Chapter 4. Conclusions

In this section we summarize the main findings of the report and discuss the implications of the findings, the limitations of the current literature, the limitations of the report, and suggested strategies for using the report to develop quality improvement tools.

Summary of Findings

The major findings and conclusions for each of the four key research questions are 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 (a) assess risks to the fetus and mother of prolonged pregnancy, and (b) assess the likelihood of a successful induction of labor?**

Consistently, tests for the assessment of risks to the fetus have lower sensitivity than specificity but higher negative predictive values than positive predictive values. This implies that the low risk of adverse outcomes is the main "driver" of high negative predictive values, and if sensitivity and specificity do not change appreciably with gestational age, that negative predictive value—the likelihood that a fetus with a normal test will have a normal outcome—decreases with advancing gestational age. Thus, false negative results will increase with advancing gestational age.

The most sensitive tests to assess the risks to the fetus of prolonged pregnancy appear to be combinations of fetal heart rate monitoring and ultrasonographic measurement of amniotic fluid volume. Direct comparison of test results across studies is difficult because of differences in patient populations and reference standards used. Published data on costs were not available.

Both ultrasound and clinical examination can be reasonably sensitive at identifying macrosomic fetuses when macrosomia is defined as greater than 4,000 grams. However, prediction of birthweights greater than 4,500 grams, the clinically more relevant threshold, is less accurate, with sensitivity ranges from 14-99 percent. There is no evidence that early detection of macrosomic infants in prolonged pregnancy improves maternal or neonatal outcomes, and modeling studies suggest that the use of ultrasound to screen for macrosomia is not cost effective.

The components of the cervical examination used to determine the Bishop score have significant inter- and intraobserver variability. The uncertainty created by this variability affects the ability of the examination to discriminate between patients likely to have a successful induction and those likely to fail.

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

Although individual randomized trials do not show significant differences in perinatal mortality between women electively induced at specific gestational ages and women followed with antepartum testing, pooled data show a significant reduction in perinatal mortality in women electively induced after 41 weeks compared with women managed with antepartum

testing. At least 500 inductions are needed to prevent one perinatal death. Cesarean section rates do not appear to differ between electively induced and expectantly managed women, either overall or in specific subgroups. In some groups, elective induction actually decreases the overall risk of cesarean section. Other maternal and perinatal outcomes do not appear to differ between groups.

Data on patient preferences for management options are lacking. Analysis of costs in the largest trial suggested that costs were reduced with elective induction; more detailed analysis based on currently used interventions and current obstetric management is needed.

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

The majority of studies of interventions for induction of labor involved women induced for a variety of indications at a wide range of gestational ages. Whether summary results from these groups are applicable to women with prolonged pregnancy is unclear.

Sweeping or "stripping" of the membranes at 38-40 weeks consistently promotes spontaneous labor and reduces the number of women requiring induction at 41 or 42 weeks.

Many studies of agents for induction are powered based on detecting differences in time to induction or differences in the proportion of women delivered within a predetermined period of time. Most do not have sufficient power to detect differences in categorical outcomes, such as cesarean section rates and adverse maternal or perinatal outcomes.

There is a consistent pattern of tradeoffs between efficacy of interventions for induction, especially as measured by time to induction or delivery within a predetermined period of time, and uterine hyperactivity, with possible increased risks of surrogate markers of fetal compromise, such as nonreassuring fetal heart rate tracings. Misoprostol appears most consistently to result in vaginal delivery within a predefined time period; however, it also appears most likely to result in very frequent uterine contractions, which may lead to fetal heart rate abnormalities.

Data are lacking on both medical and nonmedical costs of different intervention strategies.

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

We identified no published literature that showed differences among important ethnic, socioeconomic, or other subgroups.

Review of administrative data suggests that the proportion of all pregnancies extending beyond 42 weeks is similar among all racial and ethnic groups. Black women are more likely to have low birthweight infants after 42 weeks than other groups, a finding similar to observations at other gestational ages. Confirmation of these observations with more detailed data sets is needed.

Currently available literature on interventions in prolonged gestation does not address issues such as access to care or practical difficulties (for example, transportation or arranging child care) which might affect effectiveness (as opposed to efficacy) in different populations.

Research Implications

The primary research implication of our review of the literature is that much remains to be learned about the optimal management of pregnancy in women who go beyond 40 weeks gestation with otherwise normal pregnancies. It is clear that the risks of adverse outcomes increases with advancing gestational age, but the point at which this risk justifies more intensive interventions is unclear. Currently available antepartum testing strategies have good negative predictive value but poor positive predictive value. This appears to be largely due to the overall low absolute risk of adverse outcomes, since test specificity is generally better than sensitivity. The optimal test or combination of tests and the optimal timing of test initiation among women in the United States that would minimize the risk of complications associated with prolonged gestation and complications of interventions at an acceptable cost are unclear. Several interventions are available for the effective induction of labor; however, the populations studied in the published literature are heterogeneous in terms of indications for induction. Whether the benefit/risk profile of this diverse population is equivalent to that in women induced solely because of prolonged gestation, or because of abnormal antepartum testing in prolonged gestation, is unclear. Pooled results from randomized trials comparing scheduled induction and expectant management with antepartum testing show a reduced risk of perinatal mortality in women with scheduled induction after 41 weeks, with at least 500 inductions needed to prevent one death. However, the cost-effectiveness of these strategies needs to be compared using more recent data. Administrative data suggest that there are racial and ethnic differences in the epidemiology and outcomes of prolonged pregnancy; these differences need to be explored using more detailed data sets. Finally, given the complexity of decisionmaking in settings where there often are competing risks between mother and fetus, and where patients clearly have strong preferences for the process of labor and delivery, the lack of scientific data on patient preferences, quality of life, and other “subjective” measures is impressive.

Limitations of the Current Literature

Although there are a large number of randomized trials available that provide evidence addressing the key questions identified in this report, there are numerous limitations to the current literature:

- Heterogeneity of patient populations: A consistent problem with much of the literature on specific intervention agents is inclusion of women being induced for a variety of indications. Both the benefits (in terms of successful induction) and risks (in terms of fetal compromise) of induction agents might be quite different in different populations of patients. Studies either should be performed exclusively in patients with prolonged pregnancy, or subgroup analyses should be reported so that pooled estimates of efficacy in different populations can be generated.

- Appropriate endpoints: Stillbirth is, fortunately, a rare outcome even in “high-risk” populations. Most feasible studies of tests or interventions will not have sufficient power to detect differences in mortality rates. However, the clinical utility of commonly used endpoints is compromised because of inherent unreliability and susceptibility to bias (changes in fetal heart rate pattern or cervical examination), uncertainty about long-term

clinical significance (presence of meconium in amniotic fluid or Apgar scores), and the effect of variability in knowledge of preintervention test results or local practice patterns (cesarean section rates). Finally, the lack of data on patient preferences and quality-of-life measures is striking.

- Statistical issues: Even well-done studies with a priori sample size estimates often are underpowered to detect potentially clinically relevant differences in outcomes, especially when sample size estimates are based on continuous variables (such as time to delivery) and other outcomes are categorical (such as cesarean section rates). Inappropriate measures of central tendency and statistical tests are often used (for example, treating variables such as Bishop score or parity as continuous variables). This may also lead to erroneous conclusions about differences between groups.

Limitations of the Report

Literature Search

We used standard methods for identifying, reviewing, and abstracting published studies focused on the management of prolonged pregnancy. We used predefined study characteristics to identify those studies most likely to provide unbiased estimates of efficacy and test performance. We did not search the literature prior to 1980, primarily because we assumed that the lack of general availability of ultrasound for both dating and management of prolonged gestation would limit the applicability of these results to current practice. We also limited our search to articles published in English, primarily for reasons of convenience and resource constraints. It is possible that including older studies, or studies published in other languages, would have identified additional evidence that would have substantially changed our conclusions. This may be especially true for alternative or complementary therapies.

Another limitation of our exclusion criteria is that rare but severe complications of treatments may have been overlooked because they were published in case reports or small case series. Although these study designs are useful for identifying potential problems, it is difficult to quantify these risks when only numerator values are available.

Grading of Articles

We did not use one of the currently available quality scoring systems to grade the articles we reviewed. However, we believe that the rationale for each criterion we used is reasonable, and that the operational definitions are clear and reproducible. In addition, we used these grading criteria primarily to provide additional detail to other researchers. We did not use them to establish a threshold for including or excluding articles or to weight the results of a quantitative evidence synthesis such as a meta-analysis.

Other Data Sources

We used one additional data source in preparing this report, the Nationwide Inpatient Sample (NIS) (Nationwide Inpatient Sample [NIS], 1997). The NIS, like most administrative databases, is limited by a lack of clinically relevant detail. In addition, even the data recorded in these discharge abstracts were incomplete, limiting our ability to analyze them in great detail. Variability in definitions between hospitals also may lead to incorrect conclusions. The primary value of these data in the context of this report is to identify potentially important differences in outcomes between ethnic and socioeconomic groups that need to be explored further in data sets with better documentation and more complete data.

Suggested Strategies for Using this Report

The state of the currently available evidence probably does not allow for the creation of highly specific clinical guidelines or performance measures for many aspects of managing prolonged pregnancy. Consistent conclusions from the report include:

- Sweeping of the membranes consistently promotes labor. However, given the lack of data on patient preferences for undergoing this procedure or on the value of promoting labor, using performance of membrane sweeping as a quality measure is premature. However, discussion of this option with women during the late third trimester is certainly reasonable.

- Surveillance with tests that include fetal heart rate monitoring and assessment of amniotic fluid volume or elective induction both appear to be reasonable strategies beyond 41 weeks. Patients and providers should be informed that the best current evidence strongly suggests that there is a significant increase in the risk of perinatal mortality in women managed with antepartum testing compared with women who are electively induced at 41 weeks. Because this risk is small in absolute terms, and patients may have different preferences for both the outcomes and processes of labor and delivery, both options should be discussed.

- There is no evidence to justify induction of labor solely for the indication of macrosomia (defined as estimated fetal weight greater than 4,000 grams) in prolonged pregnancy.