SECTION 3: CASE STUDIES

The Neonatal “Sepsis Work-up”: Personal Reflections on the Development of an Evidence-Based Approach Toward Newborn Infections in a Managed Care Organization

Gabriel J. Escobar, MD

ABSTRACT. “Rule out sepsis” may be the most common discharge diagnosis among infants admitted to the neonatal intensive care unit. Although the frequency of sepsis, meningitis, and other confirmed bacterial infections has remained constant (between 1 and 5/1000 live births) for many years, the number of infants evaluated and treated is much higher. Each year in the United States, as many as 600,000 infants experience at least one evaluation for suspected bacterial infection during the birth hospitalization. The number treated is estimated at 130,000 to 400,000 per year. Despite massive overtreatment, delayed diagnosis still occurs.

The Kaiser Permanente Medical Care Program (KPMCP) considers developing and implementing an evidence-based approach to “rule out sepsis,” a research and operational priority. To achieve these goals, it is essential to consider two key aspects of the problem. First, it is important to adopt a phenomenologic approach that takes clinicians’ personal experience into account. This must include reflection on those aspects of experience often considered “irrational” or “subjective.” Second, incorporation of a phenomenologic approach needs to be tempered with sound epidemiologic methods.

If one considers these two aspects—physician experience and sound epidemiology—it is clear that much of the existing literature on “rule out sepsis” is of limited utility. Consequently, the KPMCP has conducted its own studies. These are aimed at characterizing the “sepsis work-up,” developing electronic datasets that would permit clinicians to simulate various strategies, and developing techniques for ongoing electronic monitoring.

This article summarizes the approach taken by the KPMCP Division of Research. It describes the results of a pilot study as well as the development and use of a dedicated neonatology outcomes database, the Kaiser Permanente Neonatal Minimum Data Set (NMDS). The NMDS database includes the Score for Neonatal Acute Physiology and permits ongoing monitoring of sepsis “work-ups” as well as confirmed cases of neonatal infection. The article also describes how the experience from the pilot as well as the NMDS was incorporated in the design of a much larger study on “rule out sepsis.” Finally, the article describes some important theoretic issues affecting decision rule development and the use of computer simulations in neonatology. These issues are 1) how one handles possible overanalyses of a dataset; 2) how one handles data points that are unstable (eg, the absolute neutrophil count, which can vary considerably depending on age and sampling conditions); and 3) the limitations of decision rules based on computer simulations. Pediatrics 1999;103:360–373; neonatal intensive care, sepsis evaluations, sepsis, meningitis, neonatal bacterial infections, antibiotic therapy, evidence-based medicine.

ABBREVIATIONS. KPMCP, Kaiser Permanente Medical Care Program; NICU, neonatal intensive care unit; SCN, special care nursery; ICD-CM, International Classification of Diseases, Clinical Modification; NMDS, Neonatal Minimum Data Set; SNAP, Score for Neonatal Acute Physiology; ANC, absolute neutrophil count; CBC, complete blood count; CART, classification and regression trees; LOS, length of stay.

Since 1991, at the Kaiser Permanente Medical Care Program (KPMCP)’s Division of Research in Oakland, CA, I have developed a research program in neonatology. This program now includes studies on severity of illness scoring,1–5 neonatal jaundice,6 severe neonatal dehydration, the effect of maternal substance abuse on rates of neonatal assisted ventilation, and the informatics of neonatal outcomes measurement.7

One major area of effort has been, and continues to be, the development of an evidence-based approach toward the neonatal “sepsis work-up.” Evaluating a newborn suspected of bacterial infection is no longer considered very interesting by science reporters, who invariably find other subjects far more fascinating. Among these are new therapies such as surfactant, nitric oxide, high-frequency ventilation, and partial liquid ventilation. The approach I am taking to the routine (“bread and butter”) “sepsis work-up” epitomizes my philosophic orientation toward research. This phenomenologic orientation values a clinician’s personal experience, not just methodologic rigor. This approach includes careful use and assessment of focused data collection efforts as well as reflection on the policy implications of research findings.

From the Kaiser Permanente Medical Care Program Division of Research, Perinatal Research Unit, Oakland, California.

This article is dedicated to the members of the Division of Research Perinatal Research Unit: Mary Anne Armstrong, Marla Gardner, Bruce Folck, Joan Verdi, Veronica Gonzalez, Diane Carpenter, and Blong Xiong. Their dedication to perinatal research has made this work possible.

Received for publication Sep 8, 1998; accepted Sep 8, 1998.

Address correspondence to Dr Escobar, Kaiser Permanente Medical Care Program, Division of Research, Perinatal Research Unit, 3505 Broadway, Rm 718, Oakland, CA 94611.

PEDIATRICS (ISSN 0031 4005). Copyright © 1999 by the American Academy of Pediatrics.
PHILOSOPHIC ORIENTATION

To the things themselves!

—EEdmund Husserl, Cartesian Meditations

One of the primary aims of a specifically phenomenologic method in philosophy is to make more thematic what is otherwise merely implicit and taken for granted in human experience. Moreover, phenomenology places special stress on firsthand or direct description, thereby minimizing recourse to the highly mediated constructions of metaphysics, natural science, and other theory-saturated disciplines. What is sought in the implementation of such a method is an accurate description of a given phenomenon as it presents itself in one’s own experience, not an explanation of its genesis through antecedent causal factors. The phenomenologist’s basic attitude is: no matter how something came to be in the first place, what is of crucial concern is the detailed description of the phenomenon as it now appears.

—Edward S. Casey, Imagining, A Phenomenologic Study

Edmund Husserl (1859–1938), who founded the school known as phenomenology, claimed to have defined a radically new approach toward philosophic inquiry. Setting aside the issue as to whether Husserl succeeded, neonatologists can learn from his methods and those of his followers. Central to the vision of phenomenology is the notion that accurate description should precede active theorizing. The starting point of phenomenologic inquiry is called époché, withholding judgment. Époché involves trying to step back and seeing things as they strike us, momentarily setting aside value judgments. For example, if one closes one’s eyes and has a mental image of a house, one characterizes this image without necessarily addressing the issue as to whether it is “real.” Assessing whether the house one has visualized is “real” is a second step. Theory formation follows accurate characterization. Moreover, as Edward S. Casey emphasizes, a phenomenologist studying subjects such as memory or imagination will focus on the common manifestations of these subjects rather than on the more spectacular or dramatic. Not surprisingly, phenomenologists are fascinated by Marcel Proust’s attempts at meticulous description of just how he remembered the taste of a madeleine. Phenomenologists are less interested in the elaborate mediated constructions of metaphysics, natural science, and other theory-saturated disciplines. What is sought in the implementation of such a method is an accurate description of a given phenomenon as it presents itself in one’s own experience, not an explanation of its genesis through antecedent causal factors. The phenomenologist’s basic attitude is: no matter how something came to be in the first place, what is of crucial concern is the detailed description of the phenomenon as it now appears.

REFLECTIONS ON PERSONAL EXPERIENCE

Back in the “days of the giants,” we seemed to have one basic approach to “rule out sepsis”: treat. It has been only relatively recently that the accepted wisdom of our pediatric training programs has begun to be questioned in a systematic manner. One subtle semantic manifestation of this involves three letters. Until 1994 or so, whenever the subject came up in our medical centers, the discussions used the term “septic work-up.” Implicit in this was the presumption that disease was present—a newborn had to prove his or her microbiologic innocence. Now it is more common to use the term “sepsis work-up,” which at least has a more neutral connotation!

The time has come to replace the fears of our youth with fresh reflections tempered by accumulated experience. Whenever I speak about “rule out sepsis,” I always hear one phrase from the audience: “I had a baby once...” One word summarizes our experience: fear. Most pediatricians remember having cared for what I now call the “nightmare” infant. Initially asymptomatic, this proverbial infant also had normal laboratory results (although no one seems to remember exactly what thresholds were considered “normal,” but more on this issue below). Sometime later, perhaps after “discharge to parents,” we all seem to recall that same infant, now purpuric, in shock, covered with oozing petechiae, its ventilator settings and dopamine infusion rates permanently etched into our memories.

There is another aspect of experience that we must consider: it is demoralizing to treat hundreds of infants with negative results while still managing to miss some. In the back of our minds, we suspect that a different approach might be possible. This approach would mandate more aggressive treatment and longer observation periods for some infants. At the same time, it seems likely that another group of infants would not need treatment at all. In my own case, being able to attempt this approach occurred in a group model managed care organization.

RESEARCH IN MANAGED CARE SETTINGS

More and more medical care is shifting to managed care settings, and cost-containment pressures have become intense. Outside academic institutions, justification of the research enterprise must go beyond intrinsic intellectual interest and needs to consider issues such as cost, volume, and outcomes. Researchers need an active constituency, not just interesting findings. From this standpoint, “rule out sepsis” was a good place to start. Clinicians welcomed my taking a careful look at an issue that kept them up at 3 in the morning and that sometimes had lifetime consequences for infants and their families. Administrators were interested in careful analyses of a frequent and expensive process. Given the constraints of starting a research program from scratch,
it would have been much harder to begin with a project that focused on the rarer events that dominate the academic literature (eg, infants of very low birth weight). The time to accumulate adequate numbers of such infants in a database would have been prohibitive.

The “sepsis work-up” was a good place to start. It is extremely common: somewhere between 150 000 and 600 000 infants are evaluated each year, and as many as 400 000 are treated with systemic antibiotics for a minimum of 48 to 72 hours, pending receipt of negative culture results. In the KPMCP, 6% to 7% of newborns are treated expectantly pending receipt of negative culture results, and the “bread-and-butter” evaluation and treatment combination accounts for 25% of all our NICU days.

EPIDEMIOLOGIC REFLECTIONS

Epidemiologists routinely take their objects of study and slice them up. In many respects, stratification is often a very good place to begin. This is a step that, until very recently, has not been managed properly in neonatal research. Problems have occurred because of errors in two directions. On the one hand, some studies lump all infants together, pooling the “800 grammer” with the “four kilo mec baby.” A different error is that of excessive reliance on case series. Ultimately, these problems revolve on the failure to pay attention to three questions considered essential by epidemiologists: What is the denominator? What is the numerator? Is there some sort of a control or comparison group?

Before tackling these questions from a methodologic point of view, I want to defend our profession from a phenomenologic and historical standpoint. I believe that the reason that the pediatricians and neonatologists who conducted the first systematic studies on “rule out sepsis” did not consider these issues is rooted in the history of our specialty. Our specialty began in an atmosphere of crisis and urgency. It was so much more important to act than to measure. The reality of our experience is that in the absence of data from studies conducted with systematic antibiotics for a minimum of 48 to 72 hours, pending receipt of negative culture results. In the KPMCP, 6% to 7% of newborns are treated expectantly pending receipt of negative culture results, and the “bread-and-butter” evaluation and treatment combination accounts for 25% of all our NICU days.

WHAT IS THE DENOMINATOR?

The use of denominators to convert counts into proportions seems almost too simple to mention. However, a proportion is one basic way to describe a group. One of the central concerns of epidemiology is to find and enumerate appropriate denominators to describe and to compare groups in a meaningful and useful way.

—Gary D. Friedman, Primer of Epidemiology

The major problem with the “rule out sepsis” literature is its failure to define, quantify, and use two methodologically correct denominators (all live births and all infants ever evaluated) instead of denominators of convenience (eg, infants with positive cultures or infants in a given weight range with a specific type of infection). For example, commonly cited monographs quote a series of articles that report presenting signs in infants with culture-proven sepsis and/or meningitis—eg, 14% of such infants presented with lethargy, 29% had trouble feeding, 30% had hepatosplenomegaly, etc. Absent from these publications and neonatology texts, however, are data on 1) just how common these presenting signs are; and 2) how many infants with a given clinical sign ultimately experienced adverse outcomes. Because the number of infants ever evaluated and the proportion of these infants with an individual clinical sign are seldom reported, it is not possible to determine the predictive value of a given clinical sign. Nonetheless, neonatal texts use the presence of these signs as a rationale for evaluation and treatment.

Given that the frequency of some presenting signs may be very high, it is important to note that cohort studies and randomized trials are not the only option. Case–control studies, the utility of which has been well-established in the epidemiology literature, also are an option. For such methods to be useful, however, we must first consider just what constitutes a “case,” which brings us to the second issue: the numerator.

WHAT IS THE NUMERATOR?

There is nothing intrinsically wrong with reporting on a case series of newborns with positive blood culture results, any more than there is on reporting a case series of infants with pneumonia. What is wrong is generalizing from such case series to all infants, or making blanket recommendations for evaluations based on such case series. Several sound epidemiologic principles justify this critique.

Most diseases do not manifest as dichotomous entities. It is far more common that a given clinical presentation can be best described as a point on a continuum. Because the phenomenon of spontaneous resolution of disease is also common, and needs to be discussed in the context of control and comparison groups, merely reporting on the characteristics of a group of infants with positive cultures is of extremely limited utility.

Inferences based on case series highlight how one set of events (eg, an infant developing septic shock) can lead us to ignore other events that may be just as common (eg, infants resolving transient bacteremia). Unfortunately, for several years, we did not take the epidemiologic approach (which recognizes both the forme fruste as well as the spontaneous cure). Instead, our fear of “missing” an infant led us to consider such infants from a purely personal perspective (“we got lucky”). The problem is compounded further because there are infants who “look great” at first but go on to “crash.”

The numerator problem can only be resolved by studies that report more than just which infants had a positive culture. Ideally, studies should 1) clearly define what is considered an infection; 2) include criteria for defining an infection in the absence of positive cultures; 3) report the numbers of infants who may have had spontaneous resolution of their
disease; and 4) report the numbers of infants with a severe clinical presentation that was not attributable to infection.

**IS THERE SOME SORT OF CONTROL GROUP OR COMPARISON GROUP?**

In developed nations, it is unusual for either sepsis or the suspicion of sepsis to be ignored. Clinicians intervene, perform evaluations, and provide treatment. When assessing screening and treatment strategies, epidemiologists always consider the effects of “no screening” and “no treatment.” This important consideration is virtually absent from the “rule out sepsis” literature, which primarily mentions the risk of death when sepsis is not treated. However, any attempt at developing rational strategies must include some thought about what occurs to infants with a given risk factor, clinical sign, or laboratory test result when treatment is not provided. This is critical if we are to assess the sensitivity, specificity, and predictive value of risk factors, clinical signs, or laboratory results.

The problem is complicated because, given the current state of knowledge, agreement as to who should and should not be treated does not really exist. Although some clinicians have called for a randomized clinical trial, I believe that some methodologic reflection on what would constitute a proper control or comparison group is indicated before we can even begin to design such trials.

If we use the term “control” as meaning a group of infants who do not receive a specific treatment, the definition is problematic in either prospective or retrospective study designs. In the context of prospective studies, withholding treatment would be unethical. In the context of retrospective studies, infants may have received antibiotics at a fairly late stage. In other words, if one uses the term “control,” it is necessary to also specify a time component. In the context of a hypothetic clinical trial, for example, one could have an infant with risk factor X randomly assigned to either having immediate antibiotic therapy (treatment arm) or observation for Y hours or less (ie, infants in the control arm would be allowed to cross-over to the treatment arm for ethical reasons).

Because such a study would be very difficult to design, let alone conduct, a more realistic strategy is to conduct retrospective studies that use a comparison group. The term comparison group stresses that the decision to treat or not is not being made by random assignment. The difficulty with using comparison groups is in teasing out the contribution of specific predictors. For example, one could hypothesize that an absolute neutrophil count (ANC) of <10,000 at 4 to 6 hours of age is a marker for sepsis. Testing this hypothesis is very difficult when analyzing retrospective data from a heterogeneous group of treated and untreated infants with varying frequencies of other signs, risk factors, and test results. Fortunately, epidemiologists have developed techniques for doing this.23 It is worthwhile to remember that in this situation, the epidemiologic viewpoint coincides with a clinician’s experience and a phenomenologist’s fascination with the common: all we can recall discharging infants without treatment and wondering just what would happen this time.

**FIRST ATTEMPTS AT QUANTIFICATION**

Our first attempt at developing an evidence approach to “rule out sepsis” in KPMCP was a study with limited aims.18 This project served as the pilot for the larger study described below. We did not attempt to come up with recommendations as to whether one should start treatment with antibiotics. We focused on 1) whether one could define a low-risk group using readily available predictors, and 2) whether definition of such a group would permit a reasonable inference that could guide a clinician in deciding when to stop antibiotic treatment after 24 hours in selected infants.

Figure 1 shows results of our first study. It shows information that is highly relevant to the development and implementation of an evidence-based approach to “rule out sepsis.” Data shown are from 10 KPMCP hospitals in Northern California during the months of September and October 1990. On a population basis, a number of event types occur with sufficient frequency that they could be captured with a cohort of 5709 births. For example, one can infer that in the KPMCP, with 2.7 million members, a newborn “crashes” approximately every 12 to 18 hours. During this 2-month time frame, ~13% of all live births spent at least some time in a “special care nursery” (SCN, meaning any location reserved for nonnormal newborns), and 260 infants ≥2500 g received parenteral antibiotics. Of these 260 infants, 41 were critically ill within 24 hours of entry into an SCN, and 5 were treated for presumed or proven maternal syphilis. This leaves 214 infants (3.7% of live births) who were of relatively low risk.

Figure 2 shows the outcomes of the 214 infants who were treated presumptively and who were not critically ill in the first 24 hours in “special care.” It demonstrates that these infants do not constitute a no-risk group: 3 had severe respiratory deterioration. Importantly, 2 of these infants had negative culture results.

Additional insights can be gained by examining the final discharge diagnoses among the 749 infants who entered “special care” (Table 1). There is no question that “rule out sepsis” is common; almost half of the SCN admissions had this diagnosis. Three other diagnoses that are not necessarily distinguishable from early sepsis or meningitis also are frequent; transient tachypnea of the newborn26–29 was present in 21% of all admissions, ill-defined feeding difficulties in 6.8%, and “probable” sepsis in 6.3%. During this period, only 28 of the 749 SCN admissions had culture-proven disease (a rate of ~5 per 1000 live births).

The frequencies of these diagnoses demonstrate that although true sepsis is rare, the number of infants with conditions that could presage sepsis is in fact quite large. Put differently, there is a rational basis for our fear. Therefore, if one is to conduct studies to define better approaches to this problem, it is critical to consider those diagnoses in which sepsis
**Fig 1.** Distribution of live births and SCN admissions in 10 Kaiser Permanente hospitals during September and October 1990. SCN was considered any setting reserved for nonnormal newborns. The study focus was on the 214 infants who were of normal birth weight, were not critically ill during their first 24 hours in an SCN, and who received systemic antibiotics during the birth hospitalization. These infants made up the majority (214/260, or 82%) of infants weighing ≥2500 g who received parenteral antibiotics. The other infants in this weight range who were treated with antibiotics included 41 who deteriorated very quickly and 5 who were treated because of presumed or proven maternal syphilis.

---

**DEVELOPING THE RIGHT TOOLS**

At the same time that our first study began, my colleagues and I also began to address our general information needs for neonatology, not just those related to “rule out sepsis.” Although KPMCP data systems are rich, clinical information systems available to us in 1991 to 1992 relied heavily on the International Classification of Diseases, Clinical Modification (ICD–CM). At that time, the ICD–CM system had a number of important limitations, some of which have since been corrected. For example, there was no code to identify an infant merely evaluated for sepsis, or one in whom no cause for symptoms was ever found (the V29 code is now used). There also was no code for an umbilical artery catheter, which virtually defines a newborn in an intensive care setting. Moreover, data quality with respect to the use of ICD–CM codes was heterogeneous.

We had to address three other specific problems. One was the obsolescence of mainframe-based information systems, which could not exploit the many advances and advantages of distributed networks. Second, the organizational response to the problem of “legacy systems” was complicated by the need to grant considerable local autonomy to individual medical centers. For example, some places used Macintosh computers, whereas others did not. Finally, the information sources available to us were not structured for correct aggregation (using live births as the denominator).

We made the strategic decision to build a neonatal database—called the Neonatal Minimum Data Set (NMDS)—from scratch, a process we have described elsewhere. Shortly after we began to collect data, Dr Douglas Richardson of Harvard University provided us with a then experimental protocol for a neonatal
severity of illness scale, the Score for Neonatal Acute Physiology (SNAP).\textsuperscript{1–4} The SNAP assigns points based on the degree of physiologic derangement a newborn experiences in the first 12 or 24 hours in intensive care. It is clinically intuitive—the higher the SNAP, the sicker the infant.

Use of the SNAP was important for two major reasons: one scientific, one phenomenologic. The scientific reason is fairly clear: traditional predictors used for risk adjustment in neonatology (birth weight, gestational age, sex) only explain a fraction of the variation observed between centers. The phenomenologic reason was that without some way of addressing illness severity, we could not address an experience-based statement made by physicians, “my infants are sicker.” Using SNAP permitted us to assess a number of aspects related to practice variation.\textsuperscript{5,7}

The NMDS now functions as a wide area network linking six level III units and two level II units in California and one level III unit in Colorado. It has an evolving reporting structure. Figure 3 shows a length of stay comparison grid based on mortality risk (SNAP, Perinatal Extension). Figure 4 shows an individual NICU’s basic utilization report. As presented in Fig 4, the NMDS database has the capability to track an individual unit’s “rule out sepsis”

* All “special care nursery” admissions to the 10 Kaiser Permanente birth hospitals (9/1/90–10/31/90) from a cohort of 5709 live births.

![Diagram of SNAP outcomes]

**Fig 2.** Outcomes of the 214 infants constituting the principal study group. Note that only 11 (or 5% of the group) infants had a positive culture result. There were four infants whose culture results were considered “questionable” by treating physicians. Also note that of the 3 infants who experienced deterioration after 24 hours in a special care setting, 2 had negative culture results.

**TABLE 1.** Most Frequent Neonatal Discharge Diagnoses*

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>No. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Rule out sepsis&quot;</td>
<td>365 (48.7%)</td>
</tr>
<tr>
<td>Jaundice</td>
<td>191 (25.5%)</td>
</tr>
<tr>
<td>Transient tachypnea of the newborn</td>
<td>157 (21.0%)</td>
</tr>
<tr>
<td>Hypoglycemia (all forms)</td>
<td>91 (12.1%)</td>
</tr>
<tr>
<td>In infants of diabetic mother</td>
<td>31 (4.1%)</td>
</tr>
<tr>
<td>All other forms</td>
<td>60 (8.0%)</td>
</tr>
<tr>
<td>Meconium aspiration</td>
<td>70 (9.3%)</td>
</tr>
<tr>
<td>Apnea + bradycardia</td>
<td>68 (9.0%)</td>
</tr>
<tr>
<td>Respiratory distress syndrome</td>
<td>67 (8.9%)</td>
</tr>
<tr>
<td>Ill-defined feeding difficulties</td>
<td>51 (6.8%)</td>
</tr>
<tr>
<td>Infant of diabetic mother</td>
<td>48 (6.4%)</td>
</tr>
<tr>
<td>&quot;Probable&quot; sepsis</td>
<td>47 (6.3%)</td>
</tr>
<tr>
<td>Infant whose mother used a drug of abuse</td>
<td>29 (3.9%)</td>
</tr>
<tr>
<td>Confirmed infections (cultures positive)</td>
<td>28 (3.7%)</td>
</tr>
<tr>
<td>Air leaks</td>
<td>21 (2.8%)</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>20 (2.6%)</td>
</tr>
</tbody>
</table>

* All “special care nursery” admissions to the 10 Kaiser Permanente birth hospitals (9/1/90–10/31/90) from a cohort of 5709 live births.
utilization patterns. This capability is critical if one aims to assess the implementation of an evidence-based approach.

DESIGNING THE NEXT STUDY

The next step in our work was conducting a study that used our accumulated experience. This study was funded by The Permanente Medical Group, Inc, Kaiser Foundation Health Plan, Inc, the Sidney Garfield Memorial Fund, the Packard Foundation’s Center for the Future of Children, and the Maternal and Child Health Bureau’s Research Program. It is called “Watchful Waiting” versus “Antibiotics A.S.A.P.” Its results are described elsewhere.31 In this article, I focus on the theoretic issues that defined how we structured data collection and analysis.

Our goal was to create a large electronic dataset that could be used to define decision rules (evidence-based treatment guidelines). One portion of the dataset (derivation dataset) would be used to develop the rules, whereas another (validation dataset) would be used to test them.

We made a conscious decision to take this approach because we did not feel that given the state of knowledge available in 1994–1995, it would be ethical or feasible to conduct a randomized trial. The results of a decision rule approach, however, are amenable to prospective tests.

DEFINING THE TARGET POPULATION

Because of limited resources, we were not able to conduct a study using prospective data collection. Instead, we opted for a compromise: prospective identification and retrospective chart review. We were aided by two factors. First, by 1995, the KPMCP had developed robust database systems that permitted us to establish ongoing downloading of key laboratory data (eg, complete blood counts (CBCs), arterial blood gases, and culture results). Second, because of the NMDS, we had a cadre of trained research assistants. These two factors permitted us to “piggyback” the study onto an existing infrastructure that includes a help desk for chart reviewers with questions about our research protocols.

We decided to use the proper denominator (all infants ever evaluated), which posed a problem: How does one define “evaluation” for sepsis? After discussions with many neonatologists, we defined the inclusion criteria as follows: an infant was considered to have been evaluated for sepsis if a physician suspected the condition and obtained a CBC and/or a blood culture. This definition was independent of either treatment or outcome. Strictly speaking, this definition is incorrect because a physician who evaluates a newborn may consider the risk to be so low as not to warrant any sort of screening. Using the “best” definition, however, would have been fatal for the study because of not being able to identify eligible subjects electronically or by chart review.

Four additional decisions merit special mention. First, we decided to track infants after neonatal discharge. Lack of any follow-up data has been a notable weakness of the “rule out sepsis” literature. Second, we decided to incorporate outcomes other than just positive culture results (eg, a category defined as

---

**Fig 3.** LOS comparisons using analysis of variance. Analyses are based on groups stratified by mortality risk using the SNAP–PE.1–3. Individual nurseries can determine whether their average LOS is significantly different from another unit within a defined birth weight and mortality risk category.

---

<table>
<thead>
<tr>
<th>LOCATION</th>
<th>N*</th>
<th>ALOS‡</th>
<th>SD**</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>68</td>
<td>10.5</td>
<td>5.66</td>
</tr>
<tr>
<td>X</td>
<td>73</td>
<td>10.9</td>
<td>10.12</td>
</tr>
<tr>
<td>Y</td>
<td>45</td>
<td>5.9</td>
<td>4.39</td>
</tr>
<tr>
<td>Z (REF)</td>
<td>39</td>
<td>5.7</td>
<td>4.96</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LOCATION</th>
<th>W vs.</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>W vs.</td>
<td>*</td>
<td>0.7789</td>
<td>0.0009</td>
<td>0.0010</td>
</tr>
<tr>
<td>X vs.</td>
<td>0.7789</td>
<td>*</td>
<td>0.0003</td>
<td>0.0004</td>
</tr>
<tr>
<td>Y vs.</td>
<td>0.0009</td>
<td>0.0003</td>
<td>*</td>
<td>0.9320</td>
</tr>
<tr>
<td>Z vs.</td>
<td>0.0010</td>
<td>0.0004</td>
<td>0.9320</td>
<td>*</td>
</tr>
</tbody>
</table>

---

N* = Total number of babies in the time period
ALOS‡ = Average Length of Stay
SD** = Standard Deviation

---

Time Period: 12.1.93 - 11.30.94, SNAP-Perinatal Extension 0 - 4, Birth Weight Range: 1500 - 2499 Grams

<table>
<thead>
<tr>
<th>LOCATION</th>
<th>N*</th>
<th>ALOS‡</th>
<th>SD**</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>91</td>
<td>3.9</td>
<td>1.9</td>
</tr>
<tr>
<td>X</td>
<td>57</td>
<td>3.6</td>
<td>2.6</td>
</tr>
<tr>
<td>Y</td>
<td>253</td>
<td>2.8</td>
<td>1.8</td>
</tr>
<tr>
<td>Z (REF)</td>
<td>258</td>
<td>3.4</td>
<td>1.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LOCATION</th>
<th>W vs.</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>W vs.</td>
<td>*</td>
<td>0.3430</td>
<td>0.0001</td>
<td>0.0240</td>
</tr>
<tr>
<td>X vs.</td>
<td>0.3430</td>
<td>*</td>
<td>0.0046</td>
<td>0.4302</td>
</tr>
<tr>
<td>Y vs.</td>
<td>0.0001</td>
<td>0.0046</td>
<td>*</td>
<td>0.0007</td>
</tr>
<tr>
<td>Z vs.</td>
<td>0.0240</td>
<td>0.4302</td>
<td>0.0007</td>
<td>*</td>
</tr>
</tbody>
</table>
“probable” infection for infants with obvious septic shock, but with negative culture results. Third, to avoid circular reasoning, we made an explicit decision not to use results of CBC or arterial blood gas studies to define patient outcomes. Finally, we made a heavy investment in collecting data based on two points in time: time of birth and time of onset. The use of two points in time merits additional
discussion. Time of birth is important with respect to “rule out sepsis” for several reasons. These include the fact that some commonly used predictors (such as the ANC) vary as a function of a newborn’s chronologic age. On the other hand, it also is useful to conceptualize time intervals based on when an infant was first labeled as being at risk for sepsis (ie, that moment when the infant was no longer unequivocally considered a “well baby”). One of the results of our first study was that we found that this time, which we have referred to as “entry time” or “onset time,” is charted carefully by nurses or physicians. In those situations where it is not charted, it can be inferred by examining the collection date and time of either the first CBC or blood culture.

Figure 5 shows a portion of the data collection form we used to abstract neonatal charts. The use of “highest” and “lowest” values for vital signs shows the influence of the SNAP and other severity scales. It also highlights our experience with our neonatal database, which uses dichotomous outcomes whenever possible.7

RECONSIDERING SOME KEY PREDICTORS

No discussion of the “sepsis work-up” would be complete without consideration of the interpretation of a test whose performance has acquired a ritual quality: the CBC. Virtually all published guidelines suggest obtaining this test. Unfortunately, the agreement ends there. There are myriad recommendations as to what is considered predictive: the ANC, immature to total ratio (I:T ratio), not to mention a host of others including the band count, total white blood cell count, platelet count, and various other combinations.32–39

Although some studies have taken the trouble to actually compute sensitivity, specificity, and positive predictive value, most of the studies that have attempted to justify their recommendations on an empirical basis pay little attention to two important questions: What constitutes a normal result? What is the effect of sampling variation on the assumption that a given result is a “hard” finding?

The most commonly cited study on the neonatal CBC is that of Manroe et al,40 whose graceful curves and scatterplots can be found in neonatology textbooks, review articles, and treatment protocols. No one takes the trouble to remind the reader that in the period from 0 to 24 hours of age—the critical decision time for most neonatal “sepsis work-ups,” especially in the era of early newborn discharge—Manroe et al based their graphs on a mere 108 infants. Nor are most readers aware that the 90th and 10th percentile envelopes were defined using visual inspection. One can forgive Manroe et al for their methodology, which antedated the personal computer; however, it is less comforting to think that so many of us have accepted these norms without question.

Setting aside the issue of blood sampling from a statistical perspective, it is important to remember that the actual physical sampling can lead to dramatic changes in CBC results. First of all, it has been well-documented that the CBC does vary depending on the infant’s age,40–42 on whether the sample is arterial or venous,43 and on whether the infant is crying vigorously.43 This means that a given test result—such as the ANC—can be visualized mentally not as a static value but as an extremely time-dependent one.

Laboratory test results also are experienced by physicians in at least three additional ways. The first way, which unfortunately cannot be drawn in a diagram, is through the twin filters of fear and fatigue: a total white
blood cell count of 14,500 is experienced differently at 11 am than at 4 am. Similarly, we will experience a given test result in one way when an infant is “crumpling” in front of us than when we get a call from labor and delivery informing us that the same result was obtained from an infant who is currently breastfeeding quite well. Finally, we tend to “overremember” CBC results that somehow “burned” us—these CBCs often are mentioned to me when I give talks, usually as part of the phrase, “I had a baby once . . . ” Unfortunately, very few people take the trouble to document these unique, “nightmare” infants, and my systematic search for case reports is yielding very few studies such as the one by Christensen and colleagues. The philosopher Husserl describes these aspects of our experience using terms such as “series of now-points” or “series of nows which possibly will be filled with other Objects.”

Ideally, norms for any test should be established using populations meeting rigorous criteria for being considered “normal.” In my opinion, the two best studies on the CBC are those of Schelonka and co-workers, who used a tightly defined population studies on the CBC are those of Schelonka and colleagues.4 The philosopher Husserl describes these aspects of our experience using terms such as “series of now-points” or “series of nows which possibly will be filled with other Objects.”

Existing studies also do not report the effect of maternal treatment with antibiotics on the neonatal CBC. Nor do they compare the sensitivity and specificity of the CBC against the other problematic predictor: asymptomatic status.

Some published reports do support the notion that asymptomatic status correlates strongly with a favorable outcome.4,44 However, much less attention has been devoted to the details that haunt us early in the morning when the telephone rings: 1) Just what is “asymptomatic”?: 2) Assuming that we agree on a definition, who applies it? Should it be a nurse, nurse practitioner, pediatrician, or neonatologist?: And—of great importance to a researcher—3) who reports it, who records it, and where is this recorded? Lost in the debates as to whether such infants are at X% or Y% risk is the fact that we do not in fact have any data as to what is “normal” in a newborn. For example, textbooks typically label a respiratory rate of ≥60 breaths per minute as a sign of illness. However, when I went back to the original studies stating that a respiratory rate of 60 was in fact dangerous, I discovered that these studies had extremely small sample sizes (which in fact can only be approximated because they are not clearly reported), were based on premature infants in the late 1950s, and did not actually demonstrate any associations between (supposed) predictors and outcomes.47–52 Although no comparable study has been performed on newborns, a study on 1007 infants showed no relationship between respiratory rate and illness severity or outcome.53 Significantly, in our collaborative effort to define SNAP-II with our colleagues at Harvard and Vancouver, the respiratory rate variable was not predictive of neonatal mortality.54,55

Similar problems exist with another controversial aspect of the neonatal “sepsis work-up”: when to perform a lumbar puncture. Here the literature shows an interesting relationship: studies that carefully stratify patients and clearly delineate predictor-outcome relationships tend to be more cautious in their recommendations. In contrast, studies that do not permit actual quantification of risk they tend to recommend more lumbar punctures.

Ignoring clinicians’ experience probably is also responsible for the poor compliance with various recommendations regarding screening for and management of group B streptococcus carriage. Although this literature is now quite voluminous and current recommendations have some basis on data from maternal randomized trials,61,62 they neglect one key component of clinicians’ experience. This is that physicians who decide to “rule out sepsis” are not thinking about one organism (Streptococcus agalactiae) but about several overlapping conditions and organisms meriting antibiotic treatment (eg, Escherichia coli and Listeria monocytogenes).

**DECISION RULES AND COMPUTER SIMULATIONS. WHAT THE CLINICIAN WANTS VERSUS WHAT THE METHODOLOGIST WILL ATTEST TO**

My colleagues and I have arrived at several conclusions as to what it would take to define and implement an evidence-based approach. This is where the desires of many clinicians inevitably conflict with the requirements of the methodologists. When I first began discussing our second study, one of my friends pointed to the 3-inch “code card” clipped onto his scrubs and said, “You mean you’re gonna collapse rule out sepsis into one of these? That would be great!” His comment synthesized what many physicians do in fact expect from my research unit: one or two simple algorithms with less than three branches—

If baby meets criterion A and mother meets criterion B, then treat.

5

Otherwise, if C holds, watch for D hours and then reassess.

5

Otherwise, send the baby home.

The concept of a decision rule is deceptively simple. Consider a hypothetic cohort of five newborns,
because some data points, for example, the ANC, are essential in any computer simulation strategy favorably.

370 SUPPLEMENT

However, examination of classification and regression trees (CART). CART software permits rapid generation of “outcomes trees.” Each branch of the tree is based on a predictor, which can be a dichotomous, categoric, or continuous variable.

One advantage of CART software is that it permits one to get a different view of the utility of a given predictor. For example, in our first study, the expert panel pointed to the ANC as an important predictor, and the final decision rule used a cutoff of 10 000. However, examination of classification and regression trees using these data showed something else: high ANCs (>15 000) in the first 12 to 24 hours of life were associated with favorable outcomes.

One also must conduct sensitivity analyses. These are essential in any computer simulation strategy because some data points, for example, the ANC, are “unstable.” Consider a hypothetic decision rule for “rule out sepsis” that incorporates the ANC. Because it is known that the ANC may vary with crying or where the blood was obtained, it is important to test models where one 1) systematically elevates or reduces ANC values, 2) assumes that the data point is missing, and 3) randomly inserts wrong values. Last but not least, one needs to incorporate “real world” constraints (for example, decision rules should not use data that are not ordinarily available or assume unlikely contingencies such as an infant being discharged immediately after the antibiotics are discontinued). Performing these additional simulations invariably results in projections that are far less optimistic!

Ultimately, however, this approach—showing where decision rules fail, not just where they succeed—is more fruitful. It brings out the real value of decision rules and computer simulations. Rules and simulations by themselves will not produce perfect solutions. Their true value is that they can be used as tools to make clinicians test their assumptions explicitly. In this context, one needs to remember that a decision rule that does not miss any infants is not necessarily perfect, whereas a decision rule that misses one infant in fact may be clinically useful.

FROM DATA TO GUIDELINE

The KPMCP’s approach to clinical guidelines also has evolved. One thing our organization has learned is that the unaided diffusion of new knowledge is very slow and very uneven. In an integrated system, it is important to have mechanisms not only for dissemination, but for standardization. Guidelines that are pushed by an individual champion without organizational “buy in” are unlikely to be implemented.

Accordingly, there is now a formal mechanism for the development, dissemination, and implementation of guidelines. This mechanism includes a formal approval process at the beginning, middle, and end of a guideline’s development process. At the beginning, a department chiefs’ group (in this case, the 12 nursery directors for Northern California) formally proposes a guideline effort to our Department of Quality and Utilization. Once approved, this results in a small grant that pays for physician time to work on the guideline, an important consideration because all our physicians are salaried and cannot simply leave their offices without affecting many others’ schedules. Funding also is provided for nonphysicians and for staff support (eg, for photocopying). During the development phase of the guideline, drafts are circulated widely so as to ensure that as many groups involved are aware that change is in the air and can comment on the drafts. Finally, once complete, it must be approved by a special group consisting of all the directors of our medical centers that reviews all proposed guidelines on a quarterly basis. Final approval means that it is possible to place the new guideline on the KPMCP intranet. In some cases, funding for implementation can be obtained as a separate grant.

Shortly after we began to examine data stratified
but who play an increasing role in care delivery: at four in the morning. Finally, the consensus must who will no longer merely transmit CBC results to us may be different from our own. It also will have to cians from other medical centers whose protocols consensus includes a willingness to work with clini- care networks, a political consensus must exist. This “bad data”). computer simulations with random insertions of results of intermediate steps and variable scenarios (eg, properly performed studies. The teams who will pre- form strategies. However, it is insufficient merely to have neration consumers. They will not endorse new guide- lines based on manifestly erroneous sampling strat- egies. However, it is insufficient merely to have properly performed studies. The teams who will prepare such guidelines must be provided with the re- sults of intermediate steps and variable scenarios (eg, computer simulations with random insertions of “bad data”).

Second, and particularly in integrated managed care networks, a political consensus must exist. This consensus includes a willingness to work with clini- cians from other medical centers whose protocols may be different from our own. It also will have to incorporate changing how we interact with nurses, who will no longer merely transmit CBC results to us at four in the morning. Finally, the consensus must include those individuals who are not clinicians at all but who play an increasing role in care delivery: administrators, computer networking specialists, and even financial analysts.

Most importantly, implementing evidence-based guidelines cannot ignore the role of what we have conveniently relegated to the category of items we consider “irrational.” No guideline effort on “rule out sepsis” can succeed if it does not consider that feeling of fear in the pits of our stomachs, the tears in parents’ eyes when we inform them that we are “placing the baby on antibiotics,” that vague sense of unease we feel when we sign our names after the phrase, “Discharge to mother.” What are some possible consequences of such shifts in practice patterns? Because the physical ex- amination of a newborn is critical to proper triage, it is reasonable to expect that nurseries that have an increased capability to offer a “second opinion” on an infant’s assessment will be on average more effi- cient and have better outcomes. Similarly, nurseries with better information distribution systems also are likely to be safer as well as more efficient. Finally, as we treat fewer term infants, this will have a signifi- cant impact on the census of many units. How this potentially de-stabilizing impact is managed by our profession remains to be seen.

One result that is not so obvious is that the re- search community will learn from this process itself. We may discover that in the same way that using the SNAP permits delineation of practice variation, well- designed studies can help us “tease out” patterns in how clinicians change their practice. This includes confronting apparently simple issues: Should umbil- ical artery catheters be placed “high” or “low”? At what level of illness severity does a sick newborn merit two umbilical catheters rather than one? How does one compare the safety and effectiveness of feeding regimes in nurseries with identical rates of necrotizing en- terocolitis? Under what conditions could one justify “prophylactic” assisted ventilation in term infants suspected of having pulmonary hy- pertension? It also must include thornier issues such as human error and why physicians may reject evi- dence-based guidelines. Perhaps we may discover the rational basis for some apparently irrational be- haviors.

And then we may return to where we should always begin: to reflecting on our own experience, not with the ossified hastiness of our youth, but with the intellectual freshness that comes from involving other disciplines.

CONCLUSION

I do not believe that we will “banish” the tradi- tional “sepsis work-up” from the nursery, but I do believe we will succeed in implementing new ap- proaches to this problem. I believe in this not just because of the compelling logic of properly per- formed analyses, but also because physicians and nurses are tired of the current approaches.

If we do succeed, it will be because of the follow- ing factors. First and foremost is the need for sound science. Neonatologists are sophisticated informa- tion consumers. They will not endorse new guide- lines based on manifestly erroneous sampling strat- egies. However, it is insufficient merely to have properly performed studies. The teams who will prepare such guidelines must be provided with the results of intermediate steps and variable scenarios (eg, computer simulations with random insertions of “bad data”).

ACKNOWLEDGMENTS

This work has been funded by the Permanente Medical Group, Inc, Kaiser Foundation Health Plan, Inc, the David and Lucile Packard’s Center for the Future of Children, the Sidney Garfield Memorial Fund, and the Maternal and Child Health Bureau’s Research Program.

I thank Dr Joseph V. Selby, Dr De-Kun Li, Dr Jeffrey B Gould, Dr Jeffrey D. Horbar, Ms. Mary Anne Armstrong, and Ms. Maria N. Gardner for reviewing the manuscript. Graphics were prepared by Ms. Gardner and Ms. Verdi.

REFERENCES


