REVIEW

Neonatal disease severity scoring systems

J S Dorling, D J Field, B Manktelow

Arch Dis Child Fetal Neonatal Ed 2005;**90**:F11–F16. doi: 10.1136/adc.2003.048488

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 their 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 many situations when a clinician, parent, nurse, manager, or researcher may wish to quantify the morbidity of a neonate. This may be to try to explain in terms of case mix differences the wide variations in mortality and other outcomes seen between different neonatal intensive care units.[1] Alternatively, it may be the estimated probability of a specific outcome in a particular infant that is of interest, or the need to identify high risk infants suitable for a particular intervention or for inclusion in a clinical trial. These and other problems shown in table 1 can be tackled by using an illness severity score.

Scoring systems involve using appropriately weighted demographic, physiological, and clinical data collected on the infant to calculate a score that quantifies its morbidity. The principle for such an approach has been long established in many branches of medicine.[2] The desirable properties of neonatal scores have been described as including: ''(1) ease of use; (2) applicability early in the course of hospitalisation; (3) ability to reproducibly predict mortality, specific morbidities, or cost for various categories of neonates; (4) usefulness for all groups of neonates to be described.''[3] However, these properties are difficult, perhaps impossible, to achieve completely.

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,[4] 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.[5] Calibration can be investigated in a number of ways, most commonly using the Hosmer and Lemeshow goodness of fit test.[6] 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 recalibrated to more closely

See end of article for authors' affiliations

Correspondence to: Dr Dorling, Neonatal Unit, Leicester Royal Infirmary, Leicester LE1 5WW, UK; JSD10@le.ac.uk

Accepted 16 March 2004

Abbreviations: Az, area under the ROC curve; CRIB, clinical risk index for babies; FIO_2, fractional inspired concentration of oxygen; NBRS, neurobiological risk score; NTISS, neonatal therapeutic intervention scoring system; PO_2, 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

Table 1 Research uses of predictive scores in neonatology

Group predictions
1. Comparing study groups for similarity of risk
2. Auditing the severity of illness in different units
3. Comparing performance of different units
4. Determining trends in results over time
5. Reviewing if infants are treated appropriately for risk (e.g. number of septic screens or ventilation days)
6. Comparing rates of complications; are some preventable?

Individual predictions
1. Giving prognostic information
2. Stratifying infants in trials (to ensure similarity of risk)
3. Determining individual treatment

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[7] 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.[8]

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.[9] 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[10] 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.[10] 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.[1] A unit tending to treat only those patients with good prognoses would be expected to have a high rate of ''good'' outcome.[11] Conversely those treating patients with poor prognoses would expect a higher rate of ''poor'' outcome. As put by Poloniecki,[12] 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.[13]

In these circumstances a score should quantify the morbidity of the infant when it first arrives into the charge of the unit, before care given 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,[9] 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.[9]

In addition to comparing mortality—for example, in Scotland and Australia[14]—disease severity scores have also been used to investigate other outcomes, such as narcotic administration,[15] blood transfusion rates,[16] and retinopathy of prematurity.[17] 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.[18] The derivation cohort contained

Table 2 Scoring systems variables

CRIB	SNAP	NTISS
Birth weight	Blood pressure	Supplemental oxygen
Gestation	Heart rate	Surfactant administration
Congenital malformation	Respiratory rate	Tracheostomy care
Maximum base deficit in first 12 h	Temperature	Tracheostomy placement
Minimum appropriate FIO_2 in first 12 h	PO_2	CPAP administration
Maximum appropriate FIO_2 in first 12 h	PO_2/FIO_2 ratio	Endotracheal intubation
CRIB II	PCO_2	Mechanical ventilation
Birth weight by gestation	Oxygenation index	Mechanical ventilation with paralysis
Maximum base deficit in first 12 h	Packed cell volume	High frequency ventilation
Sex	White blood cell count	Extracorporeal membrane oxygenation
Admission temperature	Immature total ratio	Indomethacin administration
Berlin score	Absolute neutrophil count	Volume expansion
Birth weight	Platelet count	Vasopressor administration
Grade of RDS	Blood urea nitrogen	Pacemaker on standby
Apgar score at 5 min	Creatinine	Pacemaker used
Artificial ventilation	Urine output	Cardiopulmonary resuscitation
Base excess at admission	Indirect bilirubin	Antibiotics
NICHHD score	Direct bilirubin	Diuretics (enteral)
Birth weight	Sodium	Steroids (postnatal)
Small for gestational age	Potassium	Anticonvulsant
Race	Calcium (ionised)	Aminophylline
Sex	Calcium (total)	Other unscheduled medication
Apgar score at 1 min	Glucose	Diuretics (parenteral)
NMPI	Serum bicarbonate	Treatment of metabolic acidosis
Gestational age	Serum pH	Potassium binding resin
Birth weight	Seizure	Frequent vital signs
Cardiac arrest	Apnoea	Cardiorespiratory monitoring
PaO_2/FIO_2 ratio	Stool guaiac	Phlebotomy
Major congenital malformations	**SNAP-PE**	Thermoregulated environment
Sepsis	SNAP score plus :	Non-invasive oxygen monitoring
Base excess	Birth weight	Arterial pressure monitoring
SINKIN 12 hour	Apgar score<7 at 5 min	CVP monitoring
Birth weight	Small for gestational age	Urinary catheter
Gestational age	**SNAP-II**	Quantitative intake and output
Apgar score at 5 min	Mean blood pressure	Gavage feeding
Peak inspiratory pressure at 12 h	Lowest temperature	Intravenous fat emulsion
NBRS	PO_2/FIO_2 ratio	Intravenous amino acid solution
Blood pH	Serum pH	Phototherapy
Hypoglycaemia	Multiple seizures	Insulin administration
Intraventricular haemorrhage	Urine output	Potassium infusion
Periventricular leucomalacia	**SNAPPE-II**	Transfusion
Seizures	SNAP II score plus :	Intravenous γ globulin
Infection	Birth weight $\leqslant$749 g	Red blood cell transfusion
Need for mechanical ventilation	Apgar<7 at 5 min	Partial volume exchange transfusion
	Small for gestational age	Platelet transfusion
		White blood cell transfusion
		Double blood cell transfusion
		Transport of patient
		Chest tube
		Minor operation
		Thoracentesis
		Major operation
		Pericardiocentesis
		Pericardial tube
		Dialysis
		Vascular access
		Peripheral intravenous line
		Arterial line
		Central venous line

CRIB, clinical risk index for babies; FIO_2, fractional inspired concentration of oxygen; NBRS, neurobiological risk score; NMPI, neonatal mortality prognosis index; NTISS, neonatal therapeutic intervention scoring system; PO_2, partial pressure of oxygen; RDS, respiratory distress syndrome; SNAP, score for neonatal acute physiology; SNAP-PE, score for neonatal acute physiology-perinatal extension.

812 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).[18–20] 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.[22] 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. The

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

	Age at assessment (months)	Number of infants assessed	Method of developmental assessment	Outcome	Predictive value (Az)	Reference
CRIB	12	351	Griffith's test	Major impairment	0.703	33
CRIB	18	695	Questionnaires from doctors, health visitors, and community nurses	Death or impairment	0.83	34
CRIB	18	81	Amiel-Tison method and Bayley development scales	Major disability	0.77	35
CRIB	24	398	Health visitor: standardised questionnaire	Severe disability	0.71	36
CRIB & cranial US at 72 h	18	240	Health visitor completed questionnaire	Severe disability	0.89	37

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 0.93).[25]

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, Po_2/Fio_2 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)

NTISS[27] 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 (NICHHD)

The NICHHD 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 4 Neurodisability predictive ability of nursery neurobiologic risk score (NBRS)

Score	Age at assessment	Score value	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)	Reference
NBRS	24 months	5 or more	52	100	100		39
NBRS	18 months	5 or more	81	54	49	84	37
		8 or more	56	87	71	78	
Modified NBRS	3 years	"high" NBRS	100	98	92	100	42
			Any handicap	**Major handicap**			
NBRS	12 months	<5	20	5			40
		5 to 7	41	23			
		8 or more	95	80			

validation infants from 1988 to 1991.[30] 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.[31] 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 neurodisabilty 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.[35–36] 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.[34] 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.[36]

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),[33] and 18 months (0.77).[35] Lago *et al*[35] 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*[37] combined CRIB with cranial ultrasonography in 297 infants from the original cohort surviving beyond 72 hours. CRIB scoring was performed at 72 hours, with ultrasound appearances 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.[38] 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.

NURSERY NEUROBIOLOGIC RISK SCORE (NBRS)

The NBRS was developed for neurological prediction in VLBW infants.[39] Brazy *et al* chose and weighted 13 factors, correlating these with outcome in 57 infants at 24 months of age from 1986 to 1988. A ''revised NBRS'' was developed from the seven factors accounting for almost all of the differences in outcome (see table 2). Scored at 14 days of age, taking five minutes per infant, it was highly repeatable, with all infants scoring over 5 having abnormal development at 24 months corrected age. Table 4 summarises the use of the NBRS in predicting neurodisability.

Using this score, Nunes *et al*[40] studied 77 infants at 12 months of age. Of those infants with a score of 8 or more, 80% developed a major handicap. Lefebvre *et al*[41] retrospectively collected the NBRS and outcome at 18 months in 121 infants, obtaining remarkably different results from Brazy *et al*.[39] Lefebvre *et al*'s ROC curve value of 0.79 is similar to that of CRIB.[37] Contractor *et al*[42] analysed 3 year outcomes in 56 extremely premature infants, showing that a high NBRS at discharge was associated with four times the risk of an abnormal outcome. After modifying the score (to comprise acidosis, hypoxaemia, hypotension, intraventricular haemorrhage, infection, and hypoglycaemia), they also showed very good sensitivity and specificity.[42]

Although it is a reasonable predictor of neurological outcome, the NBRS cannot be used for risk adjustment because of the delayed timing of data collection and the consequent effect of care.

CONCLUSIONS

Illness severity scores are now well accepted as essential tools when comparing healthcare providers. When using an illness severity score, it is important to remain clear about the question being investigated to be sure that the scoring system being used is appropriate. The use of an existing score, developed for another purpose, simply because it is convenient is unlikely to represent the best approach. It is also important to remember that, even the best scoring systems are not completely accurate. No mathematical formula can completely capture the complex clinical processes in a neonate. The use of scores for predicting individual outcomes is fraught with difficulty, most particularly because of variation in the approach to clinical care adopted by different units (and even clinicians in the same unit) as well as important ethical and legal concerns. It is almost certainly these issues that have, rightly, limited the extent to which scoring systems have been used for individual risk prediction and counselling.

In the future, further adequately sized studies, perhaps testing new factors, are warranted both to confirm that our current risk adjustment tools are optimal and also to check that the scores are adequately recalibrated after changes in care. Further work is needed in relation to the use of risk correction scoring systems for comparisons of later health status.

Authors' affiliations

J S Dorling, D J Field, Department of Health Sciences, University of Leicester, Neonatal Unit, Leicester Royal Infirmary, Leicester LE1 5WW, UK

B Manktelow, Department of Health Sciences, University of Leicester, 22–28 Princess Road West, Leicester LE1 6TP, UK

Competing interests: none declared

REFERENCES

1 **Signorini DF**, Weir NU. Any variability in outcome comparisons adjusted for case mix must be accounted for. *BMJ* 1999;**318**:128.
2 **Ridley SA**. Uncertainty and scoring systems. *Anaesthesia* 2002;**57**:761–7.

3 **Fleisher BE**, Murthy L, Lee S, *et al*. Neonatal severity of illness scoring systems: a comparison. *Clin Pediatr* 1997;**36**:223–7.
4 **Rosenthal GE**, Harper DL. Cleveland health quality choice: a model for collaborative community-based outcome assessment. *Jt Comm J Qual Improv* 1994;**20**:425–2.
5 **Altman DG**, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;**19**:453–73.
6 **Hosmer DW**, Lemeshow S. *Applied logistic regression*. 2nd ed. New York: Whiley, 2000:147–56.
7 **Van Erkel AR**, Pattynama PMT. Receiver Operating Characteristic (ROC) analysis: basic principles and applications in radiology. *Eur J Radiol* 1998;**27**:88–94.
8 **Iezzoni LI**, Ash AS, Shwartz M, *et al*. Judging hospitals by severity-adjusted mortality rates: the influence of the severity-adjustment method. *Am J Public Health* 1996;**86**:1379–87.
9 **Pollack MM**, Koch MA, Bartel DA, *et al*. A comparison of neonatal mortality risk prediction models in very low birth weight infants. *Pediatrics* 2000;**105**:1051–7.
10 **Stevens SM**, Richardson DK, Gray JE, *et al*. Estimating neonatal-mortality risk: an analysis of clinician judgments. *Pediatrics* 1994;**93**:945–50.
11 **Field D**, Draper ES. Survival and place of delivery following preterm birth: 1994–96. *Arch Dis Childhood Fetal Neonatal Ed* 1999;**80**:F111–14.
12 **Poloniecki J**. Half of all doctors are below average. *BMJ* 1998;**316**:1734–6.
13 Learning from Bristol: the report of the public inquiry into children's heart surgery at the Bristol Royal Infirmary 1984–1995. Bristol Royal Infirmary Inquiry 2001. www.bristol-inquiry.org.uk/final_report/Summary.pdf (accessed 30 Nov 2003).
14 **International Neonatal Network SNCNCSG**. Risk adjusted and population based studies of the outcome for high risk infants in Scotland and Australia. *Arch Dis Child Fetal Neonatal Ed* 2000;**82**:F118–23.
15 **Kahn DJ**, Richardson DK, Gray JE, *et al*. Variation among neonatal intensive care units in narcotic administration. *Arch Pediatr Adolesc Med* 1998;**152**:844–51.
16 **Bednarek FJ**, Weisberger S, Richardson DK, *et al*. Variations in blood transfusions among newborn intensive care units. *J Pediatr* 1998;**133**:601–7.
17 **Vyas J**, Field D, Draper ES, *et al*. Severe retinopathy of prematurity and its association with different rates of survival in infants of less than 1251g birth weight. *Arch Dis Child Fetal Neonatal Ed* 2000;**82**:F145–9.
18 **International Neonatal Network**. The CRIB (Clinical Risk Index for Babies) Score: a tool for assessing initial neonatal risk and comparing performance of neonatal intensive-care units. *Lancet* 1993;**342**:193–8.
19 **Rautonen J**, Makela A, Boyd H, *et al*. CRIB and SNAP: assessing the risk of death for preterm neonates. *Lancet* 1994;**343**:1272–3.
20 **Maier RF**, Caspar-Karweck UE, Grauel EL, *et al*. A comparison of two mortality risk scores for very low birthweight infants: Clinical Risk Index for Babies and Berlin Score. *Intens Care Med* 2002;**28**:1332–5.
21 **Lund GC**, Green D, Browne R, *et al*. New CRIB Score: one score for all NICU admissions. *Pediatr Res* 1997;**41**:162.
22 **Bastos G**, Gomes A, Oliveira P, *et al*. A comparison of 4 pregnancy assessment scales (CRIB, SNAP, SNAP-PE, NTISS) in premature newborns. Clinical Risk Index for Babies. Score for Neonatal Acute Physiology. Score for Neonatal Acute Physiology-Perinatal Extension. Neonatal Therapeutic Intervention Scoring System. *Acta Med Port* 1997;**10**:161–5.
23 **Parry G**, Tucker J, Tarnow-Mordi W. CRIB II: an update of the Clinical Risk Index for Babies score. *Lancet* 2003;**361**:1789–91.
24 **Richardson DK**, Gray JE, McCormick MC, *et al*. Score for Neonatal Acute Physiology: a physiological severity index for neonatal intensive-care. *Pediatrics* 1993;**91**:617–23.
25 **Richardson DK**, Phibbs CS, Gray JE, *et al*. Birth-weight and illness severity: independent predictors of neonatal-mortality. *Pediatrics* 1993;**91**:969–75.
26 **Richardson DK**, Corcoran JD, Escobar GJ, *et al*. SNAP-II and SNAPPE-II: simplified newborn illness severity and mortality risk scores. *J Pediatr* 2001;**138**:92–100.
27 **Gray JE**, Richardson DK, Mccormick MC, *et al*. Neonatal Therapeutic Intervention Scoring System: a therapy-based severity-of-illness index. *Pediatrics* 1992;**90**:561–7.
28 **Field D**, Manktelow B, Draper ES. Bench marking and performance management in neonatal care: easier said than done! *Arch Dis Child Fetal Neonatal Ed* 2002;**87**:F163–4.
29 **Horbar JD**, Onstad L, Wright E, *et al*. Predicting mortality risk for infants weighing 501 to 1500 grams at birth: a National Institutes of Health Neonatal Research Network Report. *Crit Care Med* 1993;**21**:12–18.
30 **Maier RF**, Rey M, Metze BC, *et al*. Comparison of mortality risk: a score for very low birthweight infants. *Arch Dis Child Fetal Neonatal Ed* 1997;**76**:F146–50.
31 **Garcia H**, Villegas-Silva R, Villanueva-Garcia D, *et al*. Validation of a prognostic index in the critically ill newborn. *Rev Invest Clin* 2000;**52**:406–14.
32 **Sinkin RA**, Cox C, Phelps DL. Predicting risk for bronchopulmonary dysplasia: selection criteria for clinical-trials. *Pediatrics* 1990;**86**:728–36.
33 **Buhrer C**, Grimmer I, Metze B, *et al*. The CRIB (Clinical Risk Index for Babies) score and neurodevelopmental impairment at one year corrected age in very low birth weight infants. *Intens Care Med* 2000;**26**:325–9.
34 **Pharoah POD**. Crib and impairment after neonatal intensive-care. *Lancet* 1995;**346**:58.
35 **Lago P**, Freato F, Bettiol T, *et al*. Is the CRIB score (Clinical Risk Index for Babies) a valid tool in predicting neurodevelopmental outcome in extremely low birth weight infants? *Biol Neonate* 1999;**76**:220–7.
36 **Fowlie PW**, Gould CR, Tarnow-Mordi WO, *et al*. Measurement properties of the Clinical Risk Index for Babies: reliability, validity beyond the first 12 hours, and responsiveness over 7 days. *Crit Care Med* 1998;**26**:163–8.
37 **Fowlie PW**, Tarnow-Mordi WO, Gould CR, *et al*. Predicting outcome in very low birthweight infants using an objective measure of illness severity and cranial ultrasound scanning. *Arch Dis Child Fetal Neonatal Ed* 1998;**78**:F175–8.
38 **Mattia FR**, Deregnier RAO. Chronic physiologic instability is associated with neurodevelopmental morbidity at one and two years in extremely premature infants. *Pediatrics* 1998;**102**:e35.
39 **Brazy JE**, Eckerman CO, Oehler JM, *et al*. Nursery Neurobiological Risk Score: important factors in predicting outcome in very-low-birth-weight infants. *J Pediatr* 1991;**118**:783–92.
40 **Nunes A**, Melo F, Silva JE, *et al*. Importance of J. Brazy's neurobiological index. Prediction of the number and severity of complications in very low birth weight infants. *Acta Med Port* 1998;**11**:615–21.
41 **Lefebvre F**, Gregoire MC, Dubois J, *et al*. Nursery Neurobiologic Risk Score and outcome at 18 months. *Acta Paediatr* 1998;**87**:751–7.
42 **Contractor CP**, Leslie GI, Bowen JR, *et al*. The Neonatal Neurobiologic Risk Score: does it predict outcome in very premature infants? *Indian Pediatr* 1996;**33**:95–101.