Chapter 2. Methodology

In this chapter, we describe the basic methodology used to develop the evidence report, from topic assessment and refinement through the literature search, screening, and data abstraction process. Included are descriptions of the literature search strategies and results, literature sources, screening and grading criteria, quality control procedures, and supplemental data sources.

Topic Assessment and Refinement

A national advisory panel of technical experts was convened to work with the Duke research team. The 11-member panel included representatives from obstetrics-gynecology, including maternal-fetal medicine; pediatrics; childbirth education; and midwifery. In addition to the American College of Obstetricians and Gynecologists (ACOG), other major interest organizations represented on the panel included the American College of Nurse Midwives and the Adolescent Pregnancy Prevention Coalition of North Carolina.

Prior to our first conference call, the advisory panel and the Task Order Officer at the Agency for Healthcare Research and Quality (AHRQ) received a document that summarized the incidence and prevalence of prolonged pregnancy, described the characteristics and size of the affected population, identified the most affected practice settings and providers, specified the interventions to be considered, and presented a diagram of the conceptual model/causal pathway. The panel also received the four key questions specified in the task order. Based on Duke's preliminary assessment of the literature and discussion with the advisory panel and AHRQ Task Order Officer, all parties agreed to refine the key questions as follows:

1. What are the test characteristics (reliability, sensitivity, specificity, predictive values) and costs of measures used in the management of prolonged pregnancy to assess: (a) risks to the mother and fetus of prolonged pregnancy and (b) the likelihood of a successful induction?

2. What is the direct evidence comparing the benefits, risks, and costs of planned induction versus expectant management at various gestational ages?

3. What are the benefits, risks, and costs of currently available interventions for induction of labor?

4. Are the epidemiology and outcomes of prolonged pregnancy different for women in different ethnic groups, different socioeconomic groups, or in adolescent women?

In addition to reaching consensus on the key questions, the advisory panel agreed on the patient population, practice settings, and target audiences of the report, as described in Chapter 1 of this report. The causal pathway is represented in Figure 1.

Literature Search and Selection

The comprehensive review of the literature, from identification of databases through abstraction of individual articles into evidence tables, was a multi-step, sequential process.

Literature Sources

The primary sources of literature were six of the most widely used computerized bibliographic databases: MEDLINE (1980-December 2000), HealthSTAR (1980-December 2000), CINAHL (1983-December 2000), the Cochrane Database of Systematic Reviews (CDSR) (Issue 4, 2000; Issue 1, 2001; and Issue 2, 2001), the Database of Abstracts of Reviews of Effectiveness (DARE), and EMBASE (1980-Jan 2000). Searches of these databases were supplemented by secondary searches of reference lists in all included articles, especially Cochrane review articles, and scanning of current issues of journals not yet indexed in the computerized bibliographic databases. Titles regularly scanned included the *American Journal of Obstetrics and Gynecology*, the *British Medical Journal*, the *British Journal of Obstetrics and Gynaecology*, the *European Journal of Obstetrics and Gynecology and Reproductive Medicine*, the *International Journal of Gynecology and Obstetrics*, the *Journal of the American Medical Association*, the *Journal of Maternal-Fetal Medicine*, the *Journal of Obstetrics and Gynaecology*, *Obstetrics and Gynecology*, the *Lancet*, and the *New England Journal of Medicine*. Suggestions regarding search terms and specific articles were solicited from the advisory panel during two conference calls in December 2000 and March 2001 and resulted in additions to the literature database.

Search Strategy

We developed the basic search strategies using the National Library of Medicine's MeSH key word nomenclature developed for MEDLINE. The same strategies were used to search HealthSTAR and CINAHL. A Duke University Medical Center librarian checked the strategies and assisted with their translation to the key word structure used by EMBASE. Dr. Evan Myers searched the CDSR and DARE using "postterm pregnancy," "prolonged pregnancy," and similar terms.

The initial searches were performed in MEDLINE and then duplicated in other databases. All searches were limited to articles published since 1980, in the English language, and with human subjects. The cut-off threshold of 1980 was based on the general unavailability of ultrasound prior to that date. It was judged that trials conducted and published prior to 1980 would be problematic both in terms of the accuracy of diagnosis and comparability with current testing and management strategies. The decision to restrict the literature search to articles published since 1980 was agreed to by the members of the advisory panel.

The search strategies are reproduced in Tables 2 and 3.

Screening Criteria

Inclusion and exclusion criteria were developed for the literature searches so that the yield of articles would be appropriately focused. Empirical studies or review articles were excluded after screening based on the following criteria:

- Article was not original research.
- Article did not address prolonged pregnancy.
- The study design was a single case report.

- The study design was a small case series with fewer than 20 subjects.

Each screened article was coded as addressing one of three topic areas:

1. Testing: Two or more tests were compared in terms of the accuracy or agreement of test results or the test result was correlated with some health outcome.

2. Management: The article addressed the relative effectiveness of planned induction versus expectant management or the relative effectiveness of an induction agent.

3. Testing and management: Some combination of the above.

The criteria used to include articles were:

- The study population must address prolonged pregnancy; ideally, results should be reported separately for patients with prolonged pregnancy. Because it is possible that the response of the cervix and uterus to induction agents would be quite different in different clinical scenarios (both in terms of labor patterns and potential maternal and fetal side effects), studies of induction agents that did not include any otherwise healthy women with prolonged pregnancy were excluded.

- All original research or relevant reviews must relate to at least one of the four key questions described above.

- Outcomes were included if they were health outcomes or health services use or economic outcomes related to the management of prolonged pregnancy.

- We included only randomized controlled trials (RCTs) which used active or nonactive (i.e., placebo) controls for studies involving management topics. For testing articles, we included RCTs and those cohort and large case series that allowed construction of 2-by-2 tables for estimation of sensitivity and specificity. Articles that did not meet these criteria were not necessarily excluded from the review and often provided valuable background material. However, only articles meeting the inclusion criteria were formally abstracted into evidence tables.

Included study designs were determined by the article's topic area. Study designs initially included for testing articles and testing and management articles were case reports; small case series (< 20 subjects); medium to large case series (≥ 20 subjects); nonrandomized comparison studies (cohort or case series that used historical or concomitant nonrandomized controls); and RCTs. The study design of each screened article was coded in our literature database.

For the testing articles and testing and management articles, an evidence table entry was developed for each RCT and for each cohort study or large case series for which a 2-by-2 table linking test results to important outcomes could be constructed (Evidence Table 1). The only study design considered for management articles was the RCT. Our experience in past evidence report projects in which lack of data from RCTs necessitated the evaluation of nonrandomized studies has been that drawing inferences about the effectiveness of therapeutic interventions based on

nonrandomized studies is difficult, if not impossible, because of numerous biases and lack of consistency in data provided about important confounding variables. An evidence table entry was developed for each included management trial (Evidence Tables 2 and 3).

Screening Results

The literature searches yielded 701 English-language articles. A summary of the number of articles retrieved from each data source is provided in Table 4. The titles and abstracts of these articles were reviewed against the inclusion/exclusion criteria by seven investigators, Drs. Richard Blumrick, Elizabeth Livingston, Andrea Lukes, David Matchar, Douglas McCrory, and Evan Myers and a third-year medical student, Ms. Andrea Christian. Two investigators reviewed each citation. Abstracts were available for more than three-fourths of the citations; when no abstract was available, the title and source were screened. At this stage, articles were included if requested by one member of the review team. The full text of each article passing the title-and-abstract screen was retrieved from the library for further review.

At the full-text screening stage, each article was independently reviewed by two investigators, who forwarded their decisions to Ms. Jane Kolimaga, the task order manager, for recording and comparison. If indicated, reviewers were asked to reconcile differences of opinion. Overall, the teams initially disagreed on about 25-35 percent of their decisions, and all disagreements were resolved by consensus. In the event that two investigators could not agree, Dr. Evan Myers, the principal investigator, was to be the arbiter, but this situation never arose.

The task order manager coded the records in the bibliographic database at each screening stage. A summary of the results of the title-and-abstract and full-text screenings is provided in Table 5.

Data Abstraction

Teams of two investigators performed the data abstraction for eligible articles identified at the full-text screening stage: one performed the primary data abstraction, and the second “over-read” the abstracted information. A data abstraction form was developed prior to initiation of the formal abstraction process. During the development of the form, draft forms were reviewed by the investigators and Dr. Rebecca Gray, a nonclinician abstractor/editor, for clarity and completeness; as the person who converted the abstraction forms into evidence tables, Dr. Gray helped to insure that all relevant information was captured. The two final iterations of the form were pretested by the investigators who used them to abstract relevant data from a sample article. The information from the data abstraction form was then summarized in evidence table format by Dr. Gray. The data abstraction assignments were made by Dr. Myers based on the investigators’ clinical interests (e.g., management vs. testing). Copies of the data abstraction form and the evidence table template are provided in Appendixes 1 and 2, respectively.

Outcomes recorded included:

- Direct health outcomes:
  - Maternal mortality.
  - Perinatal mortality.

- Maternal morbidity (specific measures varied between studies; included infection, hemorrhage, perineal trauma, etc.).
- Perinatal morbidity (meconium aspiration, postmaturity syndrome, shoulder dystocia, brachial plexus injury, admission to neonatal intensive care unit).

♦ Surrogate measures:

- Neonatal umbilical artery pH, Apgar scores, meconium-stained amniotic fluid, nonreassuring fetal heart rate tracing.
- Cesarean section rates, overall and by specific indication.

♦ Resource use:

- Costs.
- Time to delivery, proportion of vaginal deliveries within a prespecified time.

♦ Test operating characteristics:

- Sensitivity, specificity, positive and negative predictive values for outcomes listed above.

Quality Scoring

We evaluated each study included in the evidence tables for factors affecting internal and external validity. For management articles, the elements of the quality scale were as follows:

♦ Were patients randomly assigned to the intervention?

♦ Was the method for randomization described, and if so, was it one shown to be associated with less bias (sealed envelopes) than others (alternating date or medical record number)?

♦ Was the patient population similar to the likely patient population?

♦ Were the intervention protocols clearly described or referenced?

♦ Were the criteria used to make management decisions associated with primary outcomes (such as cesarean section) described?

♦ Statistical issues: Were sample size and power issues discussed? Were the statistical tests used appropriate for the types of data analyzed?

♦ Was the study population described in terms of:
  - Gestational age?

– Criteria used to assign gestational age?
– Bishop score or other measure of cervical ripeness?

For testing articles, we used the above criteria plus:

- Was an implicit or explicit reference standard defined?
- Was the issue of possible verification bias (patients with positive test results more likely to receive the reference standard test or treatment) addressed?
- Test reliability/variability: Was inter- or intrarater reliability of the test addressed?
- Was the study population well characterized in terms of the absence of risk factors such as diabetes, hypertension, etc.?
- Was the testing protocol described in sufficient detail to allow others to replicate it?

Scores on individual quality criteria were not aggregated into an overall score but were considered and reported individually. We preferred this approach for several reasons:

1. Previous work has shown that aggregated numeric scoring systems may not discriminate well between "high" and "low" quality studies, even for randomized trials (Jüni, Witschi, Bloch, et al., 1999; Moher, Jadad, and Tugwell, 1996).
2. Development and use of a new quality score would have required additional work for validation.
3. Identification of specific weaknesses in each study will be helpful in identifying trends, which in turn will assist with our recommendations for future research.

Our approach of describing key design components, rather than assigning a single aggregate score, is also consistent with recent recommendations from an expert panel on meta-analysis of observational studies (Stroup, Berlin, Morton, et al., 2000) and a recent review of the methodology of systematic reviews (Jüni, Altman, and Egger, 2001).

Summaries of the quality evaluation are provided in the evidence table entry for each abstracted article. A "+" indicates that a given criterion was met, a "-" signifies that the criterion was not met. The "+" and "-" notations were assigned by the primary abstractor and confirmed by the over-reader.

Quality Control Procedures

We employed quality-monitoring checks at every phase of the literature search, review, and data abstraction process to reduce bias, enhance consistency, and check the accuracy of screening:

- Medical librarian review of the literature search strategy.

- Review of literature search strategies by the advisory panel of technical experts.
- Check on completeness of the literature search results through reference list checks by the screener of each article.
- Reconciliation of all differences of opinion by reviewers on all full-text articles.
- Agreement of two reviewers for all eligible studies.
- Data abstractions completed by one investigator and reviewed (over-read) by another.
- Additional checks of evidence table entries for completeness and accuracy by a nonphysician abstractor.
- Solicitation of advice at key decision points from the advisory panel of technical experts.

Supplemental Data Sources

In order to get additional information about possible racial and socioeconomic differences in the incidence and outcomes of prolonged pregnancy, we analyzed data from the 1997 Nationwide Inpatient Sample (NIS) (Nationwide Inpatient Sample [NIS], 1997). The NIS is part of AHRQ's Healthcare Cost and Utilization Project (HCUP) and collects discharge data from a stratified sample of approximately 20 percent of U.S. hospitals. Using ICD-9 codes, we divided all deliveries into "preterm" (644.2x), prolonged (645.x), and term (all other delivery codes). We examined differences in outcomes between coded ethnic groups (white, black, Hispanic, Asian/Pacific Islander, Native American, and "other") and by insurance status (Medicare, Medicaid, private/health maintenance organization, self-pay/no insurance, "no charge," and "other") within these categories.

Supplemental Analyses

At the start of every evidence report project, we evaluate the feasibility of and need for meta-analyses, decision analyses, cost-effectiveness analyses, or a combination of all three. A decision about whether to proceed with such analyses is made based on the key questions and the state of the literature, after discussion with AHRQ and the advisory panel. We decided not to perform any supplemental analyses for this report for the following reasons:

- Studies of diagnostic and screening tests were too heterogeneous in terms of outcomes assessed to allow meaningful combination.
- Studies of individual induction agents did not provide sufficient specific information on women in the population of interest. As with diagnostic test studies, there was considerable heterogeneity in terms of outcomes reported.

- We did not identify any significant trials comparing induction to expectant management published subsequent to the most recent Cochrane review (Crowley, 2000). We also did not identify any disagreements with the methods or conclusions of that meta-analysis that were significant enough to justify repeating the analysis.

- Lack of adequate cost data precluded cost-effectiveness analysis.

- Although a decision-analytic model would be an excellent method for exploring the tradeoffs involved in decisionmaking for management of prolonged pregnancy, the considerations discussed above meant that there would be considerable uncertainty surrounding key parameter estimates. While development of such a model even in the setting of widespread uncertainty has considerable value, our past experience with exploratory models in situations where the literature had similar limitations has been that they are of somewhat limited value in further explaining the specific findings of the report.

The approach used by the Cochrane Collaboration differs from ours primarily in the consistent use of meta-analytic techniques to provide summary estimates of the effectiveness and risks of interventions considered. As stated above, we concluded that the state of the literature either could not support meaningful quantitative synthesis relevant to the specific patient population being considered, or that repeating an already well-done meta-analysis (Crowley, 2000) would not be worthwhile. Where relevant Cochrane reviews exist, we have compared their findings and conclusions with our own. Any differences between our findings and Cochrane analyses may represent different inclusion/exclusion criteria, different patient populations considered, or differences in outcomes considered. We have attempted to identify these potential sources of disagreement wherever possible.