Neonatal disease severity scoring systems

J S Dorling, D J Field, B Manktelow

Illness severity scores have become widely used in neonatal intensive care. Primarily this has been to adjust the mortality observed in a particular hospital or population for the morbidity of its infants, and hence allow standardised comparisons to be performed. However, although risk correction has become relatively commonplace in relation to audit and research involving groups of infants, the use of such scores in giving prognostic information to parents, about their baby, has been much more limited. The strengths and weaknesses of the existing methods of disease severity correction in the newborn are presented in this review.

There are various scores devised for neonates in acute physiology-perinatal extension; VLBW, very low birthweight.

DERIVATION OF ILLNESS SEVERITY SCORING SYSTEMS

Although it may be possible to derive a risk adjustment score in a particular study, investigators will often require a readymade score. They may lack the data, resources, time, funding, or expertise required to develop their own, and a previously validated score also has the advantage that it is more likely to be accepted by others. There are various scores devised for neonates in the medical literature, and some of these will be described later. The choice of which variables are to be included in the score and their relative weights is obviously vital. A balance needs to be drawn between a complex score including many variables, and therefore difficult to complete, and a simpler model that may be easier to use but not as accurate. It also needs to be remembered that no score can completely quantify the complex factors that make up an individual infant’s morbidity.

Usually, scores are created in one of two ways. “Medical” scores are derived by an expert panel using clinical knowledge to select the variables to be included in the score and their relative weights. Alternatively, collected data are used in statistical models to produce “statistical” scores by identifying which variables have strong association with the outcome of interest and their relative weights. There is evidence that, in the long run, statistical scores outperform purely medical scores and today most scores are statistical as there are often relevant data available. However, clinical knowledge may, indeed should, contribute to the choice of variables included in a final model; not just because the model is then likely to perform better with other groups of infants but because it will be seen as more reliable by users.

STATISTICAL AND RESEARCH CONSIDERATIONS

However the score is derived, it is important that it has been validated to confirm that it predicts future events, preferably in a different dataset, with an adequate accuracy (calibration). Although a detailed discussion on methods for validating a score is beyond the scope of this review, it is important to remember that, for the score to be clinically useful, the predicted and observed event rates should closely match. Calibration can be investigated in a number of ways, most commonly using the Hosmer and Lemeshow goodness of fit test. With this test the observations are categorised into groups according to their predicted risk. The number of predicted and observed outcomes within each of these groups are then compared. A well calibrated score produces no statistically significant difference between these (usually p > 0.05). Often scores are re-calibrated to more closely fit hospital-specific data.

Abbreviations: Az, area under the ROC curve; CRIB, clinical risk index for babies; FiO2, fractional inspired concentration of oxygen; NIBRS, neurological risk score; NTISS, neonatal therapeutic intervention scoring system; PO2, partial pressure of oxygen; ROC curve, receiver operating characteristic curve; SNAP, score for neonatal acute physiology; SNAP-PE, score for neonatal acute physiology-perinatal extension; VLBW, very low birthweight.
match a local population by using the score as a variable in a new statistical regression model.

The ability of a score to differentiate between infants with different outcomes (discrimination) is also important, as good calibration cannot be achieved without good discrimination. Discrimination is measured by the area under the receiver operating characteristic (ROC) curve obtained by plotting the true positive rate against the false positive rate for the full range of values. The area under this curve indicates the overall discriminatory ability of a scoring system. An ideal test would have an area of 1.0—that is, no false positives or false negatives—whereas a score no better than chance alone has the value 0.5. A value above 0.8 is often taken to indicate that the score may be useful in practice.

Reproducibility is also an important feature of scores. Scores that are to be used in risk correction must be highly reproducible, both between individuals and when an individual rescores the data. If scores are not closely reproducible, then concern must exist about the potential introduction of bias when scores are used to enable comparisons.

**USING SCORES TO PREDICT AN INDIVIDUAL'S OUTCOME**

Using data on individuals to prognosticate about outcome is commonplace—for example, a birth weight of under 500 g is often used as a reason for not starting intensive care. However, the use of more complex prognostic scoring systems in other circumstances is controversial, raising both legal and ethical concerns. From a practical point of view, there are major difficulties. Using different risk scores may give similar group predictions, but individual estimates can differ significantly, lessening the usefulness of a score in a clinical situation.

Predicting an individual’s prognosis, either for counselling or for stratifying infants into a study, requires the most up to date information on the infant’s condition regardless of the influence of the care received. Limiting the data used to those collected within the first few hours of life, when additional information is available on the infant’s later progress, is likely to reduce the precision and accuracy of any such prediction. This is a common problem with the use of scores; indeed clinical risk index for babies (CRIB) and score for neonatal acute physiology (SNAP) are limited to 12 and 24 hours respectively and are therefore poor predictors of individual outcome.

On an individual basis, clinicians may be able to prognosticate as accurately as any scoring system as they can take account of the full and changing clinical picture of a child. Stevens and colleagues showed that clinicians are good at identifying high risk infants but tend to overestimate the risk of death (in other words they provide good discrimination but poor calibration). This warrants further investigation as clinical prognostications are often used in end of life decisions. It is possible that combining clinicians’ assessments with a scoring system could improve the accuracy of risk assessment. Although this may be important in clinical practice for individuals, using clinicians’ views for group predictions and research purposes would introduce an unacceptable level of subjectivity and potential bias.

**USING DISEASE SCORES FOR GROUP PREDICTIONS AND FOR COMPARING UNITS**

For comparison of outcomes across different neonatal intensive care units, the need to adequately adjust outcomes for differences in case mix (risk adjustment) is well recognised. A unit tending to treat only those patients with good prognoses would be expected to have a high rate of “good” outcome. Conversely those treating patients with poor prognoses would expect a higher rate of “poor” outcome. As put by Poloniecki, risk adjustment tries to help answer the question, “Is it you, Doc, or your patients, who are below average?” This methodology is likely to be used increasingly for comparing outcomes over time and between units since the Kennedy report into Paediatric Cardiac Surgery.

In these circumstances a score should quantify the morbidity of the infant when it first arrives into the charge of the unit, before care can influence its condition or its score. Clearly the quality of care received antenatally or during resuscitation may be important and cannot easily be corrected for by a scoring system. Even if basic birth details such as weight and gestational age are used on their own, differing policies on who to resuscitate can affect comparisons between units. Although data collected a short time after admission (up to 24 hours) may produce better discriminating models than data collected solely at birth, including information that is influenced by care can be problematic. For example, if a score that includes the inspired oxygen concentration is used (such as CRIB), an infant given more oxygen than necessary would score more points than if it had been appropriately treated. The scoring system would thus predict a poorer prognosis for this infant. This raises the expected number of deaths for that unit and falsely makes its performance look better. Including such variables also offers the opportunity to intentionally manipulate the score and hence the predicted outcomes.

In addition to comparing mortality—for example, in Scotland and Australia—disease severity scores have also been used to investigate other outcomes, such as narcotic administration, blood transfusion rates, and retinopathy of prematurity. Although in such circumstances some scores may work well, care is required when using a score to investigate an outcome for which it was not designed. It is unlikely that the risk factors for one outcome (say, mortality) are identical with those for another (the need for blood transfusion, for example).

**SCORES USED IN PREDICTING MORTALITY**

A variety of risk adjustment scores have been derived and advocated for use in assessing neonatal mortality. Full details of each scoring system are given in the papers cited although details on which variables are used are included in table 2. Each of these scores will be briefly described.

**CLINICAL RISK INDEX FOR BABIES (CRIB)**

The CRIB score was created to predict mortality for infants born at less than 32 weeks gestation at birth and was derived using data from infants admitted to four UK tertiary neonatal units from 1988 to 1990. The derivation cohort contained...
very low birthweight (VLBW) infants, of whom 25% died. The authors used logistic regression to identify the six variables most predictive of mortality (table 2). The final score is based on a weighted sum of these six factors. In the original study, the score had good discriminatory ability (area under the ROC curve: $Az = 0.90$), considerably better than birth weight alone ($Az = 0.78$). Other studies have produced similar values for the area under the ROC curve using CRIB: $Az = 0.87–0.90$.\(^{19} 21\)

The ease of data collection is a major advantage of CRIB, as calculation takes five minutes per infant, compared with 20–30 minutes for some of the more complex scores such as SNAP, SNAP-PE, and the NTISS.\(^{23}\) A further advantage is that CRIB is assessed over the first 12 hours of life, making it less susceptible to treatment effects than some other scores.

### CRIB II

CRIB II, an improved version of CRIB, was published recently.\(^{23}\) It uses a previously published grid predicting mortality by gestational age and birth weight together with admission temperature and base excess to predict mortality. The new score was intended to improve predictions for smaller, very premature infants and to exclude variables that could be influenced by care given to the infant.
appropriateness of including admission temperature remains to be proven, as this could clearly be affected by several aspects of care. Further validation of CRIB II is awaited.

SCORE FOR NEONATAL ACUTE PHYSIOLOGY (SNAP)

SNAP, the principal alternative to CRIB, was developed using data from three units in Boston, USA in 1990.24 The derivation cohort contained 1643 infants; 154 weighed less than 1500 g at birth. This score is applicable to any infant admitted to a neonatal unit, but, because of the small number of VLBW infants in the population from which it was derived, it has reduced sensitivity to differences between the most premature infants.25 SNAP scores are based on 28 items collected over the first 24 hours of life from a variety of sources including every body system and selected blood test results. Unlike the CRIB score, where parameters are weighted according to their statistical relation to death, the variables were weighted according to expert opinion, with a score of 0, 1, 3, or 5 assigned to each variable. The original cohort was also used to extend SNAP to form the SNAP-PE score (score for neonatal acute physiology—perinatal extension) by adding birth weight, small for gestational age (weight <5th centile for gestation), and low Apgar score at five minutes.25 Although the SNAP score assesses many body systems, and is able to predict death well, it is much more difficult to collect than the CRIB score. In Richardson’s comparison, SNAP predicted death better than birth weight alone (Az 0.87 v 0.77), and SNAP-PE was even better (Az 0.91) and calibration (Hosmer-Lemeshow 0.90) for SNAPPE-II in predicting mortality.

SNAP-II AND SNAPPE-II

Because of the difficulty of data collection for the SNAP and SNAP-PE scores, the original authors have recently produced simpler versions using data from 30 North American units.26 The derivation and validation cohorts were impressively large: 10 819 and 14 610 respectively. Changes included shortening the period of data collection to 12 hours and reducing the number of variables to six (mean blood pressure, lowest temperature, P02/FiO2 ratio, serum pH, multiple seizures, and urine output). These factors were assessed as having the strongest statistical association with mortality.

As with the original SNAP score, SNAP II was also extended to produce the SNAPPE-II by adding the perinatal extension factors. SNAP-II and SNAPPE-II are likely to be as easy as CRIB to collect, and they have been developed from very large cohorts of all birth weights during the second half of the 1990s. Richardson showed good discrimination (Az 0.91) and calibration (Hosmer-Lemeshow 0.90) for SNAPPE-II in predicting mortality.

NATIONAL THERAPEUTIC INTERVENTION SCORING SYSTEM (NTISS)

NTISS27 was published in 1992 and was derived by an expert panel as a modification of the adult intensive care score, therapeutic intervention scoring system. NTISS is unusual as it is based on the treatments received by an infant rather than measuring pathophysiological factors. As treatment depends on policy and practice in units, it can vary greatly,28 and it is not possible to compare units using this type of adjustment.

NATIONAL INSTITUTE OF CHILD HEALTH AND HUMAN DEVELOPMENT (NICHD)

The NICHD score was created using factors noted at admission to seven neonatal units in the United States from 1823 infants born from 1987 to 1989 and weighing 501–1500 g.29 Logistic regression was used to select the variables, with validation using another 1780 infants. It has not been used extensively since development.

BERLIN SCORE

This German score was developed using logistic regression methods with 396 VLBW development infants and 176 VLBW

---

**Table 3** Neurodisability predictive ability of the clinical risk index for babies (CRIB) score, with and without ultrasound (US)

<table>
<thead>
<tr>
<th>Score</th>
<th>Age at assessment</th>
<th>Number of infants assessed</th>
<th>Method of developmental assessment</th>
<th>Outcome</th>
<th>Predictive value (Az)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crib</td>
<td>12 months</td>
<td>351</td>
<td>Griffith’s test</td>
<td>Major impairment</td>
<td>0.703</td>
<td>33</td>
</tr>
<tr>
<td>Crib</td>
<td>18 months</td>
<td>695</td>
<td>Questionnaires from doctors, health visitors, and community nurses</td>
<td>Death or impairment</td>
<td>0.83</td>
<td>34</td>
</tr>
<tr>
<td>Crib</td>
<td>18 months</td>
<td>81</td>
<td>Apgar method and Bayley development scales</td>
<td>Major disability</td>
<td>0.77</td>
<td>35</td>
</tr>
<tr>
<td>Crib</td>
<td>24 months</td>
<td>398</td>
<td>Health visitor: standardised questionnaire</td>
<td>Severe disability</td>
<td>0.71</td>
<td>36</td>
</tr>
<tr>
<td>Crib &amp; cranial US at 72 h</td>
<td>18 months</td>
<td>240</td>
<td>Health visitor completed questionnaire</td>
<td>Severe disability</td>
<td>0.89</td>
<td>37</td>
</tr>
</tbody>
</table>

---

**Table 4** Neurodisability predictive ability of nursery neurobiologic risk score (NBRS)

<table>
<thead>
<tr>
<th>Score</th>
<th>Age at assessment</th>
<th>Score value</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Positive predictive value (%)</th>
<th>Negative predictive value (%)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBRS</td>
<td>24 months</td>
<td>5 or more</td>
<td>52</td>
<td>100</td>
<td>100</td>
<td>84</td>
<td>39</td>
</tr>
<tr>
<td>NBRS</td>
<td>18 months</td>
<td>5 or more</td>
<td>81</td>
<td>54</td>
<td>49</td>
<td>84</td>
<td>37</td>
</tr>
<tr>
<td>NBRS</td>
<td>8 or more</td>
<td>87</td>
<td>56</td>
<td>87</td>
<td>71</td>
<td>78</td>
<td>37</td>
</tr>
<tr>
<td>Modified NBRS</td>
<td>3 years</td>
<td>“high” NBRS</td>
<td>100</td>
<td>98</td>
<td>92</td>
<td>100</td>
<td>42</td>
</tr>
</tbody>
</table>

---

Any handicap Major handicap

<table>
<thead>
<tr>
<th>Score</th>
<th>Age at assessment</th>
<th>Score value</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Positive predictive value (%)</th>
<th>Negative predictive value (%)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBRS</td>
<td>12 months</td>
<td>≤5</td>
<td>20</td>
<td>5</td>
<td>5</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>NBRS</td>
<td>5 to 7</td>
<td>41</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NBRS</td>
<td>8 or more</td>
<td>95</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

www.archdischild.com
validation infants from 1988 to 1991. It suffers from the inclusion of a number of subjective factors. The inclusion of these data items limits its role as a means of objective comparison between units.

NEONATAL MORTALITY PROGNOSIS INDEX (NMPI)
This score was derived using logistic regression to select prognostic factors collected up to 12 hours after admission from 336 Mexican infants in 1993. The model was validated in an additional cohort of 300 infants. It has not been widely used.

SCORES USED IN PREDICTING NEURODISABILITY
Three risk adjustment scores have been assessed for use in predicting later neurodisability after neonatal intensive care. With the improvements that have been seen in survival, there is increasing interest in long term outcomes after neonatal care. Methods for neurodisability risk correction would be a valuable step forward. The currently available systems are briefly detailed below and summarised in tables 3 and 4. For further information please see the cited articles.

CRIB SCORE AND NEUROLOGICAL MORBIDITY
Four publications have examined the use of the CRIB score for predicting neurodevelopmental outcome. Table 3 summarises the results from these studies. Data on the outcome of 695 infants from the derivation cohort suggested that CRIB could predict a combined outcome of death or impairment. However, in a further study containing infants from the original study, a close relation between CRIB at 12 hours and severe disability at 24 months of age was not demonstrated.

Two studies not containing infants from the original cohort revealed that CRIB discriminated poorly in the role of predicting outcome at 12 months (Az = 0.70) and 18 months (0.77). Lago et al also found that birth weight alone was similar (Az = 0.70), and gestational age alone was better (Az = 0.83) than CRIB. These studies may be difficult to interpret, as neurodevelopmental testing before 2 years probably fails to detect all affected infants.

Fowlie et al combined CRIB with cranial ultrasonography in 297 infants from the original cohort surviving beyond 72 hours. CRIB scoring was performed at 72 hours, with ultrason appearance from “around” 72 hours. Ninety nine infants had missing CRIB, ultrasound, or follow up data. A CRIB score greater than 4 with a grade 3 or 4 intraventricular haemorrhage was predictive of severe disability, but there were only five infants in this group. In comparison with birth weight (Az = 0.70) and gestational age (Az = 0.74), CRIB and ultrasonography improved the model’s discrimination (Az = 0.89). To implement this simple approach would require an alteration to current practice for collecting CRIB scores and, probably, ultrasound data. In addition interpretations of cranial ultrasound findings have been shown to vary between clinicians.

SNAP AND NEUROLOGICAL MORBIDITY
A retrospective case note review of 173 inborn infants from Minnesota examined the ability of the SNAP score to predict neurological outcome in premature infants born in 1993 and 1994 before 30 weeks gestation. A score was collected for every day of each admission to produce a “cumulative SNAP score”. This was then examined in relation to assessments at around 1 year of life and during the 3rd year of life. Although the authors did not use ROC curve analysis, they did show that the quartile of infants with the worst cumulative SNAP score had significantly lower motor development indices at 1 year as well as lower psychomotor development indices at both assessments.

REFERENCES
1 Signorini DF, Weir NU. Any variability in outcome comparisons adjusted for case mix must be accounted for. BMJ 1999;318:128.
38 Mattia FR, Deregnier RAO. Chronic physiologic instability is associated with neurodevelopmental morbidity at one and two years in extremely premature infants. Pediatrics 1999;103:e35.