EUS: a meta-analysis of test performance in suspected choledocholithiasis

Frances Tse, MD, FRCPC, Louis Liu, MEng, PhD, MD, FRCPC, Alan N. Barkun, MD, CM, MSc (Epid), David Armstrong, MA, MB Bchir, FRCPC (UK), FRCPC, Paul Moayyedi, BSc, MB ChB, FRCP (UK), PhD, MPH

Hamilton, Ontario, Canada

Background: EUS has been proposed as a less invasive means of diagnosing choledocholithiasis and may eliminate the need for ERCP and its associated risks. The literature pertaining to EUS for the diagnosis of choledocholithiasis reports widely varying sensitivities and specificities.

Objective: To more precisely estimate the diagnostic accuracy of EUS in suspected choledocholithiasis.

Design: MEDLINE and EMBASE databases were used to identify prospective cohort studies in which the results of EUS were compared with the results of an acceptable criterion standard, including ERCP, intraoperative cholangiography, or surgical exploration. Two independent reviewers extracted standardized data and assessed trial quality. A random effects model was used to estimate the sensitivity, specificity, likelihood, and diagnostic odds ratio (DOR), and a summary receiver operating characteristic curve was constructed. All predefined potential sources of heterogeneity were explored by subgroup analysis and meta-regression.

Patients: A total of 2673 patients with suspected choledocholithiasis were reported in 27 studies that satisfied the inclusion criteria.

Results: EUS had a high overall pooled sensitivity of 0.94 (95% CI, 0.93-0.96), a specificity of 0.95 (95% CI, 0.94-0.96), and an area under the curve of 0.98. Three variables appeared to yield a higher DOR: a higher disease prevalence, an adequate time interval between index test and criterion standards, and the presence of verification bias.

Limitations: Misclassification of patients by imperfect criterion standards could potentially underestimate the performance of an EUS.

Conclusions: An EUS is a noninvasive test, with excellent overall sensitivity and specificity for diagnosing choledocholithiasis. An EUS should, therefore, be used to select patients for a therapeutic ERCP to minimize the risk of complications associated with unnecessary diagnostic ERCP. (Gastrointest Endosc 2008;67:235-44.)
imaging modality should document the presence of CBDS first, before proceeding with an ERCP.

An EUS combines an endoscopy and US to provide high-resolution images of the pancreatobiliary system, without the interference of subcutaneous fat and bowel gas. By using high frequencies (7.5 and 12 MHz), the resolution of an EUS is less than 1 cm at a focal distance of 5 cm. A stone in the CBD appears as a hyperechoic focus with associated acoustic shadowing. An EUS has been proposed as a less invasive means of diagnosing CBDS that may eliminate the need for ERCP and its associated risks in certain patients. However, evidence for its use is lacking, because some patients who undergo an EUS will require an ERCP for therapy. The possible morbid implications of false-positive and false-negative results also need to be considered. Furthermore, the literature that pertains to EUS for the diagnosis of CBDS reports widely varying sensitivities of 71% to 100% and specificities of 67% to 100%, when EUS was compared with ERCP, intraoperative cholangiogram (IOC), surgical exploration, and/or clinical follow-up. The reason for this is unclear but may relate to variation in patient selection, study design, diagnostic threshold, and operator expertise. To our knowledge, there exist no published meta-analysis about the accuracy of EUS in detecting CBDS, and we, therefore, conducted a systematic review of studies on this topic.

MATERIALS AND METHODS

Search strategy
The systematic review was performed according to developed guidelines for conducting diagnostic reviews.44 We searched MEDLINE (1966 to February 2006), EMBASE (1980 to February 2006), the Cochrane Database of Systematic Reviews, the Cochrane Controlled Trials Register, and the Database of Reviews of Effectiveness. Medical subject headings “endosonography” and “choledocholithiasis” were combined with free-text terms, screening title, abstract, and subject heading for endoscopic ultrasound, endoscopic ultrasonography, common bile duct stones, and gallstones. The reference lists of all articles selected for the review were screened for potentially relevant articles that were not identified by the initial search. We also conducted a manual search of abstracts submitted to Digestive Disease Week that covered the past 5 years (2000-2005). In addition, we contacted two international content experts to provide information on any studies that may have been missed in the electronic and manual searches. The search was restricted to human and English language studies a priori.

Two reviewers (F.T. and L.L.) independently screened the titles and abstracts of all articles according to a priori defined inclusion and exclusion criteria. Agreement was measured by using kappa statistics, and differences were resolved by discussion and consensus with a third reviewer (P.M.). The complete report of all selected articles was then retrieved and reviewed by the same two reviewers.

Inclusion criteria
Intervention. Radial scanning echoendoscopes with frequencies of 7.5 to 12 MHz were used.

Criterion standards. Acceptable criterion standards included ERCP, IOC, or surgical exploration. A clinical follow-up period of at least 3 months for confirmation of a negative result was also regarded as a criterion standard to avoid partial verification bias.45

Population. The study population consisted of patients with suspected CBDS based on history, physical examination, laboratory data, and transabdominal US findings.

Study design. The study design was a prospective cohort study in which the results of EUS were compared with the results of a criterion standard as defined above.

Outcome. The outcome consisted of reporting of results in sufficient detail to allow reconstruction of a diagnostic 2 × 2 table (ie, true positive, true negative, false positive, and false negative).

Exclusion criteria
Case series and case reports were excluded from the review because of a lack of a comparison group. We also excluded case-control and retrospective designs that were previously shown to introduce bias and an overestimation of diagnostic performance.46 Studies that did not provide sufficient data for reconstruction of a diagnostic 2 × 2 table were excluded. In addition, miniprobes, intraductal US, prototypes of EUS, or microlithiasis were excluded.

Data abstraction
Two independent reviewers (F.T. and L.L.) extracted the following data from the selected studies onto standardized data forms: study characteristics (design, country, year of publication, setting, sample size, clinical context, time interval between EUS examinations, and criterion

Capsule Summary

What is already known on this topic
- ERCP is effective in detecting and extracting common bile duct stones, but the procedure is associated with a 5% to 10% complication rate and often yields negative results.

What this study adds to our knowledge
- When data are used from 2673 patients with suspected choledocholithiasis, reported in 27 studies, EUS is seen to have a high overall pooled sensitivity and specificity for diagnosing choledocholithiasis, which makes it useful for selecting patients for a subsequent “therapeutic” ERCP.
standards), population (mean age, proportion of male subjects, and prevalence of CBDS), interventions (criterion standards, and manufacturer and operating frequencies of EUS), and outcomes (number of true positives, true negatives, false positives, and false negatives for the presence of CBDS). These data were considered missing if not explicitly mentioned in the text. Discrepancies were resolved by discussion and consensus with a third reviewer (P.M.).

**Quality assessments**

A component approach was taken in the assessment of trial quality, because previous studies showed a lack of agreement between different scoring systems.\(^{47,48}\) Two independent reviewers (F.T. and L.L.) assessed the quality of each study by using 3 criteria adapted from Irwig et al\(^ {14}\) and the Quality Assessment of Diagnostic Studies tool.\(^ {49}\) Agreement was measured by using kappa statistics, and differences were resolved by discussion.

**Blinding.** The absence of blinding can bias the interpretation of the reference or index tests; therefore, if authors did not explicitly state the use of blinding, we assumed that this was not the case.

**Complete verification.** Verification bias is present if the decision to perform the reference test is based on the result of the index test. This criterion was met if verification was obtained by a criterion standard, regardless of the index test results. In many diagnostic studies with an invasive test, most of the positive results and only a small part of the negative results are verified.\(^ {45}\) Alternately, negative test results are verified by a different, less thorough, standard, such as a clinical follow-up.\(^ {45}\) In cases in which more than 10% of the study group was not subjected to the criterion standard, the study was scored as applying partial verification; in cases in which different criterion standards were used, the study was scored as a differential criterion standard.\(^ {45}\)

**Adequate time interval.** When a delay occurs between the index test and the criterion standard, the target condition may change, thereby leading to misclassification and an underestimation of diagnostic accuracy.\(^ {50}\) In fact, spontaneous passage of CBDS has been reported to occur in 21% of cases within 3 days of symptoms and an additional 20% between day 3 and 27.\(^ {51,52}\) The time interval was considered adequate a priori if the criterion standard (except clinical follow-up) was performed within 72 hours of the EUS examination. All other cases were considered inadequate. Also, if the time interval was not specified by the investigators, then we assumed that this was inadequate.

**Statistical analysis**

Agreement was assessed by using a kappa statistic, which is a measure of relative agreement over and above chance. True positive, false positive, true negative, and false negative were entered into MetaDiSc version 1.3 statistical software (MetaDiSc, Unit of Clinical Biostatistics team of the Ramon y Cajal Hospital, Madrid, Spain). The sensitivity, specificity, positive likelihood ratio (LR), negative LR, and diagnostic odds ratio (DOR) (positive LR/negative LR) were calculated for each study and then were pooled by using the DerSimonian-Laird random effects model.\(^ {53}\) Likelihood ratios above 10 and below 0.1 are considered to provide strong evidence to rule in or rule out a diagnosis, respectively. The DOR is a single indicator of test accuracy that combines the data from sensitivity and specificity.\(^ {54}\) The DOR of a test is the ratio of the odds of positive test results in the diseased patient relative to the odds of positive test results in the nondiseased patient. MetaDiSc version 1.3 was used to generate forest plots of sensitivity, specificity, and positive and negative LRs. StatsDirect version 2.5.6 (StatsDirect Ltd, Cheshire, England) was used to generate forest plots of DOR.

Heterogeneity was assessed by both \(\chi^2\) and \(I^2\) statistics.\(^ {55,56}\) The \(\chi^2\) test, with degrees of freedom = number of studies – 1, assessed whether observed differences in results were compatible with chance alone. A \(P\) value <.10 (or a large \(\chi^2\) statistic relative to degrees of freedom) was considered evidence of heterogeneity beyond chance a priori. An \(I^2\) index describes the percentage of total variation across studies that is from heterogeneity rather than chance. A value greater than 50% was considered substantial heterogeneity a priori. Implicit diagnostic threshold effects were assessed by using the Spearman correlation coefficient of sensitivity and 1 – specificity. A diagnostic effect was considered to be present in the case of a Spearman correlation of \(\rho < -0.6.\)\(^ {57}\) In addition, a summary receiver operating characteristic curve (SROC) was constructed by using DerSimonian-Laird random effects model.\(^ {53}\) The model was weighted by the inverse variance of the log of the DOR. The area under the curve (AUC) was computed by numeric integration of the curve equation by the trapezoidal method.\(^ {58}\) A perfect test has an AUC close to 1, and poor tests have AUCs close to 0.5.

We defined 5 potential sources of heterogeneity a priori: (1) the prevalence of CBDS, (2) the clinical context (suspected CBDS vs acute biliary pancreatitis vs suspected biliary obstruction), (3) interpretation of test results (blinded vs not blinded), (4) verification bias (absence vs presence), and (5) the time interval between the criterion standard and an EUS (adequate vs inadequate). All predefined potential sources of heterogeneity were explored by subgroup analysis and meta-regression of the ln(DOR) by adding covariates to the SROC model.\(^ {57}\) The results of the meta-regression model were expressed as relative diagnostic odds ratios (RDOR) of the corresponding covariate.\(^ {59}\) RDOR indicates a change in diagnostic performance of the test under study per unit increase in the covariate. Analysis was performed by using Comprehensive Meta-analysis version 2.0 (Biostat, Englewood, NJ).

The robustness of the meta-analysis to publication bias was assessed by funnel plots and various bias indicators, including the Egger and Fail-safe N tests, and the trim-and-fill method.\(^ {60,61}\) Analysis was done by using
StatsDirect version 2.5.6 and Comprehensive Meta-analysis version 2.0. *P* values < .05 were considered statistically significant.

RESULTS

Literature search

A total of 165 studies were initially identified by using the search strategy. Two independent reviewers excluded 128 studies after a preliminary review of titles and abstracts, which left 37 for detailed evaluation. Of these 37 potentially eligible studies, 25 published articles that met the inclusion and exclusion criteria were identified. Two additional unpublished studies in abstract form were identified by a manual search. In total, 27 studies were deemed appropriate for the meta-analysis. There was excellent agreement between reviewers (k statistic 0.82, 95% CI, 0.78-0.93) for inclusion. The study selection process is shown in Figure 1.

Description of studies

The studies had a total of 2673 patients, with cohorts ranging in size from 20 to 459 patients (mean *N* = 99). Studies originated from 12 different countries, but most were done in Europe (63%). All of the studies were prospective cohort studies conducted in tertiary-care settings. The mean age of participants was 60.1 years (range 10-95 years); 40% were men. The CBDS prevalence varied from 15% to 86% (mean 36%). Recruitment for all studies was based on the presenting symptoms, screened with history taking, a physical examination, basic laboratory tests (including liver enzymes), and a transabdominal US. Eighteen studies (67%) were done for the evaluation of suspected CBDS, 3 studies (11%) were for acute biliary pancreatitis, and 6 studies (22%) were for suspected biliary obstruction or disease. All studies had some type of criterion standard (as defined above) performed in all patients, and 14 of the studies (52%) used the same criterion standard on all patients. Twenty-five studies (93%) used ERCP, 12 studies (44%) used surgical exploration or IOC, and 8 studies (30%) used clinical follow-up to confirm negative EUS results. The study characteristics of the included studies are shown in Table 1.

Methodologic quality assessment

Study quality was generally low. Only 9 studies (33%) satisfied all 3 criteria used. The criterion that was satisfied least often was blinded interpretation of both the index test and the criterion standard. Overall, there was excellent agreement between reviewers (k statistic 0.75, 95% CI, 0.68-0.82). The results of the quality assessment of included studies are summarized in Table 1.

Figure 1. Flow diagram of the studies identified in the systematic review.
Blinding. Because EUS was always performed first, before the criterion standards, interpretation of the EUS results was presumed to be without knowledge of the results of the criterion standards. However, the criterion standards may be interpreted with full awareness of the results of the EUS. A total of 11 studies (41%) reported blinded interpretation of both test results.

Complete verification. All studies used an acceptable criterion standard, as defined above, for all patients, thus avoiding partial verification bias. Seven studies...

<table>
<thead>
<tr>
<th>Source</th>
<th>No. patients</th>
<th>Clinical context</th>
<th>Criterion standard</th>
<th>Prevalence of CBDS (%)</th>
<th>Blinding</th>
<th>Complete verification</th>
<th>Adequate time interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liu et al, 2005</td>
<td>139</td>
<td>2</td>
<td>ERCP, surgery, F/U</td>
<td>35</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Kondo et al, 2005</td>
<td>30</td>
<td>1</td>
<td>ERCP</td>
<td>86</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Aube et al, 2005</td>
<td>45</td>
<td>1</td>
<td>ERCP, IOC</td>
<td>36</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Meroni et al, 2004</td>
<td>47</td>
<td>1</td>
<td>ERCP, IOC</td>
<td>15</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Ainsworth et al, 2003</td>
<td>163</td>
<td>3</td>
<td>ERCP</td>
<td>33</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Napoleon et al, 2003</td>
<td>238</td>
<td>1</td>
<td>ERCP, F/U</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Buscarini et al, 2003</td>
<td>459</td>
<td>1</td>
<td>ERCP, surgery, F/U</td>
<td>53</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Scheiman et al, 2001</td>
<td>28</td>
<td>3</td>
<td>ERCP</td>
<td>18</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Berdah et al, 2001</td>
<td>68</td>
<td>1</td>
<td>ERCP, F/U</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Materne et al, 2000</td>
<td>50</td>
<td>3</td>
<td>ERCP, surgery, F/U</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Polkowski et al, 1999</td>
<td>50</td>
<td>1</td>
<td>ERCP</td>
<td>32</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Chak et al, 1998</td>
<td>36</td>
<td>2</td>
<td>ERCP</td>
<td>36</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>de Ledinghen et al, 1999</td>
<td>32</td>
<td>1</td>
<td>ERCP, IOC</td>
<td>31</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Sugiyama et al, 1998</td>
<td>35</td>
<td>2</td>
<td>ERCP</td>
<td>43</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Montariol et al, 1998</td>
<td>215</td>
<td>1</td>
<td>IOC</td>
<td>19</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Canto et al, 1998</td>
<td>64</td>
<td>1</td>
<td>ERCP</td>
<td>30</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Norton and Alderson, 1997</td>
<td>50</td>
<td>1</td>
<td>ERCP, surgery, F/U</td>
<td>48</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Burtin et al, 1997</td>
<td>68</td>
<td>3</td>
<td>ERCP</td>
<td>49</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Sugiyama and Atomi, 1997</td>
<td>142</td>
<td>1</td>
<td>ERCP</td>
<td>36</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Aubertin et al, 1996</td>
<td>50</td>
<td>1</td>
<td>IOC, F/U</td>
<td>36</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Prat et al, 1996</td>
<td>121</td>
<td>1</td>
<td>ERCP</td>
<td>64</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Shim et al, 1995</td>
<td>132</td>
<td>1</td>
<td>ERCP</td>
<td>21</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Amouyal et al, 1994</td>
<td>62</td>
<td>1</td>
<td>ERCP, IOC</td>
<td>52</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Edmundowicz et al, 1992</td>
<td>20</td>
<td>1</td>
<td>ERCP, IOC</td>
<td>20</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Dancygier and Nattermann, 1994</td>
<td>41</td>
<td>3</td>
<td>ERCP</td>
<td>39</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Fernandez-Esparrach et al, 2005</td>
<td>158</td>
<td>3</td>
<td>ERCP, surgery, IOC</td>
<td>39</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Fazel et al, 2002</td>
<td>40</td>
<td>1</td>
<td>ERCP</td>
<td>40</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

F/U, Follow-up.
*0, absent; 1, present.
†The patient sample in which EUS was studied, stratified by the most highly suspected pretest diagnosis (1, CBDS; 2, biliary pancreatitis; 3, biliary obstruction/disease).
(26%) used different criterion standards according to the results of the EUS. In these studies, negative test results were verified by clinical follow-up for a variable period of time (range 3–36 months), and positive results were confirmed by invasive criterion standards (ERCP, IOC, or surgery), thus introducing differential verification bias.

**Adequate time interval.** Six studies (22%) had verification of diagnosis delayed for more than 72 hours after an EUS (range 6-22 days).

**Meta-analysis**

**Diagnostic accuracy.** The Forest plots of sensitivity and specificity of EUS for the diagnosis of CBDS are shown in Figure 2. Point estimates were plotted with 95% CIs for each cohort. Pooled sensitivity, specificity, positive LR, negative LR, and DOR were 0.94 (95% CI, 0.93-0.96), 0.95 (95% CI, 0.94-0.96), 22.41 (95% CI, 12.53-40.08), 0.09 (95% CI, 0.06-0.12), and 312.15 (95% CI, 163.42-596.24), respectively. Tests of heterogeneity were highly significant for all measures. When restricted to the 25 study results published in the articles, heterogeneity was considerably reduced (I² index <50%) for all measures except sensitivity (I² index 57.7%) and positive LR (I² index 51.6%).

**Diagnostic threshold effect.** The Spearman correlation coefficient was −0.075 (P = .720), which indicated the absence of a substantial diagnostic threshold effect. The overall accuracy of EUS in a SROC plot is shown in Figure 3. The symmetric curve shows a trade-off between sensitivity and specificity. The area under the SROC was 0.98, which indicated high accuracy.

**Sources of heterogeneity.** To identify factors associated with heterogeneity, we performed subgroup analysis and meta-regression. When stratified by the most highly suspected pretest diagnosis (clinical context), studies that included patients with suspected acute biliary pancreatitis had a DOR 2.5 and 3.6 times higher than those that included suspected biliary obstruction and suspected CBDS, respectively. However, the CIs of the 3 subgroups overlapped. Studies that had an adequate time interval between the index test and the criterion standard showed a DOR 5.2 times higher than studies that had an inadequate time interval. Studies with verification bias produced a DOR 2.4 times higher than those without such bias. When stratified byblindness, no major difference was seen in the DOR. Because the prevalence of CBDS greatly varies in different populations, any cutoff point chosen for high and low prevalence would be arbitrary. Therefore, a subgroup analysis stratified by low or high prevalence was not performed. Univariate subgroup analyses by t tests and analysis of variance showed no statistically significant difference between subgroups.

A meta-regression was performed to simultaneously evaluate multiple covariates in the same analysis. The outcome of the regression analysis as the RDOR are shown in Table 2. In summary, a higher prevalence of CBDS, an adequate time interval, and verification bias were associated with a higher RDOR. Blinding and clinical context were not statistically significant in the regression model.

**Publication bias.** A funnel plot was asymmetric; smaller studies with low DOR estimates were missing, which indicated a potential for publication bias. The Egger test for publication bias was also statistically significant (P = .0078). The fail-safe N test indicated that it would take an additional 3685 studies with no significant findings for the combined two-tailed P value to no longer be significant (P > .05). By using the random effects model,
the DOR and 95% CI for the combined studies was 237.30 (95% CI, 158.31-355.70). After adjusting publication bias with the trim-and-fill method, the imputed DOR was 124.51, 95% CI, 84.36-183.75.

DISCUSSION

This systematic review of EUS for patients with suspected CBDS found that an EUS had excellent diagnostic accuracy, with an AUC of 0.98. Our findings suggest that EUS results can be taken as conclusive evidence for both ruling in (positive LR 22.41) and ruling out (negative LR 0.09) the diagnosis of CBDS. Because an EUS offers high resolution (0.1 mm), it is likely that the superiority of an EUS compared with an MRCP or an ERCP is primarily evident in the detection of small stones. Unfortunately, we were not able to perform a subgroup analysis of stone size on EUS performance, because the information on diagnostic accuracy according to stone size was not recorded in the included studies. For the 3 studies that reported details about stone size,18,19,41 the accuracy of EUS appeared to be independent of stone size, being highly sensitive even for stones less than 5 mm in diameter; whereas the sensitivities of an MRCP or an ERCP decreased with stone size. However, the clinical relevance of EUS-detectable small stones has been questioned, because these stones are presumably capable of passing spontaneously into the intestine.

Although we reported excellent diagnostic accuracy for EUS, our systematic review also revealed the poor quality and reporting of the primary studies. Differential verification bias was present in about a third of the studies, because these studies used invasive criterion standards (ERCP, IOC, and surgery) for patients with positive test results and clinical follow-up as the criterion standard for patients with negative test results. In addition, only 41% of studies reported that clinicians who determined the criterion standards were blinded to EUS results. Both of these methodologic weaknesses could lead to an overestimation of diagnostic performance.45,59 In contrast, a delay in verification of a diagnosis was associated with a considerable underestimation of diagnostic accuracy. There was also much variability in the time interval between the index test and the criterion standard.

Heterogeneity was evident in the results and could be partially explained by our subgroup analysis and meta-regression. It might be argued that pooling studies with heterogeneous samples is inappropriate. However, it does provide a useful baseline from which to explore sources of heterogeneity. The meta-regression highlighted 3 variables that appear to yield higher diagnostic odds ratio: a higher disease prevalence, an adequate time interval between the index test and criterion standards, and the presence of verification bias. Post hoc analysis after withdrawing the two unpublished studies resolved heterogeneity for all measures, except sensitivity (I^2 index 57.7%) and positive LR (I^2 index 51.6%). Unexplained heterogeneity is probably caused by factors that were inadequately reported in the primary studies and, therefore, could not be explored in our meta-analysis. These factors include characteristics of the patients recruited (comorbidities and pretest probability of CBDS) and the training and experience of the endosonographers, as well as stone size and location in the biliary tree (proximal vs distal).

Our review has limitations. We only included English language studies. We also did not hand search the leading journals. The funnel plot of the log DOR was asymmetrical, which suggests that publication bias may be present. Nevertheless, because the fail-safe number was 3685
studies, we feel confident about our findings, because it is highly unlikely that there are 3685 additional unpublished studies with nonsignificant findings. Also, the imputed DOR of 124.41 based on the trim-and-fill method was still highly significant.

Another potential limitation of our review is the use of imperfect criterion standards. Cholangiography, either an ERCP or an IOC, was used as the criterion standard. Although this was usually combined with a sphincterotomy and duct instrumentation, or surgical exploration, it is well recognized that small stones, especially in dilated bile ducts, may be missed by direct cholangiographic methods. Furthermore, it can be difficult to differentiate air bubbles, introduced inadvertently during an ERCP, from small stones. Misclassification of patients by imperfect criterion standards could potentially underestimate the performance of an EUS.

Technical limitations of EUS are few and include the following: upper-GI stenosis, previous gastrectomy, or Billroth II resection. Visualization is limited to the 8-cm to 10-cm depth from the probe, and imaging can be obscured by pneumobilia, surgical clips, calcifying pancreatitis, or a duodenal diverticula. Of importance, the accuracy of an EUS is highly operator dependent. Because all studies were primarily conducted in referral centers, the results may not be generalizable to community-based practices.

Based on local availability and expertise, an EUS could be incorporated into the diagnostic algorithm of patients with suspected CBDS. Based on a number of prospectively validated clinical algorithms, patients can be risk stratified into low, intermediate, and high risk for CBDS. Patients with the lowest risk of having CBDS do not need a cholangiography. For patients at intermediate risk of having CBDS, an EUS can be used to select patients who require a therapeutic ERCP. Patients at high risk for CBDS would benefit most from an ERCP. However, even patients classified as high risk were found to have CBDS in only 66% to 70% of cases. An argument, therefore, could be made to perform an EUS first in these patients to reduce an unnecessary diagnostic ERCP with its attendant risks.

In conclusion, currently available literature demonstrated that an EUS has excellent diagnostic accuracy for CBDS. An EUS, therefore, should be used to select patients for therapeutic ERCP to avoid the risk of complications and death associated with an unnecessary diagnostic ERCP. However, real-life results may be different than what could be anticipated based on test performance characteristics. Clinicians may be reluctant to direct clinical actions based solely on EUS results; consequently, an ERCP will often be requested to confirm the findings, thereby adding to risks, costs, and patient inconvenience. Future studies should determine the impact and cost-effectiveness of an EUS on clinical decision making or patient outcomes.

ACKNOWLEDGMENTS

The authors wish to thank the two international content experts for this study: Pascal Burtin (France) and Laurent Palazzo (France).

DISCLOSURE

The authors report that there are no disclosures relevant to this publication.

REFERENCES


Received July 26, 2007. Accepted September 20, 2007.
Current affiliations: Division of Gastroenterology (F.T., LL., D.A., P.M.), McMaster University Medical Centre, McMaster University, Hamilton, Ontario, Division of Gastroenterology (A.N.B.), Montreal General Hospital Site, McGill University Health Centre, Montreal, Quebec, Canada.
Reprint requests: Frances Tse, MD, Division of Gastroenterology, McMaster University Medical Centre, 1200 Main St West, Hamilton, Ontario, Canada, L8N 3Z5.

Online Audio and Podcasting

Audio and Podcasts of article abstracts published in Gastrointestinal Endoscopy are now available online. Recordings are edited by Steven Edmundowicz, MD, Senior Associate Editor, and performed by Deborah Bowman, MFA, Managing Editor of Gastrointestinal Endoscopy.
Log on to www.giejournal.org to listen to recordings from the current issue.