Evaluating the Medical Evidence for Quality Improvement

Roger F. Soll, MD

Advances in neonatal care have led to significant improvement in survival and quality of life of newborn infants. That said, Neonatal-Perinatal Medicine has also participated in the introduction of untested drugs and interventions that have led to disastrous complications. The past several decades in Neonatal-Perinatal Medicine have seen their share of therapeutic misadventures: the epidemic of retinopathy attributable to the indiscriminate use of supplemental oxygen; gray baby syndrome attributable to chloramphenicol; and an increase in the incidence of kernicterus attributable to the introduction of sulfonamide drugs. To avoid these disasters, clinicians must practice medicine based on the strong support of the available evidence. This review builds on the principles of evidence-based medicine (EBM) discussed in a previous review of the subject and discusses how to practice EBM and how EBM can be used in quality improvement. Examples from Neonatal-Perinatal Medicine are given for which these principles were, or were not, adhered to.

On evaluation of the use of commonly used interventions, a tremendous variation in practice is discerned. A review of the Vermont Oxford Network database of infants weighing 401 to 1500 g demonstrates large variation in a variety of common

The author is Professor of Pediatrics at the University of Vermont College of Medicine, President of the Vermont Oxford Network, and Coordinating Editor of the Cochrane Neonatal Review Group. He has received no direct support regarding any of the interventions discussed in the article.

a Division of Neonatal-Perinatal Medicine, University of Vermont College of Medicine, Smith 552A, 111 Colchester Avenue, Burlington, VT 05401, USA
b Neonatal Intensive Care Unit, Fletcher Allen Health Care, University of Vermont College of Medicine, Smith 552A, 111 Colchester Avenue, Burlington, VT 05401, USA
c Vermont Oxford Network, University of Vermont College of Medicine, Smith 552A, 111 Colchester Avenue, Burlington, VT 05401, USA
d Cochrane Neonatal Review Group, University of Vermont College of Medicine, Smith 552A, 111 Colchester Avenue, Burlington, VT 05401, USA

* Division of Neonatal-Perinatal Medicine, University of Vermont College of Medicine, Smith 552A, 111 Colchester Avenue, Burlington, VT 05401.

E-mail address: Roger.Soll@vtmednet.org
practices. Practices such as the use of nasal continuous positive airway pressure (CPAP) and high-frequency ventilation vary greatly between centers. In very low birth weight infants, the median use of nasal CPAP is 67%, but the interquartile range is broad (52%–79%). Similar variation in practice is observed with more recent technological innovations such as high-frequency ventilation (HFV). The median use rate of HFV is 22%, but the interquartile range is 9% to 28%. If these and other practices are based on the best evidence, why is our practice so varied? And if these practices truly matter, why are some of us putting our infants at risk?

To avoid these problems, we need to learn to evaluate the medical evidence and base our practice on the best available evidence. EBM is the integration of clinical expertise, patient values, and the best evidence into the decision-making process for patient care. EBM requires personal commitment, institutional commitment, and societal commitment. These factors are reflected in personal practice choices as well as the development of guidelines that may influence practice on a local or national level. EBM cannot be expected to provide answers to all our questions, but EBM will allow us to improve the quality of care by identifying and promoting practices that work, while eliminating those that are ineffective or harmful. In his primer on EBM, Sackett and colleagues proposed 6 simple steps that are essential to practicing EBM. These steps are detailed in this article.

FORMULATING THE QUESTION

Although it seems obvious, formulating the question is critical to the practice of EBM. Questions must be searchable (ie, having a reasonable expectation that some research has attempted to answer the question) and clinically relevant. The question must be explicit regarding the patient or problem being considered, the intervention being considered, the comparison intervention, and the clinical outcome of interest. The question could deal with any aspect of care, including etiology, diagnosis, prevention, treatment, or prognosis. For the purpose of this review, the focus is on evaluating the efficacy of new therapies.

The clinical question should be structured in the PICO format (Patient or Problem, Intervention, Comparison, Outcomes). For the “Patient or Problem,” the specific characteristics of the patient (or in the case of guidelines or recommendations, the population or group of patients) in whom the intervention takes place must be defined. Issues such as the patient’s age, disease severity, or coexisting conditions need to be considered. For this discussion, the “Intervention” will involve treatment choices. Both the “Intervention” and the “Comparison” need to be specifically defined. Are you considering treatment or nontreatment, or the comparison between 2 different drugs or surgical interventions? The “Outcome” chosen should be of consequence either to the patient or society, not merely a surrogate for clinical or physiologic improvement.

For example, imagine you had a 2-week-old, 27-week gestational age infant who remained on assisted ventilation with worsening respiratory status. You wonder whether corticosteroid administration might improve your ability to extubate this infant and enhance his chances for survival. The key components for your clinical question would include:

<table>
<thead>
<tr>
<th>Patient or Problem:</th>
<th>2-week-old ex–27-week gestational age infant on assisted ventilation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervention:</td>
<td>Corticosteroids</td>
</tr>
<tr>
<td>Comparison:</td>
<td>No corticosteroids</td>
</tr>
<tr>
<td>Outcomes:</td>
<td>Ability to wean from ventilation and survival</td>
</tr>
</tbody>
</table>
The 4-part clinical question could be formulated as follows: “In a 2-week-old, 27-week gestation baby on assisted ventilation with worsening respiratory status, does the administration of corticosteroids compared with not giving corticosteroids decrease the risk of developing chronic lung disease?”

FINDING THE EVIDENCE

Once you have formulated the clinical question, the next step is to search for relevant evidence that may help answer the question. This process has changed dramatically since the advent of computer databases and the Internet. Long gone are the days of sitting in the library poring through volumes of Index Medicus. Today, there are multiple resources for access to the general medical literature and some resources that are unique to Neonatal-Perinatal Medicine. Searching the medical literature is now widely available through the Internet, including several bibliographic databases such as the Cochrane Library database, MEDLINE, EMBASE, and CINAHL. MEDLINE is the premier biomedical database produced by the National Library of Medicine. MEDLINE includes journals published from 1966 to the present, indexes over 5200 international biomedical journals, and contains more than 18 million references. A variety of MEDLINE search engines are available, including Ovid and PubMed. PubMed is freely available on the Internet through the National Library of Medicine (http://www.ncbi.nlm.nih.gov/pubmed/). Practitioners must learn effective search strategies in MEDLINE, including the use of the National Library of Medicine subject index (MeSH headings), and a variety of limits, including publication types (eg, clinical trial, letter, review) and text words to create the most efficient literature searches. Haynes and colleagues10 offer simple strategies to locate the best studies of treatment, diagnosis, prognosis, or etiology with the greatest precision. Depending on your needs, you may want your search to be more sensitive (including the greatest number of relevant articles but also including some less relevant articles) or more specific (including mostly relevant articles but omitting a few relevant articles). Search engines such as PubMed have incorporated these “filters” into specialized search engines (refer to the “Clinical queries using research methodology filters” in PubMed).

Other databases include EMBASE and CINAHL. EMBASE (published by Elsevier Science) covers the biomedical literature from 7000 journals, and is particularly strong in pharmaceutical and toxicologic studies. CINAHL is the Cumulative Index to Nursing and Allied Health Literature, and represents the most comprehensive resource for Nursing and Allied Health Sciences publications. For a comprehensive literature search, such as the search needed to create a systematic overview, knowledge and use of these search engines is essential. However, for the busy clinician, certain shortcuts may be useful. Although not comprehensive, quick questions can be answered by simple searches through Google Scholar or other readily available search engines.

There are a variety of other resources that may be important in the practice of EBM, particularly in Neonatal-Perinatal Medicine. The Cochrane Library of Databases (which includes the Cochrane Database of Systematic Reviews, the Databases of Abstracts of Reviews and Effectiveness, and the Cochrane Control Trials Registry) is maintained by the Cochrane Collaboration, an international initiative that designs, prepares, maintains, and disseminates systematic reviews of health care interventions. The Cochrane Review Group creates systematic overviews for the Cochrane Library.11 These reviews are available through the Cochrane Library or through the National Institute of Child Health and Human Development (NICHD) Web site (http://www.nichd.nih.gov/COCHRANE/).
CONDUCTING THE SEARCH

Each database has its own specific way to search the literature. A familiarity with the most basic portals and databases (e.g., using PubMed to search MEDLINE) is critical for success. To successfully pair down the search, an understanding of the use of the Boolean operators “AND” and “OR” is essential. If you are combining 2 terms “AND” allows only articles containing both terms to be retrieved while “OR” allows articles containing either term to be retrieved. In addition, certain specific limits (such as publication type, publication dates, or study population) may be very useful in refining your search.

Given the PICO question asked earlier, a simple search of MEDLINE using PubMed might include the following: chronic lung disease AND corticosteroids OR steroids OR glucocorticoids OR prednisone OR dexamethasone. Limits might include “clinical trial” or “randomized controlled trial” and an age group, for example, “all infants.” More sophisticated searches can be done, but these require greater expertise and, frequently, the help of a research librarian.12

APPRAISING THE EVIDENCE

Once you have located potentially relevant articles through your search, you must learn how to evaluate the available evidence. Although all the articles may contain some useful information, all are not equally valid. The quality (also known as the strength or validity) of the evidence ranges from extensively researched issues using methodologically sound study designs (such as systematic reviews or large, well-designed randomized controlled trials [RCTs]) to the opinion of respected authorities (based on clinical evidence, descriptive studies, or reports of expert committees). Many groups have attempted to help practitioners and guideline developers evaluate the strength of the evidence. One useful approach on which to base the “validity” or “strength” of the available evidence is the hierarchy of evidence created by the US Preventive Services Task Force (http://www.ahrq.gov) (Box 1).13

You may not always find the highest level of evidence to answer your question. In the absence of the best evidence, you then need to consider moving down the pyramid to other types of studies. The Task Force has placed the “opinions of respected authorities” at the base of the pyramid, implying that this represents the lowest level of evidence. While that may be true, it must be remembered that many

---

**Box 1**

Judging the strength of the evidence

I: Properly powered and conducted RCT; well-conducted systematic review or meta-analysis of homogeneous RCTs

II-1: Well-designed controlled trial without randomization

II-2: Well-designed cohort or case-control analytical study

II-3: Multiple time series with or without the intervention; dramatic results from uncontrolled experiments

III: Opinions of respected authorities, based on clinical experience; descriptive studies or case reports; reports of expert committees

of the most important breakthroughs in medical care began with the observation and opinions of experts. The often quoted story of Sir Alexander Fleming’s observation that a substance produced by the fungus Penicillium notatum had bactericidal properties led to the eventual discovery of the antibiotic penicillin. However, even as great an idea as a new antibiotic needs to be vigorously tested, as demonstrated by the disastrous introduction of sulfonamides in newborns leading to an increase in the incidence of kernicterus.

Other study designs contribute to our understanding of new interventions. Case series and case reports consist of reports on the treatment of an individual patient or groups of patients. In the hierarchy of the US Preventive Task Force pyramid, these studies would be classified as II-3. No statistical comparisons can be done because these are only reports of cases that do not use a control group for comparison. That is not to say that important scientific insight has not been gained from such studies. If a specific new intervention is associated with a bizarre and previously unseen complication (eg, the effects of thalidomide when given to pregnant women), such reports provide essential and actionable information.

Well-designed cohort or case-control analytical study are combined in the Task Force pyramid (classified as II-2). Case-control studies are studies in which patients who have a specific condition are compared with people who do not have the condition. The evaluation is fairly straightforward; individuals who have the outcome of concern are compared with individuals who have not experienced that outcome to reveal the frequency that each group has been exposed to the intervention in question. Cohort studies compare 2 “cohorts” of patients from the same time period, one of whom has received a particular treatment and the other who has not. The classic example of a cohort study in perinatal medicine is the report of decreased mortality in women cared for by the midwifery service at the Second Obstetric Clinic of the Vienna General Hospital, associated with hand-washing practices. Although potentially useful, cohort studies are prone to bias. Unlike RCTs, the groups may not be similar and may differ in ways other than exposure to the intervention under study.

The most valid evidence comes from RCTs. The methodology of RCTs seeks to minimize bias at all points of the study and thereby gives the most accurate and reproducible estimates of effect. Random allocation of study subjects is essential to minimize bias at the time of study entry (selection bias) and provides the basis for all traditional statistical comparisons used in the analysis of trial results. Bias can also occur after patient allocation. Bias can occur regarding the exposure to the intervention (performance bias), completeness of follow-up (exclusion bias), and measurement of outcomes (detection or assessment bias). RCTs are limited in their ability to evaluate the long-term consequences of therapy and have no use in evaluating issues, such as complex processes of care or environmental issues, as seen with exposure to chemical or industrial hazards. Even with these limitations, it is critical that clinicians familiarize themselves with the methodological issues involved in the proper conduct of RCTs to appreciate the validity of the evidence.

Methodological quality of RCTs does impact on the interpretation of results. Both groups in the trial must be treated the same except for administration of the experimental treatment. If “cointerventions” (interventions other than the study treatment) exist, they should be applied in a similar fashion to both groups. In trials in which treatment allocation has been inadequately concealed, treatment effects are commonly overestimated. In trials without a placebo, there may be greater reported effects. All subjects entered into a trial should be accounted for at the end of the trial (intention
to treat analysis). Failure of a subject to receive the assigned intervention may be due to the difficulty of delivering the intervention (eg, immediate intubation and surfactant treatment in the delivery room) or patient noncompliance (eg, refusal to take a foul tasting medicine or stopping taking a medicine because the patient perceives no benefit). These situations will skew the results of the study by leaving subjects in the study that had a more positive response to treatment. Similarly, loss to follow-up may skew results. Subjects who do not make themselves available for follow-up could theoretically be quite different from those that do (either having experienced a poor response or adverse event, or perhaps because they are completely cured and see no reason to seek further medical advice). All of these issues must be formally addressed to understand the validity of the trial (Box 2).8

<table>
<thead>
<tr>
<th>Box 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessing the validity and importance of a clinical trial</td>
</tr>
</tbody>
</table>

**Are the results of this therapy study valid?**

1. **Was the assignment of patients to treatment randomized?**

   The assignment of patients to either group (treatment or control) must be done by in a random fashion; this might include a coin toss (heads to treatment/tails to control) or use of randomization tables, often computer generated.

2. **Were all the patients who entered the trial properly accounted for at its conclusion? Was follow-up complete?**

   All patients who started the trial should be accounted for at the end of the trial. If patients are not accounted for, the validity of the study may be jeopardized. A good study will have better than 80% follow-up for their patients. Patients may drop out of a study for various reasons, including because of adverse events related to the treatment. If these patients are not included in the results, they can make the treatment look better than it really is (and vice versa).

3. **Were patients analyzed in the groups to which they were (originally) randomized?**

   Patients should be analyzed within their assigned groups. This is called “intention to treat” analysis. Patients who are noncompliant with treatment or follow-up should not be eliminated from the study analysis. Excluding noncompliant patients leaves behind those that may be more likely to have a positive outcome, creating bias.

4. **Were patients, clinicians, and study personnel “blind” to treatment allocation?**

   “Blinding” or “masking” means that the study subject and study personnel do not know which treatments were received. Blinding eliminates bias and any preconceived notions as to how the treatments should be working. When it is difficult or unethical to blind patients to a treatment, such as a surgical treatment, then a “blinded” researcher is needed to interpret the results.

5. **Were the groups similar at the start of the trial?**

   Based on random assignment, the treatment and the control group should be similar for most prognostic characteristics except for whether they received the experimental treatment. If there are known factors that highly influence outcome, equal distribution of these factors can be addressed by stratification at the time of randomization.

6. **Aside from the experimental intervention, were the groups treated equally?**

   Both groups must be treated the same except for administration of the experimental treatment. All “cointerventions” must be applied in a similar fashion to both study groups.
Systematic reviews are extremely useful tools to evaluate the entirety of the high-order evidence available to address a specific question. To perform a systematic review, an extensive literature search must be conducted to identify all eligible methodologically sound studies. The studies are reviewed, assessed, and the results summarized according to the predetermined criteria. The reviews published in the Cochrane Library are a prime example of this process.

Systematic reviews are potentially even more powerful than individual RCTs. Systematic reviews limit the bias inherent to traditional overviews by conducting a comprehensive search of all potentially relevant articles and using explicit, reproducible criteria in the selection of articles for review. Qualitative systematic reviews summarize the data but do not perform further statistical analyses. Quantitative systematic reviews, or meta-analyses, are systematic reviews that use statistical methods to combine the results of multiple RCTs. The statistical methods for pooling results are similar to the statistical methods used in analyzing the data from multicenter trials. Pooling the results of previous similar RCTs increases the statistical power lacking in the individual smaller trials, and enables the clinician to have greater security in accepting or rejecting treatment differences demonstrated by the trials.

Meta-analysis has its critics. Any attempt at pooling results from various studies will not only incorporate the biases of the primary studies but will also add additional bias attributable to study selection and the inevitable heterogeneity of the selected studies. LeLorier and colleagues compared the findings of 12 large RCTs with the results of meta-analyses conducted earlier on the same topics. There was only fair agreement between the meta-analyses and the large clinical trials. However, differences in the point estimate between the RCTs and the meta-analyses were statistically significant in only 12% of the comparisons. The discrepancies noted by LeLorier may be attributable to a variety of biases that may be incorporated in the meta-analysis. There are several plausible explanations of how a meta-analysis might obtain a positive result that is not confirmed by subsequent large, well-designed RCTs. Publication bias, the tendency for investigators to preferentially submit studies with positive results, and the tendency of editors to choose positive studies for publication skews the medical literature toward favorable reports of treatment. Unless the authors of the meta-analysis have done a scrupulous search of all available resources, these studies will not be located and the meta-analysis stands a good chance of reporting a falsely positive finding. This problem is further compounded by the greater chance that this false positive finding will itself be published. Meta-analysis can offer false negative conclusions because of inappropriate study selection. If the studies selected are not groupable (heterogeneous), the positive effects observed with one specific treatment or in one specific population may be lost. To minimize bias, the authors of the meta-analysis and the readers of the review must demand the same methodological quality from these analyses that they would from an individual RCT. It is essential that all meta-analyses include a prospectively designed protocol, a comprehensive and extensive search strategy, strict criteria for inclusion of studies, standard definitions of outcomes, and standard statistical techniques.

Once the evidence has been found, the importance of the evidence must be assessed. Clinical trials may use a variety of statistical techniques in reporting their results. Just because a reported difference is “statistically significant,” does not make the finding automatically clinically relevant. To assess whether the results of a trial are, in fact, clinically relevant, one needs to calculate some simple statistics from the study findings. The relative and absolute risk reductions are useful statistics in understanding the clinical impact of the therapy. The relative risk reduction (RRR) is the control event rate (CER) minus the experimental event rate (EER) over the
Relative Risk (RR) is the risk of the outcome in the treated group (Y) compared with the risk in the control group. RR = Y/X. Relative Risk Reduction (RRR) is the percent reduction in risk in the treated group (Y) compared with the control group (X). RRR = 1 – Y/X × 100% or 1 – RR × 100%.

Absolute Risk Reduction (ARR) is the difference in risk between the control group (X) and the treatment group (Y). ARR = X – Y.

Number Needed to Treat (NNT) is the number of patients that must be treated over a given period of time to prevent one adverse outcome. NNT = 1/(X – Y) or 1/ARR.

Many studies publish P values to evaluate statistical significance. In reality, all that the P value tells the reader is whether the event was likely to be random or not. Confidence intervals are more useful in quantitating the uncertainty of the estimated value of an effect measurement. A 95% confidence interval (CI) reflects 95% certainty that the true value of the measure lies within the bounds of the interval.

**APPLICABILITY**

Once the appropriate research articles are in hand and the data have been summarized in ways that are clinically relevant, it is necessary to decide how to integrate this evidence into clinical practice or, in the case of institutions, into practice guidelines. In the individual patient, one must assess whether the results of the randomized trial apply to the treatment of that particular individual. A variety of characteristics (age, sex, comorbidity, and disease severity) may be substantially different from the characteristics of the patients enrolled in the trial. These differences may make it unclear that the extrapolation of the results of the evidence to the individual patient is appropriate. A typical example would be the use of a treatment in a patient with a condition similar, but not identical, to the condition of patients treated in the trial. This situation
frequently needs to be addressed with respect to issues of timing or disease severity. In these situations, Sackett and colleagues ask us to use common sense. They pose the question in reverse, asking “Is my patient so different from those in the trial that the results cannot help me make my treatment decision?” In reality, there are few situations in which you would expect that an intervention would produce qualitatively different results in patients who do not strictly fit eligibility criteria. Only in these situations should you consider rejecting the results.

Another common mistake in applying the evidence occurs with inappropriate subgroup analyses. If there was some prior reason for expecting differential responses to the intervention among different subgroups of patients, this analysis should be prospectively planned as part of the study. If not, these analyses are nothing more than “fishing” expeditions. It is very likely that differences will emerge, but it is impossible to distinguish between real effects and chance events. Unless there is some persuasive biologic reason to believe that the treatment would be totally ineffective or detrimental to the patient (compared with patients enrolled in the study), one should assume a similar direction of effect in the patient’s illness.

**GRADE Recommendations**

Evidence must be used not only in the care of individual patients but also in the creation of clinical policies or guidelines. In creating policies or guidelines, one must consider the evidence regarding the impact of the disease, barriers to implementing the clinical policy, safety, acceptability, and cost effectiveness. Even in situations where the evidence is extensive, formulating a clinical policy can be a difficult task. Decisions regarding clinical policies and guidelines involve a series of compromises that take both the individual and societal values into account.

The process of developing guidelines lacks any standardization or consistency. Guyatt and colleagues recommend use of the GRADE system (Grades of Recommendation, Assessment, Development, and Evaluation), a method for categorizing and communicating information regarding the quality of evidence. The GRADE system has been adopted by organizations worldwide. To achieve transparency and simplicity, the GRADE system classifies the quality of evidence into 1 of 4 levels: high, moderate, low, or very low. High-quality evidence means that the current evidence is methodologically sound and provides a very precise estimate of effect, such that further research is very unlikely to change the estimate of effect; moderate quality denotes that further research is likely to have an important impact on interpretation of the estimate of effects and may change the estimate; low quality notes that future research is very likely to have an important impact on our interpretation of the estimate of effect and is likely to change the estimate; and very low quality denotes that the estimate of effect is very uncertain. There are limitations to such a simplified approach. Quality of evidence is a continuum; any discrete characterization involves some degree of arbitrariness. However, Guyatt and colleagues argue that given the difficulties in disseminating evidence-based recommendations, the advantages of simplicity, transparency, and clarity outweigh these limitations.

Making GRADE recommendations requires a sophisticated sense of the methodological issues that affect the validity of RCTs. Evidence derived from RCTs is the highest quality evidence, but as noted in the previous discussion, confidence in the evidence may be decreased for several reasons, including methodological limitations of the study, inconsistencies of results across studies, lack of precision (from appropriately small sample size), and reporting or publication bias. The GRADE system offers 2 grades of recommendations: either “strong” or “weak.” Strong recommendations can be made in situations where the risk/benefit ratio is clear (where the benefits
clearly outweigh the risks or vice versa). On the other hand, when the balance of risks and benefits are less certain, either because of low-quality evidence or because the evidence suggests that the risks and benefits are closely balanced, only weak recommendations can be made. Using the GRADE system potentially provides a system for rating quality of evidence and strength of recommendations that is explicit, comprehensive, transparent, and pragmatic.

EVALUATING PERFORMANCE: THE ROLE OF QUALITY IMPROVEMENT

In Sackett’s primer on EBM, the final step to the practice of EBM is to evaluate our own performance. Of all the steps in evidence-based medicine, this seems to receive the least attention. However, this is as important a step as there is in the practice of EBM. EBM helps us to find and understand the evidence of “efficacy” of a given intervention; the results that were found in an idealized experimental situation. If these results are valid, we need to give careful attention to how we will integrate these interventions in our practice and monitor our practice to assure that these new practices are “effective.”

The original concept of evaluating performance implied that this performance was based on an individual patient or provider. However, in Neonatology, it makes more sense to see this in the context of the complex system of the Neonatal Intensive Care Unit (NICU). The Vermont Oxford Network has applied a team approach to health care improvement, with the goal of improving the effectiveness and efficiency of neonatal care. The various Neonatal Intensive Care Quality (NIC/Q) projects have used a collaborative model of quality improvement and benchmarking that was originally developed by industry and has now been applied successfully to health care.

The major components of the NIC/Q project include:

1. Multidisciplinary collaboration within and among hospitals
2. Feedback of information from the network database regarding clinical practice and patient outcome
3. Training and quality improvement methods
4. Site visits to project NICUs
5. Benchmarking visits to superior performers within the network
6. Identification and implementation of “potentially better practices”
7. Evaluation of the results.

The original project involved ten NICUs and was funded by the Center for the Future of Children of the David and Lucille Packard Foundation, and was performed in collaboration with the Rand Corporation. In the original NIC/Q project, teams from the 10 hospitals worked together in cross-institutional improvement groups. These groups participated in an intensive series of large group meetings, site visits, and conference calls. Two major clinical quality indicators were chosen: nosocomial infection and chronic lung disease. Six NICUs focused on reducing nosocomial infection and 4 units focused on reducing chronic lung disease. The “potentially best practices” that were proposed were based on an evidence review and careful analysis of other practices at “best performing” centers. In the original NIC/Q project, reductions were seen in both chronic lung disease and nosocomial infection. In the second NIC/Q collaborative, a more formalized system for teaching quality improvement skills was put in place. Four key habits for clinical improvement became the focus of this project: the habit for change, the habit for clinical practices and processes, the habit for collaborative learning, and the habit for evidence-based practice. The openness and urgency for change (the habit for change), the belief that productive work is accomplished through
a variety of well thought out processes (the habit for clinical practice and process), and
the willingness to assist in collaborative learning (the habit for collaborative learning) all
address behavioral issues. Of importance is that the habit for evidence-based practice
is an integral part of these improvement projects. Teaching EBM, including all of the
previously discussed issues regarding finding evidence and validating evidence,
was incorporated in this stage. The success of these improvement projects is studied
in the Plan Do Study Act cycles (PDSA cycle) used to test and implement the changes.
Each individual cycle’s specific plans are developed (Plan), conducted (Do), results are
evaluated (Study), and actions are taken based on what has been learned (Act). Evidence provides the cornerstone for all of these plans. As simple as these steps
seem, organizations need training and assistance to integrate them into their routine
behavior, and these changes can lead to measurable improvement in clinical
outcomes.

Examples of EBM in Neonatal-Perinatal Medicine

When EBM tells us what to do to: antenatal steroids

The ability for EBM to influence perinatal practice is clearly illustrated in a variety of
situations. An often quoted example of this impact is the history of the adoption of
antenatal corticosteroids as standard of care for women with impending preterm
delivery. Following preclinical studies by Liggins and Howie, RCTs of antenatal corti-
costeroids undertaken during the 1970s and 1980s provided consistent evidence of
benefit for preterm infants. Despite availability of this evidence, most women with
impending preterm delivery during that era did not receive antenatal steroids and
further controlled trials continued to be approved, funded, and undertaken. A system-
atic review published by Crowley and colleagues in the 1990s confirmed that ante-
natal steroids substantially reduced the risk of respiratory distress syndrome, neonatal
mortality, and neonatal morbidity, without increasing the rate of adverse maternal
outcome. This original meta-analysis included a total of 18 RCTs (enrolling more
than 3700 infants) studying the effect of antenatal corticosteroids on promoting lung
maturity. This review has since been updated by Roberts and Dalziel in 2006 and
now includes 21 studies. In a cumulative meta-analysis, Sinclair asked the question,
“At what point in the history of trials of antenatal corticosteroids for fetal lung matura-
tion was the aggregated evidenced sufficient to show that this treatment reduces the
incidence of respiratory distress syndrome (RDS) and neonatal death?” In this anal-
ysis, trials were ordered by their date of publication and meta-analyses were per-
formed sequentially. As previously noted, the initial trial of Liggins and colleagues
demonstrated a significant reduction in the risk of RDS and the risk of neonatal death.
As each new trial was added to the cumulative meta-analysis, the risk reduction
remained statistically significant. The point estimate of the risk reduction changed little
with the addition of each new trial; however, the 95% CI narrowed, giving increased
precision to the estimate to effect. One is hard pressed to justify the need for so
many clinical trials. Despite overwhelming evidence from RCTs, use of antenatal
steroids in very low birth weight infants remained low throughout the 1980s. The
lack of acceptance of the data on antenatal steroids was, in part, attributable to inap-
propriate subgroup analyses. Clinicians were concerned that antenatal corticoste-
roids were ineffective in twin gestation, male infants, prolonged rupture of
membranes, and a variety of other clinical situations. The meta-analysis conducted
by Crowley and colleagues evaluated the effect of antenatal steroids in several of
these subgroups, and established that antenatal steroids were effective in a broad
range of clinical situations and were not affected by issues such as multiple gestation
and gender.
The evidence from the systematic review provided the cornerstone of the National Institutes of Health’s consensus statement regarding the use of antenatal corticosteroids. The consensus statement recommends the use of antenatal corticosteroids for women at risk in a broad range of gestational ages with few exceptions. Wright and colleagues demonstrated improved understanding of the impact of corticosteroids by obstetricians in the NICHD Network before and after the National Institutes of Health Consensus Development Conference and by increased use of steroids within the NICHD Network. In 1999, more than 79% of at-risk infants were treated with either a partial or full course of antenatal corticosteroids compared with 16% in 1987.

*When EBM tells us what not to do: postnatal corticosteroids*

Corticosteroid therapy may decrease lung injury after birth through a variety of mechanisms, including stabilization of lysosomal membranes, decrease in inflammatory response, and decrease in pulmonary edema. Corticosteroids potentially could improve lung function, decrease the need for supplemental oxygen, decrease the need for ventilator support, decrease lung injury and chronic lung disease, and ultimately decrease mortality.

Multiple trials have been conducted to assess the use of postnatal corticosteroids for the prevention of chronic lung disease. The best estimates of the risks and benefits of postnatal corticosteroid therapy are provided by the systematic overviews of Halliday and Erhenkranz published in the Cochrane Library. Twenty-eight RCTs enrolling a total of 3740 participants are included in this analysis. Although the trials of postnatal corticosteroid represent a heterogeneous group of studies, differing in time of treatment, steroid preparation, steroid dosage, and length of treatment, the meta-analysis provides a useful springboard to understand the strengths of weaknesses of the therapy. Interpretation of the results are further complicated by the fact that many of the trials allowed for later “rescue” treatment of control infants who had evolving chronic lung disease. This approach complicates any interpretation of the data; late or selective treatment in the control group will allow for an underestimate of both the potential benefits and the potential risks. That said, the meta-analysis of these trials clearly shows that there are strong benefits to the early use (age <8 days) of postnatal corticosteroids in preterm infants at risk of developing chronic lung disease. Fewer infants require supplemental oxygen at 28 days or at 36 weeks postmenstrual age. Although mortality alone seems unaffected by early treatment, more infants survive with chronic lung disease at 36 weeks postmenstrual age (22 studies, 3320 infants; typical RR 0.89, 95% CI 0.84, 0.95). The typical estimate of effect suggests that, for every 4 infants treated with steroids early in the course of evolving chronic lung disease, one additional infant without chronic lung disease will survive.

Although postnatal corticosteroids lead to clinical improvement, there is great concern regarding both the short-term and long-term complications of therapy. Many complications of corticosteroid therapy have been reported, including hypertension, hyperglycemia, infection, hypertrophic cardiomyopathy, gastrointestinal bleeding, gastrointestinal perforation, and decrease in somatic growth. Of most concern are the long-term problems regarding growth and neurodevelopment. Although fewer studies contribute to the meta-analysis, it is clear that there is an increased risk of cerebral palsy (12 studies, 1452 infants; typical RR 1.45, 95% CI 1.06, 1.98) and poor neurodevelopmental outcome.

The findings concerning poor neurodevelopmental outcome have led to strong statements from the American Academy of Pediatrics and the Canadian Pediatric Society. These statements recommend severe restriction in the use of postnatal corticosteroids in the prevention and treatment of chronic lung disease, noting that
outside the context of a randomized, controlled trial, the use of corticosteroids should be limited to exceptional clinical circumstances (eg, an infant on maximal ventilator support and oxygen support).” These evidence-based recommendations have led to significant change in clinical practice. The exposure of at-risk babies has decreased dramatically. In the Vermont Oxford Network, up to 28% of very low birth weight infants were exposed to postnatal corticosteroids before the Academy statement (in 1997) and only 8% were exposed after (though there is variation in practice with the interquartile difference from 2% to 11%).

When EBM provides abundant evidence but the course of action is still unclear: prophylactic indomethacin

Applying EBM to practice becomes more complicated when the value of the outcome is less clear (virtually any time the outcome is something other than mortality) and there are actual or theoretical concerns of competing risk. Such is the case regarding the use of prophylactic indomethacin in the prevention of intraventricular hemorrhage. In current practice, 26% of all very low birth weight infants experience intraventricular hemorrhage, 9% of which are of the more severe grades. Prophylactic indomethacin has been evaluated in the prevention of patent ductus arteriosus and in the prevention of intraventricular hemorrhage. Indomethacin, a cyclooxygenase inhibitor of prostaglandin synthesis, has been demonstrated to modulate cerebral blood flow, decrease serum prostaglandin levels, and promote germinal matrix maturation.

Multiple clinical trials have suggested that indomethacin lowers the risk of intraventricular hemorrhage in very low birth weight infants. Fowlie and Davis conducted a systematic review of 19 RCTs of prophylactic indomethacin involving 2872 infants. The meta-analysis suggests a decrease in the risk of patent ductus arteriosus (typical RR 0.44, 95% CI 0.38, 0.50) and severe intraventricular hemorrhage (typical RR 0.66, 95% CI 0.53, 0.82) associated with prophylactic indomethacin administration. In clinical terms, one needs to treat 5 infants with prophylactic indomethacin to prevent one patent ductus arteriosus, 12 infants to prevent one intraventricular hemorrhage, and 26 infants to prevent one severe intraventricular hemorrhage. However, use of prophylactic indomethacin is not widespread because of concern regarding possible side effects of treatment, including cerebral ischemia and necrotizing enterocolitis. In addition, the long-term effects of indomethacin are not well known.

Schmidt and colleagues conducted a large, pragmatic trial of indomethacin, which showed similar effects to those of previous smaller trials regarding the short-term outcomes. However, little difference was noted in neurodevelopmental follow-up. This result leaves families and clinicians with difficult decisions to make. This situation is potentially ideal for applying decision analysis. Decision analysis is useful in situations whereby competing risks allow for a probabilistic quantitative framework to aid in decision making. Decision analysis requires one to structure the problem, assign probabilities to chance events, assign utility or value to all outcomes, evaluate the utility of each strategy, and perform a sensitivity analysis. At each step along the way, there are a variety of assumptions. In this case, 3 possible decisions can be evaluated: prophylactic indomethacin for all at-risk infants, cranial ultrasound screening for baseline intraventricular hemorrhage and indomethacin to at-risk infants without severe hemorrhage, or indomethacin administration only to infants with symptomatic patent ductus arteriosus. The decision tree incorporates estimates derived from the clinical literature regarding the baseline risks of intraventricular hemorrhage, patent ductus arteriosus, and the theoretical risks associated with indomethacin therapy. The results of the decision analysis help inform the clinician regarding the decision to use prophylactic indomethacin. Obviously, if there are no ischemic complications,
the decision analysis will support the results of the meta-analysis and prophylactic indomethacin will be favored. However, concerns about possible side effects or the risk of side effects in the face of limited developmental improvement will lead different families and clinicians to make different decisions.

When we ignore EBM: high-frequency ventilation

In an attempt to prevent chronic lung disease, techniques of ventilation that theoretically decrease lung injury have been introduced. HFV is perhaps one of the most promising of these newer technologies. High-frequency ventilators apply continuous distending pressure and deliver small tidal volumes (less than the anatomic dead space) superimposed on an extremely rapid rate. Experimental work in animal models suggests that ventilation strategies using high-frequency oscillatory ventilation may prevent lung injury in preterm infants. However, individual clinical trials demonstrate little clinical benefit and there are possibilities of adverse effects, including increased risk of intraventricular hemorrhage and poor neurologic development.

Cools and colleagues reviewed the RCTs comparing elective use of high-frequency oscillatory ventilation (HFOV) to conventional ventilation (CV) in preterm infants mechanically ventilated for pulmonary dysfunction. Studies that enrolled preterm or low birth weight infants with pulmonary dysfunction mainly due to RDS who were thought to require intermittent positive pressure ventilation (IPPV) were considered eligible for inclusion in the review. Studies were only included if randomization to either elective HFOV or CV occurred early in the course of RDS soon after mechanical ventilation was begun.

In their search of the literature, the investigators found 17 RCTs (enrolling 3652 infants) that met the criteria. The size of the studies varied considerably, ranging from 43 infants to 273 infants. Although all studies included preterm infants, the upper limit for birth weight and gestational age differed between the studies. The age and randomization varied from less than 1 hour to 9 hours of age. In addition, a heterogeneous group of ventilators was used to deliver HFOV.

In an overall analysis, surprisingly few changes in clinical outcome were noted. No individual trial reported a decrease in pulmonary air leak. In fact, the meta-analysis of 12 trials suggests a slight increase in the risk of pulmonary air leak (12 studies, 2766 infants; typical RR 1.19 95% CI 1.05, 1.34). None of the individual trials demonstrated any difference in mortality, and the overall analysis demonstrated no difference in the risk of death at 28 to 30 days (9 studies; typical RR 1.09, 95% CI 0.88, 1.35) or at approximately term equivalent age (15 studies; typical RR 0.98, 95% CI 0.83, 1.14). The effect of elective HFOV on the combined outcome death or chronic lung disease at 36 to 37 weeks postmenstrual age or discharge was marginal (9 studies; typical RR 0.92, 95% CI 0.85, 1.00). No differences in mortality were noted in a variety of subgroup analyses, including evaluation of ventilation strategy (high lung volume strategy).

Some of the possible risks of HFOV were not apparent in the meta-analysis. No effect was seen in the risk of intraventricular hemorrhage or white matter damage. Few trials have addressed long-term neurodevelopmental status. Although only 2 trials reported on neurodevelopmental outcome at 1 to 3 years, an increased risk of neurodevelopmental problems was reported (typical RR 1.26, 95% CI 1.01, 1.58).

Based on the results of the systematic overview, Cools and colleagues concluded that there was no clear evidence that elective HFOV, as compared with CV, offered important advantages when used as an initial ventilation strategy.
to preterm babies with acute pulmonary dysfunction. Yet HFV has come into widespread use. As noted earlier, 22% of all very low birth weight infants in the Vermont Oxford Network are placed on HFV at some point in their hospital stay.\(^5\) It is difficult to understand the diffusion of expensive technology in the face of so little positive evidence.

**Can EBM tell us when to begin new therapies: cooling therapy for hypoxic ischemic encephalopathy**

Mild hypothermia for infants with moderate to severe hypoxic ischemic encephalopathy has shown great promise in individual RCTs. Jacobs and colleagues\(^51\) have conducted a meta-analysis of these trials that demonstrates a strong effect on mortality and severe neurodevelopmental problems. Eight RCTs were included in this review, comprising 638 term infants with moderate to severe encephalopathy and evidence of intrapartum asphyxia. Therapeutic hypothermia resulted in a statistically significant and clinically important reduction in mortality (typical RR 0.74, 95% CI 0.58, 0.94), neurodevelopmental disability (typical RR 0.68, 95% CI 0.51, 0.92), and the combined outcome of mortality or major neurodevelopmental disability to 18 months of age (typical RR 0.76, 95% CI 0.65, 0.89). The point estimate of these important clinical outcomes suggest that for every 7 infants with moderate to severe hypoxic ischemic encephalopathy treated with cooling therapy, one additional infant will survive without neurodevelopmental disability.

Debate remains, however, regarding whether there is enough evidence for widespread introduction of this therapy. Concerns include the relatively small sample size of the completed studies and the uncertainty regarding longer term follow-up. It will be interesting to track the dissemination and “effectiveness” of therapeutic hypothermia as the neonatology community attempts to integrate this new therapy into practice.

**SUMMARY**

Neonatal-Perinatal Medicine has developed sufficient resources to help support the practice of EBM. Commitment to practicing evidence-based medicine, including mastering the 5 steps recommended by Sackett and colleagues, is essential to developing expertise in EBM. An institutional commitment and collaborative learning between institutions is critical to successfully practice EBM in the complex world of Neonatal-Perinatal Medicine.

**ACKNOWLEDGMENTS**

I would like to thank Susan Hayward for her assistance in preparing the manuscript.

**REFERENCES**