ABSTRACT. We can learn what is achievable with current technologies by comparing our neonatal intensive care unit outcomes with others. Because neonatal intensive care units may vary with respect to their case-mix, risk adjustment is essential to making fair comparisons in any research that does not equalize risks through randomization. Risk adjustment first requires strict definition of each specific outcome. Then each risk factor is measured and weighted accordingly. Severity of illness scores are a special form of risk adjustment. The leading newborn illness severity scores rely on physiology-based items from bedside vital signs and laboratory tests. The mechanics of score development are discussed including item selection, definition, collection, and potential biases. The process of weighting risk factors usually involves building multivariate models. Issues of derivation, validation, discrimination, calibration, and reliability affect the utility of all scores. Once a comparison is appropriately risk-adjusted, there are important cautions about interpretation, including the source of the reference (benchmark) population, sample size, and biases from incomplete risk adjustment. Nonetheless, these findings can spur quality improvement efforts that can lead to dramatic, system-wide improvements in outcomes. Pediatrics 1999;103:255–265; severity of illness, pediatrics, neonatology, intensive care units, neonatal intensive care, research ethics, health services research, outcomes research, benchmarking.

ABBREVIATIONS. SNAP, Score for Neonatal Acute Physiology; CRIB, Clinical Risk Index for Babies; NICU, neonatal intensive care unit; RDS, respiratory distress syndrome; ECMO, extracorporeal membrane oxygenation; APACHE, Acute Physiology and Chronic Health Evaluation; ICU, intensive care unit; PRISM, Pediatric Risk of Mortality.

The purpose of this article is to provide a general overview on the measurement of illness severity and methods of risk adjustment used for neonatal intensive care. This is not intended as a primer for developing an illness severity score. Definitive treatments of risk adjustment methodologies are available, to which we acknowledge our intellectual debt. Other very good discussions of methodology are included with the descriptions of adult and pediatric intensive care scoring systems and a review of adult scores as well as the original descriptions of the Score for Neonatal Acute Physiology (SNAP) and the Clinical Risk Index for Babies (CRIB). Our two previous reviews of neonatal illness severity focused first on the variety of neonatal outcomes scores and then on the range of applications of SNAP and CRIB in their first 5 years of use. Our focus here is on the mechanics of risk adjustment, the shortcomings inherent in the design of any risk adjustment strategy, and the application of these to inter-neonatal intensive care unit (NICU) comparisons for benchmarking, quality improvement, and health services research.

PURPOSES OF RISK ADJUSTMENT

What Is Risk?

The word “risk” implies the possibility of a bad event such as untimely death or financial loss. Mathematically, risk represents a probability between 0 (never) and 1 (certainty). Intrinsic to this concept is a sense of vulnerability as displayed in the phrases “high-risk pregnancy” or “high-risk newborn”. Although such uses are common, they gloss over the critical question “Risk of what?” Of mortality? Complications? Developmental delay? High-cost? Prolonged length of stay? This clarification is not simply linguistic; operationalizing risk requires specifically defining the adverse outcomes in concrete, measurable terms.

What Is Risk Adjustment?

Risk adjustment is the process of sorting patients in each comparison group into different strata of risk, and then making comparisons separately for each stratum. The aim of adjustment is to permit fair comparisons between groups. For example, if NICUs have different mortality rates, it is unfair to declare that the one with the higher mortality has poor quality of care. One must first identify the intrinsic mortality risk of the infants admitted to each unit and then compare outcomes within narrow risk strata (for example, birth weight-specific mortality). Risk adjustment is a method of comparing only between
cases of similar risk, and then aggregating the comparison across all classes of risk.

**Purposes of Risk Adjustment**

Comparisons of outcomes are inevitable. Risk adjustment permits such comparisons to be more informative. In clinical trials, randomization distributes risk evenly between the two groups, so that risk adjustment is either unnecessary, or done posthoc in cases where randomization was not fully successful. Risk-adjusted comparisons are required when analyzing outcomes that are difficult or impossible to randomize such as quality of hospitals, clinical practice styles, regional organization of care, trends over several years, utilization of resources, medical prognosis, and financial payments. For example, a reimbursement system that paid a fixed cost per case would penalize those who cared for a higher risk population. The term “case-mix” has emerged in the medical cost literature to reflect the fact that, within a population, individual patients may have a range of risks, and that the aggregate outcome reflects the aggregate risks. Case-mix is a useful concept when comparing the performance of hospitals and clinicians.

**Measuring Risk**

There are usually many risk factors for a single outcome. Using NICU mortality as an example, some risk factors have a known or presumed biological relationship to mortality (respiratory distress syndrome, congenital anomalies, acidosis) others may represent vulnerability (gestational age, birth weight, male gender, fetal growth restriction) and others are merely proxies for other risks (race, socioeconomic status). Illness severity is itself a complex construct discussed below. Often a risk factor may include several characteristics that influence outcome simultaneously. For example, multiple birth may represent a higher likelihood of immaturity, of growth restriction, or of respiratory distress syndrome (RDS) in the second twin, but it may also represent older white mothers who could afford fertility treatment. Each factor must be carefully defined and measured. Failure to consider all the implications of a given risk factor will lead to biases. For example, socioeconomic status derived from insurance status may have very different implications than socioeconomic status inferred from education or from census tract of residence. It is important to recognize that there is a cascade of risk from obstetrics into neonatal care. Thus, an obstetric population with high rates of medical complications, such as teen pregnancy, drug abuse, or infectious morbidity, will result in a higher risk population of newborns including prematurity, infections, small for gestational age, or asphyxia. Thus, quantifying obstetric risk is an important component of neonatal risk. However, as stated above, for each outcome, different combinations of risk factors may play different roles.

**Defining Outcomes**

Equal attention must be paid to precisely defining outcomes. Although NICU mortality would seem clear enough (see below), measuring other outcomes is a challenge in itself. For example, there is extensive debate about the definition of chronic lung disease—censoring of early deaths, ascertainment (differential availability of 28 day, 36 and 42 week data), and reliability (chest radiograph readings). How chronic lung disease is defined inevitably affects the number and strength of risk factors, and thereby the demands for risk adjustment in any comparison. Similar challenges affect definitions of necrotizing enterocolitis, nosocomial infection, intraventricular hemorrhage, and periventricular leukomalacia. Measuring costs of care are equally challenging because of the variability in resource use and distortions in costing procedures. Weak measures of outcome will invariably vitiate the value of any comparison, regardless of how well risks are adjusted.

**NICU MORTALITY EXAMPLE**

**Hypothetical Example: NICU A Versus NICU B**

As a hypothetical example, we present two NICUs in Table 1. The crude mortality rate for NICU A is nearly three times that of NICU B (Table 1A). A cursory review might conclude that NICU A had poor quality of care, but a knowledgeable clinician would be skeptical if NICU A were an all-transport children’s hospital and B were a maternity hospital NICU. Thus, the first step is to adjust for marked differences in the birth weight distribution (Table 1B, “Admissions”). This stratified analysis substantially reduces the higher mortality rates at NICU A. The relative risks range from 1.7 to 2.0 with the exception of the normal birth weight stratum. Again the clinician would be suspicious of a referral bias among the normal birth weight infants, anticipating transport to NICU A of all the sickest patients in the region for inhaled nitric oxide, extracorporeal membrane oxygenation (ECMO) or surgery. The birth weight and severity stratified comparison is presented in Table 1C and further reconciles but does not completely eliminate the disparities. It also documents that NICU A treats a greater proportion of sicker infants, especially for normal birth weight infants.

**Interpreting Differences—Searching for Biases**

Persistent disparities after risk adjustment must be considered “unexplained” in which one potential difference may be quality of care, but the burden is on the researcher to seek alternative explanations. For example, it is possible that measured illness severity is systematically biased in the outborn population. This phenomenon, called lead time bias, may occur if the process of transport has partially stabilized the infants before the baseline measurement of severity. The measured derangements are less extreme, and the scores are lower, yet the infants are still severely ill. When their subsequent mortality rate reflects their true severity, it will appear as if these very sick infants did worse than expected. Alternatively (or in addition) it could be that the insults sustained due to suboptimal resuscitation and stabilization in community hospitals has a lasting effect,
beyond measured severity. Continuing the example, a diligent researcher might also discover that NICU B was underreporting deaths. It might be reporting only infants cared for in the NICU and not including infants transferred out for ECMO. Similarly there might be a local policy of letting infants with prenatally diagnosed lethal anomalies remain with their mothers in the delivery room (and not become a NICU admission). Other possibilities could include a different spectrum of parental preferences for withdrawal of care or unmeasured differences in population rates of drug use. Thus, risk adjustment must always be considered potentially incomplete.

**TABLE 1.** Comparison of Two Hypothetical NICUs

<table>
<thead>
<tr>
<th></th>
<th>NICU A</th>
<th></th>
<th>NICU B</th>
<th></th>
<th>P</th>
<th>Risk Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Deaths</td>
<td>Admissions</td>
<td>%</td>
<td>Deaths</td>
<td>Admissions</td>
<td>%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>151</td>
<td>1000</td>
<td>15.1</td>
<td>52</td>
<td>1000</td>
<td>5.2</td>
</tr>
</tbody>
</table>

**1B. Birth Weight Specific Mortality**

<table>
<thead>
<tr>
<th>Birth Weight</th>
<th>NICU A</th>
<th></th>
<th>NICU B</th>
<th></th>
<th>P</th>
<th>Risk Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1000 g</td>
<td>50</td>
<td>100</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000–1499 g</td>
<td>20</td>
<td>200</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1500–2499 g</td>
<td>6</td>
<td>300</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;2500 g</td>
<td>75</td>
<td>500</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>151</td>
<td>1000</td>
<td>15.1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**1C. Birth Weight and Severity Adjusted Mortality**

<table>
<thead>
<tr>
<th>Severity Group</th>
<th>NICU A</th>
<th></th>
<th>NICU B</th>
<th></th>
<th>P</th>
<th>Risk Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>BW &lt;1000 g</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>0</td>
<td>0</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>20</td>
<td>50</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severe</td>
<td>30</td>
<td>50</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000–1499 g</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>2</td>
<td>50</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>6</td>
<td>80</td>
<td>7.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severe</td>
<td>12</td>
<td>70</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1500–2499 g</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>3</td>
<td>150</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severe</td>
<td>3</td>
<td>50</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;2500 g</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>5</td>
<td>100</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>10</td>
<td>100</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severe</td>
<td>60</td>
<td>300</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>151</td>
<td>1000</td>
<td>15.1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Effects of Sample Size**

This example also illustrates the serious problem of limited sample size. One thousand admissions is about 2 years at most NICUs, yet after double stratification by birth weight and severity, there are numerous cell sizes less than 5 cases, including several zeros, thereby mathematically limiting comparisons. Given the pace of recent mortality improvements in neonatal intensive care, it is likely that the baseline mortality risks would already be shifting downwards during the 2 years of hypothetical data collection. Thus, it is possible that very real quality dif-
OTHER OUTCOMES BEYOND MORTALITY

The success of neonatal intensive care has reduced overall NICU mortality to <5% in many NICUs. Thus, outcome comparisons must focus on morbidities and complications. For infants <1500 g there are the “alphabet soup” of outcomes: bronchopulmonary dysplasia or chronic lung disease, intraventricular hemorrhage, periventricular leukomalacia, retinopathy of prematurity, and necrotizing enterocolitis as well as complications such as nosocomial bacteremia and air leak. Even these are decreasing to single-digit incidence in many NICUs. For higher birth weight infants (1500–2499 g) serious morbidities are frankly uncommon. Such infants remain in the NICU primarily for growth and maturation. As a consequence, there is an urgent need to define “intermediate” outcomes; conditions that are more common and place the infant serious risk, but are not inevitably damaging. Examples might include poor growth by 28 days, hypernatremia, prolonged stay, and hospital readmission. Although relevant comparisons are emerging for standard outcomes, virtually no benchmarks exist for such intermediate outcomes. The ultimate goal is long-term outcome, such as the incidence of cerebral palsy, school performance, and health status. Measuring each of these requires rigorous attention to definition and measurement, beyond the challenge of finding the resources for prolonged follow-up.

ILLNESS SEVERITY

What Is Illness Severity?

Illness severity is an instinctive, self-evident concept; the challenge has been how to operationalize its measurement. Illness severity is itself a risk factor for many outcomes, but it is not synonymous with risk. For example, it is possible for a 1200-g newborn to be minimally ill or severely ill, with birth weight and severity each contributing independently to mortality risk. It is also possible for severity to correlate inversely with risk, such as high cost among ECMO patients, where early deaths cost less. The reader is referred to the extensive literature on adult and pediatric measures of illness severity noted in the beginning of this article, which has influenced the approach to measuring illness severity in newborns.

Approaches to Measuring Severity

There are three general approaches to measurement of illness severity: diagnosis, therapy, and physiology:

Diagnosis-based Measures

It is easy to recognize that an infant with RDS is more ill than an infant without the diagnosis. However, this approach overlooks the broad spectrum of severity within the International Classification of Diseases (9th revision) diagnosis 769, the code for RDS. Similar criticisms can be applied to pneumonia, sepsis and perinatal asphyxia. Furthermore, there is enormous variation in coding among hospitals, often in response to reimbursement incentives. The range of individual diagnoses is vast, and may result in small, fragmented comparisons. Diagnosis–Related Groups were an attempt to combine similar risk categories, but the results were dismal for newborns. Diagnosis–Related Groups were an attempt to combine similar risk categories, but the results were dismal for newborns. The revised Diagnosis–Related Groups are useful for length of stay (for which they were devised) but have limited utility for other applications. For example, neonatal death is a separate category, precluding computations of mortality risk by category. Finally, diagnoses available on administrative records reflect discharge diagnoses, not admission diagnoses. Thus, complications occurring during care or resulting from clinician practices (e.g., pneumonia, NEC, nosocomial infection) are included in the baseline risk, thereby “excusing” them.

Therapy-based Measures

An alternative way to measure severity is to count the number and types of therapies applied to the patient. For example, patients requiring mechanical ventilation and pressors could be compared between two hospitals. The flaw in this is evident; clinicians with more interventionist practice styles would be scored as having sicker patients. Paradoxically, those patients in whom such treatments were not obligatory, but rather used at clinician discretion, would have substantially better outcomes, and the interventionist clinicians would appear to have better overall results.

Physiology-based Measures

The two major neonatal severity measures, SNAP and CRIB, have focused on measuring and scoring physiologic derangements, following the example of the Acute Physiology and Chronic Health Evaluation (APACHE) in adult intensive care units (ICUs) and the Pediatric Risk of Mortality (PRISM) in pediatric ICUs. The rationale is that, regardless of disease or diagnosis, derangements from physiologic norm increase the likelihood of adverse outcome, and that the greater the derangements, the greater risk. This rationale obtains for each organ system, and the composite severity can be represented by the weighted sum of derangements across all organ systems. For both SNAP and CRIB, these physiology scores are then combined with other risk factors (including birth weight, gestational age, low Apgar scores and the presence of serious congenital anomalies) to provide an overall risk of mortality.

MEASURING PHYSIOLOGICAL ILLNESS SEVERITY

Item Selection

The rationale for, and examples of good and bad items are presented in Table 2. First and foremost, the items must be predictive, that is, they must vary with changes in severity. For example, BP, pH, oxy-
genation and urine output all vary with changes in severity. Beyond this, items must be available, measurable, frequent, accurately recorded and reliable. Availability is essential; lung compliance and central venous pressure are closely linked to severity and outcome, but require special measurements and are not routinely collected, so they could be used only in prospective research settings. Items must be measurable. Apnea has been a particularly difficult item, where duration and frequency are erratically reported, and mechanical ventilation interferes with ascertainment. Certain items, such as tests for disseminated intravascular coagulopathy, may be highly predictive but so infrequent that their contribution is not worth the cost of data collection. Items must be accurately recorded. Urine output is both important and measurable, but may have critical gaps. Reliability means that a second abstractor will obtain identical results. Reliability is enhanced by selection of well-defined, simple items, and by specification of the single identifiable values (for example, the highest or lowest).

Timing of Data Collection

One intrinsic weakness of physiologic scores is that they require a minimum time period to collect baseline data. The longer the period, the more complete the data but also the more contaminated it becomes with the effects of successful (or unsuccessful) treatment and thus no longer reflects “admission” severity. In addition, the score should only be used to predict events occurring after the scoring period. CRIB uses a 12-hour baseline period from birth.16 The original SNAP used a 24-hour period,14 but has demonstrated that a 12-hour baseline is sufficient.45 SNAP-II uses a 12-hour period.46–48 Evaluations of more abbreviated time periods in adult and pediatric intensive care suggest that time periods <6 hours cause serious deteriorations in performance of the scores.49

Number and Diversity of Items

The larger the number of score items, the more robust the score, both for detecting smaller changes (dynamic range), and in the face of sparse data. Sampling items across each organ system also ensures validity in a wider variety of disease processes. However, because each item comes at a cost of data collection, more parsimonious scores have an advantage. The original SNAP captures 34 items and takes approximately 20 minutes of chart review. From this, one can calculate six organ-system subscores. SNAP is applicable in term infants where there is a wide range of diagnoses and organ system failures (eg, renal failure or neurologic disease in addition to respiratory illnesses). By contrast, CRIB was designed for ease of data collection. In the population ≥31 weeks, most illness is adequately captured by sampling only three items in the respiratory/metabolic systems (worst base deficit, highest, and lowest appropriate oxygen requirement). The revised SNAP-II scores six items, covering six systems (pH, temperature, BP, oxygenation, urine output, presence of multiple seizures) yet has performance equivalent to SNAP and CRIB for both very low birth weight infants and for larger infants. The drive for the most parsimonious scores will be reversed when electronic medical records become widespread. Then, the negligible costs of data collection will give way to incorporating large volumes of data using sophisticated algorithms.50 Indeed, in pediatric intensive care, two different PRISM scores have been defined; a parsimonious admission day score11 and a richer, more complex physiology score for sequential scoring.12

Weighting of Items

The severity of each physiologic derangement must be assigned a score weight, so its contribution to overall severity is proportionate. There are two methods of assigning such item weights; clinical and empiric. Empiric weights are derived from the coefficients in multivariate models (discussed below). Clinical weights are assigned by an expert panel based on clinical judgment.51 The clinical weighting method used in SNAP was a nonlinear scale of 0, 1, 3, and 5 points where 0 reflects no physiologic derangement, 1 a mild arrangement requiring close monitoring, 3 for derangements requiring immediate change in therapy, and 5 reflecting life-threatening values. The item weights for CRIB were empirically derived to create a simple additive integer scale (see below). SNAP II also uses empirically derived integer weights.46–48 Knaus et al17 noted that the empirically derived weights for APACHE III were usually fairly close to the clinician-estimated weights from APACHE II.

Biases in Item Selection and Data Collection

One of the most fundamental issues for validity of any illness severity score is the handling of missing values. For example, an infant may not have a pH recorded during the scoring period. This could occur for several reasons each of which would have different implications for validity. First is the likelihood that the infant was too healthy; the clinician presumed normality and did not order the test. Second, the clinician may have desired a pH test, but was on transport and testing was not available. Third, the test was ordered, but the results were lost (laboratory error, lost report) or critical sections of the medical record are missing. Fourth, the clinician might have anticipated the result would be poor, and intervened with bicarbonate treatment without obtaining the test. Fifth, the parents could have requested a do-not-resuscitate order, so no tests were drawn. Sixth is the possibility that the clinician should have obtained the test, but through lack of skill, failed to do so. SNAP and CRIB make the first assumption; that a competent clinician judged that the test was probably normal and not necessary. Restricting scoring to within the NICU, SNAP minimizes the risk of lack of availability of testing. The problem of lost results is usually dealt with by excluding the score as incomplete, because it is not feasible to sustain the clinician’s assumption of normality. Although this exclusion avoids computing a biased score, higher rates of lost results on one institution may seriously distort
analyses. The fourth type of “missing” values usually obtains scorable results but they are obtained later and tend to understate the worst level of derangement. This practice style bias is difficult to detect. A related bias is excessive testing, where the large numbers of results permit finding and scoring a slightly more extreme value. This practice tends to overstate the worst level of derangement. The moribund bias is usually evident on case review and should be considered inappropriately scored, because the goal of treatment is comfort only. These cases should be counted separately. Finally the incompentence bias is possible but is rarely subtle. Lead time bias is discussed above in the example.

Risks Not Measured by Neonatal Severity Scores

SNAP and CRIB are “generic” severity scores; that is, they can be used on all cases in the NICU regardless of diagnoses with the exception of moribund infants offered comfort care. The assumption of derangements of homeostasis leading to a final common pathway toward death or damage is generally reasonable. However, for certain classes of patients, these assumptions may break down. The most obvious are congenital anomalies, where standard physiology no longer obtains. An infant with a cyanotic congenital heart disease may show only a single abnormal value (low oxygenation) yet may have a life-threatening illness. CRIB includes a qualitative marker for congenital anomalies,16 and SNAP has investigated supplemental use of this information.52 Despite the “common pathway” assumption, different diagnoses may also have different implications, specifically regarding treatability of the disease. Equal levels of physiologic derangement in infants with septic shock versus RDS may have very different prognoses. Such diagnosis-related adjustments have been developed for APACHE-III53 and PRISM-III,11 but remain beyond capacity of neonatal scoring systems at present.

MODEL BUILDING
What Makes a Good Model?

Much of the literature on prediction models derives from econometrics because of the need to make economic forecasts using massive amounts of data and often very complex interactions. Harvey54 lists the criteria for judging a good model (Table 3). These concepts apply equally well to outcome prediction models in medicine. The law of parsimony is that the simplest model should be presumed the best model. According to Friedman,55 “A hypothesis (model) is comparison of its validity of a hypothesis (model) is comparison of its predictions with experience.”

Types of Regression Models

The type of regression model used depends on the nature of the outcome to be tested (Table 4). Multiple linear regression assumes that the dependent variable is randomly distributed, and that the effects of the independent variables are additive. Thus, beyond selecting the appropriate regression model, the functional form of the model must be correctly specified, eg, linear versus nonlinear regression models. In logistic regression models, the additive coefficients are logs, so the model effects (odds ratios) are multiplicative. Use of these more complex models must be justified by substantially better predictive performance. Other mathematical techniques such as recursive partitioning57 provide robust models and make fewer assumptions.

Sample Size

Regression models are data-intensive and require very large sample sizes. A general rule of thumb is that one needs 5 to 10 outcomes (ie, deaths) for each predictor variable in the derivation model.58 Thus, a data set with 40 deaths could afford only 4 to 8 predictors. Given the rapidly diminishing mortality and morbidity in NICUs, future risk adjustment
models will certainly require large-scale collaboration to derive.

Derivation and Validation Data Sets

The model derivation process described above is designed to maximize predictive power, but runs the risk of overfitting the idiosyncrasies of that particular data set. Thus, a properly derived model requires testing on a separate validation data set. The most common way is the split half method, where a random half (or better, 70%– 80%) of the cases are used to build the model, and the rest are used to validate it. A more robust test of validity is to apply the derived model to a novel data set. Required sample sizes for the validation data sets are also large to have tight confidence intervals on the estimates. Measures of calibration and discrimination (below) are only legitimately reported on the validation set.

Prediction Is Not Causation

It is important to remember that regression analysis does not necessarily imply causation. Regression analysis only deals with the dependence of one variable on other variables. In the words of Kendall and Stuart39: “A statistical relationship, however strong and however suggestive, can never establish causal connection: our ideas of causation must come from outside statistics, ultimately from some theory or other.” The selection of known risk factors based on biologic plausibility is supportive, but not definitive. In addition, the success of regression analysis depends on the availability of the appropriate data.

Validity and Reliability

In order for an illness severity score to be usable, it needs to be both valid and reliable (ie, it is consistent with theory and can be reproduced). Table 5 displays the types of validity by which one judges a model.60 Reliability is usually measured as test retest and intrarater reliability. These are enhanced by having clearly defined, objective variables in the predictive score (see above on item selection). However, even if an illness severity score satisfies these criteria, it may still only be valid for the population from which it was derived. In order for it to be used in other populations, it must be shown to be valid for them as well. Consequently, the score needs to be validated in several different populations. CRIB,16 SNAP,14 and SNAP-II46–48 have all been tested in several different populations and found to be robust predictors of mortality.

Model Comparisons

Calibration (Goodness of Fit)

Calibration is a measure of how well a predictor estimates the outcome (ie, whether it underestimates or overestimates risk). Goodness of fit tests allow us to compare the calibration of different models used to derive illness severity scores. For example, if the score predicts a 50% mortality risk, then half of the infants with this score should have died. The R² statistic is used to compare explained variance in linear regression. The likelihood reduction statistic is conceptually similar and is used to compare logistic regression models. It is computed as the ratio of χ² attributable to the covariate divided by the χ² of the logistic model without the covariate. The most commonly used overall goodness of fit test is the Hosmer-Lemeshow goodness of fit test60 although several others exist.39

 Discrimination (Area Under the ROC Curve)

Physicians usually like to know whether a diagnostic (or predictive) test result is positive or negative. However, a dichotomized test result loses some of the information value of a test. It is self-evident that a very positive test result is more certain than a borderline positive result. Setting the cut point for “positive” at a very high level will avoid most false-positives (ie, the dichotomized test will have a high specificity), but may miss some true-positive cases with lower test results (ie, the dichotomized test will have a low sensitivity). To overcome this problem, one could consider a variety of cut-points as the test becomes increasingly positive. The area under the

TABLE 4. Types of Regression Models

<table>
<thead>
<tr>
<th>Types of Regression</th>
<th>Uses</th>
<th>Example</th>
<th>Rationale/Advantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>Interval outcomes</td>
<td>Weight gain</td>
<td>• Simple, understandable</td>
</tr>
<tr>
<td>Logistic</td>
<td>Binary outcomes</td>
<td>Survival/death</td>
<td>• Coefficients represent linear (additive) effect of each predictor</td>
</tr>
<tr>
<td>Log-linear</td>
<td>Skewed distributions</td>
<td>Days of mechanical ventilation</td>
<td>• Coefficients represent odds ratios</td>
</tr>
</tbody>
</table>

TABLE 5. Types of Validity

<table>
<thead>
<tr>
<th>Types of Validity</th>
<th>Examples</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face validity</td>
<td>Low pH, low BP, low pAO, predict mortality</td>
<td>Model includes plausible predictors from experience/theory</td>
</tr>
<tr>
<td>Criterion validity</td>
<td>SNAP-II correlates with CRIB and SNAP-I</td>
<td>New score correlates with established scores that measure similar things</td>
</tr>
<tr>
<td>Construct validity</td>
<td>Mortality risk → transfusions</td>
<td>Model accrues validity by repeated demonstration of plausible predictions—ie, it behaves in ways the abstract “construct” (mortality risk) should.</td>
</tr>
<tr>
<td></td>
<td>Mortality → IVH rate</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mortality risk → costs</td>
<td></td>
</tr>
</tbody>
</table>
**Data Reliability**

Accuracy and reliability of data are crucial to any effort to make fair comparisons of performance between NICUs. Because data are required from multiple institutions engaged in any such effort, strict definitions for data elements have to be followed. Data abstractors need to be appropriately trained, provided with code books with definitions, and to have access to a single coordinating source for clarification of definitions. Data should preferably be collected prospectively and error-checking mechanisms need to be in place to identify missing or erroneous data for corrections need to be in place. Random reabstraction of a sample of data to check for reliability will also provide an opportunity to reinforce standard procedure. Failure of any of these steps may bias data collection and invalidate inter-NICU comparisons. In addition, the association between risk and outcome may change over time as the technology improves.\(^43\) Illness severity measures must therefore be recalibrated regularly both over time and location to ensure that they remain valid and reliable. This is important for maintaining predictive power, validity, and reliability.

**USE OF RISK ADJUSTMENT FOR BENCHMARKING**

**Benchmarking—Against What?**

The simplest form of comparison is a time trend within one’s own institution. However, a NICU with poor performance (e.g., high rates of nosocomial bacteremia) can show steady improvements each year yet not recognize how serious the problem is. Thus, the most informative comparisons are against several similar NICUs. Although using the network average as the gold standard is both politically neutral and mathematically sound, it misses opportunities to identify the NICUs with the best performance (e.g., lowest risk-adjusted bacteremia rates).\(^65\) These sites both illustrate achievable goals and may represent “best practices” that can be emulated. Benchmarking refers to both the best performing sites and the process of comparing outcomes. Given the complexity of NICU care, it is likely that different NICUs will be top performers in some outcomes but not others.

**Observed Versus Expected**

The output of multivariate models discussed above is an expected probability of outcome (e.g., mortality) for the individual with a given set of risk factors (birth weight, gestational age, illness severity, gender, etc). If these individual probabilities are summed (e.g., five infants with mortality risk of 0.01, 0.09, 0.15, 0.30, 0.70, which would equal 1.25), the resulting number is the expected mortality for the population, given its level of risk. This can then be compared with the actual (observed) mortality (e.g., for the same five infants: lived, lived, lived, died, died or one death observed). The ratio of observed to expected ratio (1:1.25 or 0.80) is called the standardized rate, that is, it is the risk-adjusted performance relative to the “standard” population (standard being the population from which the multivariate equation was derived).

**Effects of Sample Size**

It is intuitively clear from the above example of five cases that the better-than-standard rate is not strong proof of superior performance because of very small numbers. It is less obvious that rates based on whole NICUs may still be too limited for confident comparisons. Horbar\(^66\) used computer simulations to examine the probability of correctly identifying substandard NICUs in a hypothetical network. A NICU with a birth weight-specific mortality rate of 130% of the network average was correctly identified as substandard 50% of the time if it had 35 admissions per year, 59% of the time if it had 70 admissions, and

<table>
<thead>
<tr>
<th>Hospital A</th>
<th>Hospital B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Use of Rescue Surfactant</strong></td>
<td><strong>Overall Mortality Rate (%)</strong></td>
</tr>
<tr>
<td>Mortality (Process)</td>
<td>100%</td>
</tr>
<tr>
<td>Mortality (Outcome)</td>
<td>21%</td>
</tr>
<tr>
<td>Use of Prophylactic Surfactant</td>
<td>100%</td>
</tr>
<tr>
<td>Surfactant Rate (Process)</td>
<td>16%</td>
</tr>
<tr>
<td>Surfactant Rate (Outcome)</td>
<td>17%</td>
</tr>
<tr>
<td><strong>Reduction in Mortality in</strong></td>
<td><strong>Hospital B versus A</strong></td>
</tr>
<tr>
<td>Hospital B versus A (Process)</td>
<td>5%</td>
</tr>
<tr>
<td>Hospital B versus A (Outcome)</td>
<td>4%</td>
</tr>
<tr>
<td><strong>Extra Lives Saved per 100 Eligible Infants Treated in</strong></td>
<td><strong>Hospital B versus A</strong></td>
</tr>
<tr>
<td>Extra Lives Saved (Process)</td>
<td>5</td>
</tr>
<tr>
<td>Extra Lives Saved (Outcome)</td>
<td>4</td>
</tr>
<tr>
<td><strong>Sample Size</strong></td>
<td><strong>Needed to Detect a Difference in Use of Surfactant</strong></td>
</tr>
<tr>
<td>Extra Lives Saved (Process)</td>
<td>8</td>
</tr>
<tr>
<td>Extra Lives Saved (Outcome)</td>
<td>1970</td>
</tr>
</tbody>
</table>

* Sample size calculated assuming 80% power with \(P = .05\).

* Based on a meta-analysis of five RCTs comparing a policy of prophylactic versus rescue surfactant in 1207 infants, which suggested a relative risk reduction of 0.05 (95% confidence interval 0.01–0.09). Soll RF, Morley CJ. Prophylactic surfactant versus treatment with surfactant. (Cochrane Review) In: The Cochrane Library, Issue 2. Oxford: Update Software; 1998. Updated quarterly.
only 77% of the time if it had 280 admissions. A NICU with average birth weight-specific mortality was wrongly identified as substandard 16%, 12%, or 2% of the time if it had 35, 70, or 280 admissions, respectively. This means that several years of data may be required, by which time both technology and performance may have changed.

Subsequent evidence has supported this prediction. First, in a multicenter cross-sectional comparison of birth weight-adjusted mortality rates between 13 NICUs in the United Kingdom, confidence intervals remained wide and overlapping after adjusting for illness severity using CRIB. Second, Parry et al performed a longitudinal comparison of hospital mortality over 6 years among 9 NICUs. Whether ranked by crude or risk-adjusted mortality rates, the apparent performance of individual NICUs fluctuated substantially from year to year, and was therefore insufficiently stable to represent a credible argument for intervention. When all six years were aggregated, one hospital did perform significantly better than expected. The reasons for its success can only be speculated on. The authors concluded that annual comparisons of risk-adjusted mortality rates may not be reliable indicators of performance and that any action prompted by these underpowered comparisons would have been equally likely to have been beneficial, detrimental, or irrelevant.

Mandated Benchmarking in the United Kingdom

The United Kingdom General Medical Council recently disciplined two pediatric cardiac surgeons for failing to take adequate action over their excessive mortality rates. In response, the Secretary of State for Health announced plans to introduce compulsory, risk-adjusted audit in every hospital and every specialty. The imminent introduction of regular, public benchmarking for individual units marks a watershed for clinical audit in the United Kingdom. The usefulness of public disclosure of NICU outcomes is being debated in California (Dr Jeff Gould, personal communication, August 19, 1998). Because most tests assess a hospital’s performance against a state norm, the use of rankings is not appropriate. Given the above discussion on sample size, the most that may be possible is to indicate if a hospital is performing significantly better or worse than the state norm. Neonatologists must insist that the most reliable methods of audit are adopted to insure that infants consistently get high standards of care, and that unacceptable deviations from good practice are quickly corrected. It is also crucial that individual NICUs are not wrongly labeled as failures that could have unintended negative effects, damaging their ability to recruit staff and placing their patients at increased risk.

Benchmarking by Comparing Process Rather Than Outcome

Instead of outcomes, it may be more useful to compare the implementation of measures of process, particularly those proven effective in randomized, controlled trials. Mant and Hicks showed that comparing the appropriate use of proven treatments for myocardial infarction would substantially reduce the number of patients and the time required to identify significant differences in hospital performance. However, comparisons are only valid if the include all patients eligible to receive the treatments, at all hospitals. Take the example of prophylactic surfactant. Given the evidence in favor of it, an audit could rapidly detect the differences in proportions of eligible infants receiving prophylactic surfactant among hospitals, long before finding differences in mortality rates. Table 6 suggests that differences in performance between hospitals A and B could be detected much more rapidly and reliably by comparing the proportion of eligible infants receiving prophylactic surfactant (process) than by comparing mortality rates (outcome). Similar process benchmarking could be applied to use of antenatal corticosteroid administration.

Benchmarking by Identifying Effective Organizational Policies

Another alternative to monitoring performance would be to undertake prospective studies of outcomes in large groups of hospitals and test whether improvements in outcome are associated with differences in prespecified organizational characteristics or policies. This approach has been adopted in the United Kingdom Neonatal Staffing Study, a prospective cohort study of risk-adjusted outcomes in relation to volume of patients, staffing policy, and daily staff workload. Investigators selected a random, stratified cohort of 54 neonatal units using a factorial design that compares performance of groups of NICUs (see Table 7). Prospective studies of this kind could demonstrate organizational characteristics likely to improve outcome more reliably and rapidly than comparisons of outcome in individual NICUs. Institutions could then become more accountable to the public by showing that their policies are based on reliable evidence.

**SUMMARY**

We can learn what is achievable with current technologies to by comparing our NICU outcomes with
others. Such comparisons can highlight particularly good or poor results. These findings should spur quality improvement efforts that can lead to dramatic, system-wide improvement in outcomes. The value of such comparisons, however, rests on both accurate data and rigorous risk adjustment. The emergence of electronic medical records holds the promise of more accurate, data-intensive risk adjustment algorithms, and a greater richness in intermediate outcome markers. However, risk adjustment cannot substitute for careful definition and data collection. The methodologies of risk adjustment will inevitably require periodic recalibration as the association between risk and bad outcome is progressively blunted by successful technologies. We need better systems of capturing longer term outcomes against which to measure risk. Finally, we must also remain cautious not to overinterpret and to act prematurely on performance rankings, which are subject to both measurement biases and the play of chance.

REFERENCES

1. Iezzoni LI. Risk Adjustment for Measuring Healthcare Outcomes. 2nd ed. Chicago IL: Health Administration Press; 1997
2. Iezzoni LI. The risks of risk adjustment. JAMA. 1997;278:1600–1607
23. Parker RA, Lindstrom DP, Cotton RB. Improved survival accounts for most, but not all, of the increase in bronchopulmonary dysplasia. Pediatrics. 1992;90:663–668

Supplement


<table>
<thead>
<tr>
<th>Updated Information &amp; Services</th>
<th>including high resolution figures, can be found at: <a href="http://pediatrics.aappublications.org/content/103/Supplement_E1/255">http://pediatrics.aappublications.org/content/103/Supplement_E1/255</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>References</td>
<td>This article cites 65 articles, 15 of which you can access for free at: <a href="http://pediatrics.aappublications.org/content/103/Supplement_E1/255.full#ref-list-1">http://pediatrics.aappublications.org/content/103/Supplement_E1/255.full#ref-list-1</a></td>
</tr>
</tbody>
</table>
| Subspecialty Collections | This article, along with others on similar topics, appears in the following collection(s):  
**Medical Education**  
[http://classic.pediatrics.aappublications.org/cgi/collection/medical_education_sub](http://classic.pediatrics.aappublications.org/cgi/collection/medical_education_sub)  
**Quality Improvement**  
[http://classic.pediatrics.aappublications.org/cgi/collection/quality_improvement_sub](http://classic.pediatrics.aappublications.org/cgi/collection/quality_improvement_sub)  
**Fetus/Newborn Infant**  
[http://classic.pediatrics.aappublications.org/cgi/collection/fetus:newborn_infant_sub](http://classic.pediatrics.aappublications.org/cgi/collection/fetus:newborn_infant_sub) |
| Permissions & Licensing | Information about reproducing this article in parts (figures, tables) or in its entirety can be found online at: [https://shop.aap.org/licensing-permissions/](https://shop.aap.org/licensing-permissions/) |
| Reprints | Information about ordering reprints can be found online: [http://classic.pediatrics.aappublications.org/content/reprints](http://classic.pediatrics.aappublications.org/content/reprints) |