Chapter 2. Methodology

This report is the product of a systematic literature review of the evidence on the diagnostic and therapeutic effectiveness of endoscopic retrograde cholangiopancreatography (ERCP) with a specific focus on four clinical conditions: (1) common bile duct stones; (2) pancreaticobiliary malignancy; (3) pancreatitis; and (4) abdominal pain of possible pancreaticobiliary origin. In addition, the evidence describing patient, procedure, or operator determinants of complications of ERCP is systematically reviewed. Also reviewed is the evidence on the prediction of common bile duct stones.

The protocol for this review was designed prospectively as much as possible to define: study objectives; search strategy; patient populations of interest; study selection criteria; outcomes of interest; data elements to be abstracted and methods for abstraction; and methods for study quality assessment.

The key questions guiding the scope of this report have been outlines in the Introduction. This chapter of the report describes the search strategies used to find articles, the criteria and methods for selecting eligible articles, the methods for data abstraction, the methods for quality assessment, and finally, the peer review and technical assistance received during the project.

Search Strategy for the Identification of Articles

The National Library of Medicine (NLM) conducted a comprehensive literature search for journal articles on ERCP from the PubMed/MEDLINE, BIOSIS, EMBASE, and SCISEARCH databases with a publication date from 1980 forward until the final search date of August 13, 2001. Articles which had been indexed to the NLM Medical Subject Heading (MeSH®) "cholangiopancreatography, endoscopic retrograde" as well as those containing the following list of ERCP synonyms and textword combinations were retrieved:

Endoscopic retrograde cholangiopancreatogr?
Endoscopic retrograde cholangio-pancreatogr?
Endoscopic retrograde pancreatocholangiogr?
Endoscopic retrograde pancreato-cholangiogr?
ERCP
ERCPs
Endoscopic retrograde cholangiogr?
ERC and endoscop?
ERC and cholangiogr?
Endoscopic cholangiogr?
Endoscopic retrograde pancreatogr?
ERP and endoscop?
ERP and pancreatogr?
Endoscopic pancreatogr?
Endoscopic cholangiopancreatogr?
Endoscopic cholangio-pancreatogr?

ECP and endosc?
ECP and cholangiogr?
Endoscopic pancreatocholangiogr?
Endoscopic pancreato-cholangiogr?
EPC and endoscop?
EPC and pancreatogr?

Textwords are words appearing in the titles, abstracts, and subject term lists of the online record of the articles.

The "?" is a truncation symbol used to permit retrieval for variant word endings, as cholangiopancreatography, cholangiopancreatographic, etc.

Excluded from the search results were articles that:

- were written in a foreign language
- did not have abstracts as a part of the online record in any of the databases searched
- did not include human subjects
- contained reports of only a single case

Citations without abstracts were not reviewed, as citations that have no abstracts have little or no yield in producing articles eligible for inclusion in the evidence report.

There was not a method developed to systematically identify studies published in abstract form only. However, if an abstract of potential importance was identified, it was included it if it was published in 1999 or after, with the reason that abstracts published before 1999 should have been published in full manuscript form by now.

Secondary Search Strategy

The literature search for the supplemental question (Topic 1c), for the indirect comparison of single arm studies of for ERCP-guided fine needle aspiration (FNA) and EUS-guided FNA for Topic 2, and for additional studies selected by the secondary selection criteria for Topics 3 and 4, did not follow the same search process. The literature review process for these supplemental questions was based on a focused identification and selection of key articles addressing the clinical issue of interest. Reference lists from these articles, were then reviewed, focused MEDLINE searches were performed, and related articles identified. It was thought that this approach led to retrieval of the important studies addressing the questions of interest.

The Technical Advisory Group and individuals and individuals providing peer review also were asked to inform the project team of any studies relevant to the key questions addressed in this evidence report that were not retrieved by either of the search strategies.

Search Results

The online searches of the PubMed, EMBASE, BIOSIS, and SciSEARCH databases in conjunction with additional citations identified through manual searching yielded a total of 5,698 titles and abstracts for review. During application of Phase I of the selection process, 789 articles were selected for review in full text. Approximately 117 of these articles were identified as review articles. Primary and secondary selection criteria were applied to articles identified as potential clinical trial reports. This process yielded a total of 149 included studies for the review of evidence. Citations for the excluded articles and the reason(s) for exclusion are listed in Appendix A.

Study Selection Criteria

Primary Selection Criteria

The criteria which applied to all topic areas in this report were:

1. Full-length report in peer-reviewed medical journals.
2. Published in the English language.
3. Study reported outcomes relevant to this systematic review.
4. Where there were multiple reports of a single study, only the report judged to be most recent and complete, based on number of included patients and length of follow-up, was included. If additional relevant outcomes were included in the duplicate reports, these data were abstracted and added to the data from the primary report with citation to the supplementary articles.
5. Was prospective in design, or if retrospective, enrolled consecutive patients or with appropriate sampling methods (i.e. case-control sampling method).

For diagnostic performance topic areas, studies were included if the study:

1. Compared ERCP and at least one of the relevant diagnostic alternatives or compared two ERCP alternatives. Relevant diagnostic alternatives included endoscopic ultrasound, MRCP, intraoperative cholangiography, or other diagnostic tests as advised by the TAG. Studies reporting only non-breath hold MRCP imaging techniques were not included in this review as these do not represent the current state-of-the-art MRCP techniques.
2. Subjected all participants to both ERCP and the relevant diagnostic alternative;
3. Addressed a relevant patient population;
4. Included at least 25 subjects;
5. Reported sufficient information to be able to calculate 2x2 contingency tables of diagnostic performance.

For therapeutic outcome topic areas, studies were included if they:

1. Compared ERCP strategies with at least one of the relevant therapeutic alternatives. Relevant therapeutic alternatives included surgical methods to remove common ducts stones,

surgical methods of bypassing malignant biliary obstructions, and surgical and medical methods of treating pancreatitis and pancreatitis-associated conditions.
2. Addressed a relevant patient population;
3. Included at least 25 subjects in each treatment group being analyzed separately; however, this criterion was relaxed to require 25 subjects in the trial for pancreaticobiliary malignancy and abdominal pain of possible pancreaticobiliary origin.
4. Reported on at least one relevant outcome measure;
5. Was a contemporaneous comparison study or if it was a noncontemporaneous study, the populations and treatment setting were comparable;

For Part V, a study was included if it:

1. Included an analysis of the relationship between patient, procedure, or operator factors and ERCP complications;
2. Enrolled at least 100 patients if a cohort study, or at least 25 cases if a case-control study;
3. Addressed potential confounding variables in either the selection of subjects or analysis.

For Part I, Section 3, a study was included if it:

1. Reported the association of individual risk factors of interest and the presence of a common bile duct stone. Based on a consensus from the TAG, these individual risk factors were jaundice, liver function test results, and an ultrasound finding of a dilated common bile duct.
2. Reported the association of a prediction rule or model predicting likelihood of having a common bile duct stone and the presence of a common bile duct stone;
3. Enrolled at least 100 patients;
4. Reported sufficient information to be able to calculate 2x2 contingency tables of diagnostic performance in the prediction of presence or absence of a common bile duct stone.

Secondary Selection Criteria

Due to a paucity of literature which met the primary selection criteria for Part III, Section 2 and Part IV, Section 2, additional selection criteria were created so that these questions could be examined. There was a lack of literature which provided comparative data on the value of ERCP treatment for these conditions. Thus studies were included from the primary search strategy and sought out using the secondary search strategy if the study was:

1. a randomized controlled trial or otherwise concurrently controlled study of an ERCP intervention compared to a relevant therapeutic alternative, regardless of sample size;
2. a single arm observational study (subject serves as own control) of ERCP intervention in treatment of chronic pancreatitis or chronic abdominal pain of possible pancreaticobiliary origin with a minimum size of 25 subjects; where the studies selected a well-defined population with a predictable natural history absent intervention based on thorough baseline evaluation; and where the study used an appropriate well-designed outcome measure. Baseline evaluation had to be obtained over a sufficient time period (approx. 3 months) and follow-up data needed be obtained over at least 6 months. Studies reporting exploration of subgroup differences in observed results were also included.

3. A single arm observational study of an ERCP intervention on pancreas divisum, subject to the above conditions in #2, but regardless of sample size.

In addition, there was an absence of direct comparative data for ERCP-guided fine needle aspiration (FNA) and EUS-guided FNA. Thus, an indirect comparison of single-arm studies was attempted. Studies of EUS-FNA that included at least 25 subjects for the evaluation of suspected pancreaticobiliary malignancy were identified and included.

Outcomes of Interest

For diagnostic performance studies, the outcomes of interest include:
Test performance characteristics (sensitivity, specificity) as well as predictive values in diagnosing clinically relevant findings.

For therapeutic outcome studies, the primary outcomes of interest include:
1. Measures of technical success (e.g., removal of stone, relief of obstruction, cyst drainage, need for repeat procedure or placement of stent)
2. Measures of clinical success (e.g., survival, quality of life, performance scores, relief of jaundice, relief of infection, symptom scores, or pain scores)
3. Resource utilization (e.g., hospitalization, perioperative care, return to work, intensity of post-procedure care)
4. Procedure-related morbidity (e.g., stent-related problems, cholangitis, sepsis, sedation-related outcomes, bleeding, perforation, pancreatitis, long-term effects of sphincterotomy, mortality)

For Part V:

Measures of relative risk or predictive value associated with patient, procedure, or operator factors associated with ERCP complications.

For Part I, Section III:

Test performance characteristics (sensitivity, specificity) and predictive values in predicting the presence or absence of common bile duct stone(s).

Methods of the Review

Article Selection

Selection of articles was a two-stage process. All abstracts retrieved by the two search strategies were reviewed. First, titles and abstracts were reviewed using the primary and secondary study selection criteria. A single reviewer marked each citation as either: (1) eligible for review as full-text articles; (2) ineligible for full-text review; or (3) uncertain. Studies were excluded at this stage only if information revealed in the abstract showed that the study did not meet selection criteria. A second reviewer reviewed all citations marked as uncertain by the first reviewer, and a consensus decision was reached.

Using the primary and secondary study selection criteria, a single reviewer then reviewed the full-text article and determined whether selection criteria were met. The reviewer marked each full-text article as either (1) included in systematic review; (2) excluded from systematic review; or (3) uncertain. A second reviewer reviewed all articles marked as uncertain by the first reviewer, and a consensus decision was reached.

Records of the results of this evaluation were kept for each full-text paper retrieved including the reason for exclusion of each excluded study. Any disagreement about the inclusion or exclusion of a particular article was resolved by consultation with the Program Director or one or more members of the Technical Advisory Group.

Data Abstraction

Prior to the start of data abstraction, data elements were defined for abstraction from each selected article in consultation with the Technical Advisory Group. However, since some of the therapeutic key questions were not fully defined before articles were selected, many elements had to be defined based on the articles that ultimately met selection criteria. These data elements were abstracted from the articles that met final selection criteria. The data elements addressed:

1. Critical features of the study design (for example, patient inclusion/exclusion criteria, controlled or uncontrolled studies, randomized or non-randomized trials, number of subjects, or blinding, reference standard for diagnostic studies);
2. Treatment protocols;
3. The specified key outcomes.

For key questions assessing diagnosis, sensitivity, specificity, positive and negative predictive values, and prevalence of condition were all abstracted, including statistical analysis when available. Studies were grouped for presentation by categories according to diagnostic test, reference standard, clinically relevant patient subgroup, or other category of interest. For key questions assessing therapy, all outcomes that corresponded to the outcome categories that were specified in the protocol were abstracted, and studies were grouped by treatment alternative, clinically relevant patient subgroup, or other category of interest. Templates for evidence tables were then created in Microsoft Word.

Due to the anticipated heterogeneity in reported outcome measures, data were not abstracted into an electronic database. One reviewer performed primary data abstraction of all data elements into the evidence tables, and a second reviewer performed accuracy checks on the evidence tables. Disagreements were resolved between the two reviewers, or if necessary, consultation with the Program Director or relevant members of the Technical Advisory Group. If small differences occurred in quantitative estimates of data from published figures, the values abstracted independently by the two reviewers were averaged.

Quality Assessment

In consultation with the AHRQ Task Order Officer and Technical Advisory Group, a general approach to grading evidence on therapeutic studies developed by the U.S. Preventive Services Task Force (provided by Dr. Mark Helfand) was applied. Criteria for assessment of study quality for diagnostic tests were developed using the following as resources: Irwig, Tosteson, Gatsonis, et al. (1994) and the Cochrane Methods Working Group on Systematic Review of Screening and Diagnostic Tests (1996). Criteria for assessment of study quality for cross sectional analyses with multivariable regression analysis were developed with reference to Concato, Feinstein, Holford, et al. (1993).

The issues about reference standards are complex in this particular topic, and quality assessment did not take this into account. Instead, these issues are discussed in the "Review of Evidence" for each section (as applicable).

Quality criteria for therapeutic studies:

1. Initial assembly of comparable groups
   -for randomized controlled trials: adequate randomization, including first concealment and whether potential confounders were distributed equally among groups
   -for cohort studies: consideration of potential confounders with either restriction or measurement for adjustment in the analysis; consideration of inception cohorts
2. Maintenance of comparable groups (includes attrition, crossovers, adherence, contamination)
3. Comparable performance of and clear definition of interventions with equivalent attention and quality of care
4. Comparable measurements: unbiased, reliable, and valid (i.e. masking of treatment assignments)
5. Appropriate analysis of outcomes. Intent-to-treat analysis for randomized, controlled trials, consideration of confounding variables in nonrandomized studies. All important outcomes considered

Summary ratings of therapeutic studies based on above criteria:

Good: Meets all criteria: Comparable groups are assembled initially and maintained throughout the study (follow-up at least 80 percent); reliable and valid measurement instruments are used and applied equally to the groups; interventions are spelled out clearly; all important outcomes are considered; and appropriate attention to confounders in analysis. In addition, for randomized controlled trials, intention to treat analysis is used.

Fair: Generally comparable groups are assembled initially but some question remains whether some (although not major) differences occurred with follow-up; measurement instruments are acceptable (although not the best) and generally applied equally; some but not all important outcomes are considered; and some but not all potential confounders are accounted for. Intention to treat analysis is done for randomized controlled trials.

Poor: Groups assembled initially are not close to being comparable or maintained throughout the study; unreliable or invalid measurement instruments are used or not applied at all equally among groups; and key confounders are given little or no attention. For randomized controlled trials, intention to treat analysis is lacking.

Quality criteria for diagnostic accuracy studies:

1. Enrollment of representative subjects. Appropriate spectrum of patients, unbiased enrollment, few eligible patients not enrolled, appropriate accounting of all potentially eligible subjects.
2. ERCP interpreted independently of diagnostic alternative.
3. Diagnostic alternative interpreted independently of ERCP.

Issues regarding the suitability and interpretation of different reference standards were not abstracted as quality measures but are discussed in each section of the report as needed. Study selection criteria required use of a reference standard in order to construct a 2 X 2 contingency table for diagnostic performance operating characteristics.

Summary ratings of diagnostic accuracy studies based on above criteria:

Good: Excellent documentation of prospective enrollment, identification and accounting of eligible and enrolled patients, few exclusions. Both ERCP and diagnostic alternative interpreted without knowledge of other test.

Fair: Had fair enrollment of patients, not too many exclusions, interprets reference standard independent of diagnostic test; and a good spectrum of patients, though reported details may have been incomplete.

Poor: Studies that had fatal flaws (e.g., Uses inappropriate reference standard; diagnostic test improperly administered; biased ascertainment of reference standard; very small sample size or very narrow selected spectrum of patients) were not eligible for inclusion in this systematic review. Thus, no included studies were assigned a Poor rating.

Quality Ratings for Multivariable Logistic Regression Analysis Studies

The most relevant criteria that provided discrimination of quality differences between studies were the degree of overfitting present in the multivariable models, the nature of statistical reporting, and the use of procedures to establish internal validity. Degree of overfitting was assessed using the ratio of the number of endpoints divided by the number of candidate variables in the model. Studies were classified as: Satisfactory, ratio $\geq$ 10; Mild, ratio = 7 to <10; Moderate, ratio = 4 to <7; Severe, ratio <4. The nature of statistical reporting was considered satisfactory when the study reported both magnitude of effect estimates as well as associated confidence intervals or p-value for statistically significant findings. If either of these elements was not reported, studies were considered unsatisfactory. The degree of internal validity was

evaluated by the use of procedures (e.g., test-validation split samples or bootstrapping) to guard against overfitting the model and spurious results.

Summary ratings of multivariable logistic regression analysis studies based on above criteria:

Good: Studies use procedures to guard against overfitting the model and spurious results; degree of overfitting is not severe for at least one analysis, and statistical reporting is satisfactory.

Fair: degree of overfitting is not severe for at least one analysis, and statistical reporting is satisfactory, but no use of procedures to guard against overfitting the model and spurious results.

Fair Minus: severe degree of overfitting for all analyses

Technical Assistance and Peer Review

The development of the evidence report was subject to extensive expert review including input from the Technical Advisory Group (TAG), the panel of designated peer reviewers, and the Medical Advisory Panel of the Technology Evaluation Center of the Blue Cross and Blue Shield Association.

The Technical Advisory Group (TAG) included the panel chairperson for the NIH State-of-the-Science conference, Sidney Cohen, MD, who is a gastroenterologist and Professor of Medicine at Jefferson Medical College, and two gastroenterologists with expertise in ERCP, Glen Eisen, MD, MPH, Associate Professor of Medicine/Gastroenterology at Vanderbilt University Medical Center, and Michael Kimmey, MD, Professor of Medicine, Division of Gastroenterology, University of Washington. TAG members provided on-going guidance and review on all phases of this project including review of the draft report.

The draft report was also reviewed by a panel of external peer reviewers that included experts in gastroenterology, surgery, radiology, and oncology. Comments were elicited from external peer reviewers using a structured comment form, compiled, and submitted with description of disposition of comments to the Agency for Healthcare Research and Quality. (Appendix B lists the members of the Technical Advisory Group and external expert reviewers).

In addition, two sections of the draft report were reviewed by the Blue Cross and Blue Shield Association Technology Evaluation Center (TEC) Medical Advisory Panel (MAP). This interdisciplinary panel comprises experts in technology assessment methods and clinical research, and also includes managed care physicians from Blue Cross and Blue Shield and Kaiser Permanente health plans.