lobularity had better agreement, these could be improved through the use of objective image analysis, which describes the variation in the overall echogenicity across the parenchyma. Last, irregularity of the duct contour must also be better defined.

It is possible that the degree of agreement will be lower when EUS is performed by less experienced endosonographers. The group in this study was highly experienced, and many members have published research studies in the field of EUS and CP. The degree of agreement between less experienced endosonographers and trainees is the subject of active study.

The optimal criteria for the diagnosis of CP by EUS are unknown. Most endosonographers in this study evaluated a series of individual features and then summed the number of features present. The threshold for diagnosing CP on the basis of EUS can be varied (e.g., 3 or more, 4 or more, or 5 or more features). The sensitivity and specificity of EUS compared with ERP depend on what threshold is chosen. If a low threshold is used (more than 1 to 2 features), the sensitivity will be extremely high, but the specificity will be low. If a higher threshold is used (more than 5 to 6 features), the sensitivity will be low, but the specificity will be high. An appropriate threshold can be chosen on the basis of the question at hand. If the purpose of an EUS is to exclude disease, a low threshold will give the best negative predictive value. For example, a patient with only 0 to 1 features of CP by EUS has a greater than 90% chance of having a normal ERP\textsuperscript{15} and presumably does not have CP. In such a case, further testing is unlikely to identify CP and may be avoided by performing EUS as an initial diagnostic test. If the purpose of an EUS is to establish the diagnosis with a high degree of certainty, a threshold will give a greater than 80% chance that the patient will have an abnormal ERP.\textsuperscript{15} Those with fewer features may require further evidence to establish the diagnosis of CP.

All EUS features of CP may not be equally important. For example, the presence of intraductal calcifications alone is highly suggestive of CP even in the absence of other features. The presence of calcified stones was rated the single most important predic-
Reliability of EUS for the diagnosis of chronic pancreatitis

M Wallace, R Hawes, V Durkalski, et al.

In summary, this study supports the hypothesis that EUS is a reliable method for the diagnosis of CP. EUS has inherent advantages over more invasive tests such as pancreatography. Given the current lack of a reference standard for the diagnosis of CP, no single test, including EUS, is likely to provide a completely accurate and reliable diagnosis. There is still a need to improve the reliability, especially for some individual features of CP, and to improve the diagnostic criteria based on individual features. Despite these limitations, the current methods used to diagnose CP by EUS offer good overall agreement and are comparable to other standard diagnostic endoscopic, radiologic and clinical tests.

REFERENCES