ORIGINAL ARTICLE

Assessing mortality risk in very low birthweight infants: a comparison of CRIB, CRIB-II, and SNAPPE-II

L Gagliardi, A Cavazza, A Brunelli, M Battaglioli, D Merazzi, F Tandoi, D Cella, G F Perotti, M Pelti, I Stucchi, F Frisone, A Avanzini, R Bellù, and the NNL study group

See end of article for authors' affiliations
....

Correspondence to:
Dr Gagliardi, Division of Neonatology and Paediatrics, Ospedale della Versilia, Via Aurelia 335, I-55043 Lido di Camaiore, Lucca, Italy; l.gagliardi@neonatalnet. org

Background: Illness severity scores are increasingly used for risk adjustment in clinical research and quality assessment. Recently, a simplified version of the score for neonatal acute physiology (SNAPPE-II) and a revised clinical risk index for babies (CRIB-II) score have been published.
Aim: To compare the discriminatory ability and goodness of fit of CRIB, CRIB-II, and SNAPPE-II in a cohort of neonates < 1500 g birth weight (VLBWI).
Methods: Data from 720 VLBWI, admitted to 12 neonatal units in Lombardy (Northern Italy) participating in a regional network, were analysed. The discriminatory ability of the scores was assessed measuring the area under the receiver operating characteristic curve (AUC). Outcome measure was in-hospital death.
Results: CRIB and CRIB-II showed greater discrimination than SNAPPE-II (AUC 0.90 and 0.91 v 0.84, p < 0.0004), partly because of the poor quality of some of the data required for the SNAPPE-II calculation—for example, urine output—but also because of the relative weight given to some items. In addition to each score, several variables significantly influenced survival in logistic regression models. Antenatal steroid prophylaxis, singleton birth, absence of congenital anomalies, and gestational age were independent predictors of survival for all scores, in addition to caesarean section and not being small for gestation (for SNAPPE-II) and a five minute Apgar score of ⩾ 7 (for SNAPPE-II and CRIB).
Conclusions: CRIB and CRIB-II had greater discriminatory ability than SNAPPE-II. Risk adjustment using all scores is imperfect, and other perinatal factors significantly influence VLBWI survival. CRIB-II seems to be less confounded by these factors.

Survival of very low birthweight infants (VLBWI) depends on birth weight (BW) and gestational age (GA), but also on other perinatal factors and physiological conditions of the individual infants, in particular disease severity in the first hours of life.[1] Illness severity scores were thus developed with the aim of quantifying the clinically obvious fact that infants of the same GA and BW differ in their risk of dying.[2][3] Although the initial goal of computing the risk of death for individual subjects has not been realised,[4] illness severity scores are increasingly used to allow fair comparisons of outcome across different hospitals by "adjusting outcome rates for initial severity of the illness"[4]—that is, for risk adjustment. The availability of reliable and valid instruments to measure the severity of illness may allow the conduction of unbiased comparisons in benchmarking[5] and quality of care studies. Moreover, they can better define populations of neonates within clinical trials, outcome evaluations, or resource utilisation studies.[6]

CRIB (clinical risk index for babies)[7] and SNAPPE (score for neonatal acute physiology—perinatal extension)[8] are the most commonly used scores, and their performance has been extensively validated.[2][3][6][9–12] However, both scores have some limitations and were developed almost a decade ago, before widespread use of surfactant and antenatal steroids, when mortality was higher.

SNAPPE[8] (developed and mainly used in the United States and Canada) can be applied to neonates of all BW and all GA, whereas CRIB[7] (developed in the United Kingdom and mainly used in Europe) can only be applied to VLBWI. CRIB, on the other hand, has the disadvantage of using some data that can be determined by the clinician. Apart from this, published data on which score is best are conflicting. Only three studies have compared CRIB and SNAPPE by applying them to the same set of neonates. The first was on 222 VLBWI admitted to a single neonatal intensive care unit

(NICU) in Finland,[13] and concluded that CRIB is superior to SNAPPE in its ability to predict in-hospital death. The second was on 476 VLBWI in eight units in the United States,[6] and found non-significant differences between the two scores, with SNAPPE being only slightly better. The third was a retrospective study on 280 VLBWI in two NICUs in Sweden, and found similar results for the two scores.[14]

Both CRIB and SNAPPE have recently been updated. In 2001 a revised and shortened version of SNAPPE, called SNAPPE-II, was published,[15] which uses only nine items collected over 12 hours from admission instead of the original 34 collected over 24 hours. In 2003, CRIB was also updated, using only data (five items) available up to one hour from admission.[16]

The aim of this study was to compare the ability of CRIB, CRIB-II, and SNAPPE-II to predict in-hospital mortality in a cohort of VLBWI admitted to 12 NICUs participating in a regional network in Italy in 1999–2001.

PATIENTS AND METHODS

The study was a prospective study planned to compare CRIB with SNAPPE-II; after publication of the CRIB-II score, relevant data were analysed and an assessment of CRIB-II was included. A cohort of infants (BW < 1500 g, GA 23–32 weeks), admitted to the 12 level III NICUs in Lombardy (Northern Italy) that participated in the Network Neonatale Lombardo in 1999–2001, were included in the analysis. Neonates of GA < 23 weeks or BW < 400 g, those with lethal congenital anomalies, those who died in the delivery room or

....

Abbreviations: AUC, area under the curve; BW, birth weight; CRIB, clinical risk index for babies; GA, gestational age; HL, Hosmer-Lemeshow; NICU, neonatal intensive care unit; SGA, small for gestational age; SNAPPE, score for neonatal acute physiology—perinatal extension; VLBWI, very low birthweight infants

were moribund on admission to the NICU (arbitrarily defined as dying within 10 hours of life and receiving only comfort care), and late admissions (> 12 hours from birth) were excluded.

The outcome measure was in-hospital death. Neonates transferred from participating NICUs were tracked to ascertain their outcome; those transferred after 20 days of life or transported back to level I or level II units were considered to be alive. The outcome of two transferred infants remained unknown, and they were excluded from the analysis. The final sample was 720 VLBWI.

Calculation of scores

Data used to calculate the scores were prospectively collected, abstracted from charts by one or two trained observers at each centre, and recorded on a customised web database (www.neonatalnet.org) as part of a regional neonatal network. CRIB[7] was calculated from six items: BW; GA; highest and lowest fractional inspired oxygen (F_{IO_2}) needed to keep a normal arterial oxygen saturation (88–95%) excluding the delivery room; worst base excess; congenital anomalies. The data collection window was the first 12 hours of life. SNAPPE-II[15] was calculated from nine items: BW; being small for GA (SGA); Apgar at five minutes; urine output; lowest mean blood pressure; worst Pao_2/F_{IO_2} ratio; lowest pH; occurrence of seizures; lowest temperature. The data collection window was the first 12 hours after admission to the NICU. SNAPPE-II data were abstracted following the original description of SNAPPE[8] (data collection was started in 1999, when only abstracts on SNAPPE-II had been published); weights were given to the values after SNAPPE-II publication. To minimise errors in data collection, original values were recorded, and coding and scores were calculated by computer.[17]

Following the original SNAPPE recommendations,[8] items not recorded in charts were treated in the score calculation as a normal value. For CRIB, the only item that could be missing was worst base excess: in this case, following the same line of reasoning and CRIB description,[7] a normal value was assumed.

CRIB-II[16] is calculated from five items: sex; BW; GA; worst base excess; temperature at admission. Given that CRIB-II was published after collection of the data, missing data were not assumed to be normal and the case was deleted. Thus 720 neonates were available for the three way comparison.

Statistical methods

Discrimination—that is, the ability of the scores to correctly predict life or death—was assessed by calculating receiver operating characteristic curves and their associated area under the curve (AUC).[18] An AUC value of 0.5 indicates no ability to discriminate, and larger values indicate increasing ability. A value of 0.8 is considered good.

As a model could have a high AUC, yet systematically overestimate or underestimate risk for some groups of infants, the Hosmer-Lemeshow (HL) test[19] was used to measure the goodness of fit of the models. This test compares observed and expected mortality across several strata (usually 10) of the score. A non-significant p value of the HL statistic indicates a model with a constant discriminatory ability across strata.

The effect of adding other variables to the scores was assessed with multiple logistic regression models, using inhospital death as the dependent variable, and calculating as above the AUC and the HL test. Because of the sampling method used (cluster sampling), methods that allow robust estimation of variance were used.

All calculations were carried out with the statistical package Stata 7 (College Station, Texas, USA).

RESULTS

Table 1 gives the basic characteristics of the neonates.

The median CRIB value was 1 (interquartile range (IQR) 1–4) for infants who survived and 10 (7–14) for those who did not. The corresponding values were 7 (5–9) and 14 (11–16) for CRIB-II and 20 (9–33) and 56.5 (39–71) for SNAPPE-II.

Figure 1 shows the receiver operating characteristic curves comparing the discriminating ability of CRIB, CRIB-II, and SNAPPE-II. CRIB and CRIB-II showed significantly greater discrimination than SNAPPE-II (AUC 0.90 (0.015) for CRIB, 0.91 (0.016) for CRIB-II, and 0.84 (0.020) for SNAPPE-II; p < 0.0004), but had a worse goodness of fit (HL p value = 0.045 for CRIB, 0.04 for CRIB-II, and 0.52 for SNAPPE-II).

Exclusion of babies weighing 400–499 g (n = 15) yielded an AUC of 0.898 for CRIB, 0.905 for CRIB-II, and 0.835 for SNAPPE-II (p = 0.0015).

All the scores did not fully estimate the risk of death, because in addition to the scores, several other variables were significantly associated with survival in multiple logistic regression models. Antenatal steroid prophylaxis, caesarean section, singleton birth, an Apgar score $\geqslant 7$ at five minutes, not being SGA, and not having any congenital anomaly, were significantly associated (or tended to significance) with a better survival in VLBWI (tables 2–4). The influence of these factors was greatest with SNAPPE-II, and least with CRIB-II.

A higher GA was associated with a better survival in CRIB and SNAPPE-II models, and with a worse survival in the CRIB-II model. For the other factors, the direction and magnitude of the odds ratios were similar. Only congenital anomalies, which are included in CRIB but not in SNAPPE-II and CRIB-II, had a clearly smaller odds ratio in CRIB. For CRIB and SNAPPE-II, it must be noted that some factors are highly significant even though they are already included in the score (GA and congenital anomalies for CRIB, SGA, and Apgar score at five minutes for SNAPPE-II), probably indicating poor weighting for these items. For both scores, the addition of these simple factors substantially increased both discrimination (CRIB-based model: AUC = 0.931; SNAPPE-II-based model: AUC = 0.913) and goodness of fit (CRIB: HL p value = 0.20; SNAPPE-II: p = 0.53). For comparison, a logistic model including only these variables and BW, but without CRIB, CRIB-II, or SNAPPE-II, had an AUC of 0.907 with a good fit (HL test p value = 0.49).

DISCUSSION

This study addressed two important questions: what are the results of one widely used (CRIB) and two new (CRIB-II and SNAPPE-II) scores in an unselected sample of VLBWI $\leqslant 32$ weeks in terms of discrimination, goodness of fit, and comparative performance; is risk adjustment using these

Table 1	Characteristics of the sample studied (median and range or percentage)
Birth weight (g)	1090 (400–1499)
Gestational age (weeks)	29 (23–32)
Birth weight <5th centile	11.2%
Location of birth: inborn	89.5%
Males	49.9%
Steroid prophylaxis	78.5%
Caesarean section	80.4%
Twin (any order)	28.8%
Use of mechanical ventilation	58.5%
Use of surfactant	48.4%
In hospital mortality	16.7%
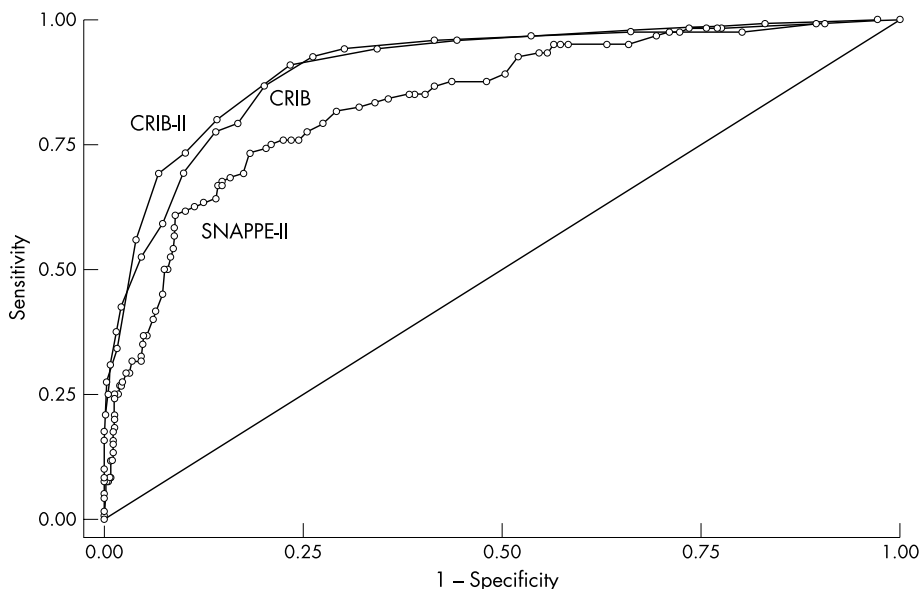
Figure 1 Receiver operating characteristics curves for the clinical risk index for babies (CRIB), CRIB-II, and score for neonatal acute physiology—perinatal extension (SNAPPE-II). The area under the curve: CRIB, 0.903; CRIB-II, 0.907; SNAPPE-II, 0.837.

scores complete, so that their use effectively removes the confounding due to severity of illness and allows unbiased comparisons between different groups of neonates.

This is the first study to compare CRIB, CRIB-II, and SNAPPE-II for their ability to correctly predict in-hospital death in VLBWI. We show that CRIB and CRIB-II performed better than SNAPPE-II; however, the addition of other perinatal factors considerably improved discrimination of all scores.

The worse SNAPPE-II performance in VLBWI cannot be ascribed to its unexpectedly poor performance, because the discriminatory ability obtained in this study for all scores was in line with published data.[6][7][13–16] In fact, in our study the scores in VLBWI were as in the original papers.[7][15][16] The cohort in which our study was carried out (infants of BW 400–1499 g and GA $\leqslant$ 32 weeks) was slightly different from the original CRIB cohort[7] (infants of GA < 31 weeks or BW < 1500 g) and CRIB-II cohort[16] (infants of GA $\leqslant$ 32 weeks). On the other hand, SNAPPE[8] and SNAPPE-II[15] were developed and validated on all acute neonatal admissions, irrespective of GA or BW. Although these differences should not influence the risk adjusted mortality prediction (our inclusion criteria defined a subset of those with the three scores), some SNAPPE-II items may have a lower sensitivity in VLBWI. Richardson et al[15] have acknowledged that urine output is less useful in VLBWI than in bigger neonates,

probably because of difficulties in obtaining complete and precise spontaneous urine collection in a relatively short timespan (12 hours from admission) in tiny babies. A recent report confirms that urine output is an unreliable item.[20] A similar problem is found with seizures: in our sample, only five babies had convulsions in the first 12 hours. Also this item is probably more useful in larger babies, where admission to NICU reflects a larger variety of causes.

For CRIB and SNAPPE-II, a possible cause of reduced discrimination is the weight given to levels of physiological derangement for some items: the fact that low Apgar score and being SGA (for SNAPPE-II) and congenital anomalies and GA (for CRIB) are highly significant and increase discrimination when added to the models implies that these factors, already present, are inadequately weighted by the scores.

For SNAPPE-II, the problem could again be because the weights were derived from samples in which VLBWI represented a minority, and for CRIB because the weights were derived a decade ago, when a different mix of risk factors (with different relative importance) probably influenced mortality.

Moreover, in CRIB some items are determined by the care team. For instance, CRIB takes into account both highest and lowest appropriate F_{IO_2} used during the first 12 hours, and it is possible that prophylactic surfactant administration could

Table 2 Odds ratios for CRIB based logistic regression model of mortality

	Odds ratio (95% CI)	p Value
Gestational age (per week)	0.70 (0.61 to 0.80)	< 0.001
Steroid prophylaxis	0.47 (0.22 to 0.99)	0.049
Vaginal delivery	2.13 (0.79 to 5.73)	0.135
Multiple birth	2.12 (1.27 to 3.54)	0.004
5 min Apgar < 7	1.64 (1.01 to 2.67)	0.045
Being SGA	2.44 (0.98 to 6.08)	0.056
Congenital anomalies	2.50 (1.42 to 4.42)	0.002

Odds ratio for CRIB (per point change): 1.352 (1.21 to 1.51), p < 0.001.
CRIB, Clinical risk index for babies; SGA, small for gestational age.

Table 3 Odds ratios for SNAPPE-II based logistic regression model of mortality

	Odds ratio (95% CI)	p Value
Gestational age (per week)	0.62 (0.55 to 0.71)	< 0.001
Steroid prophylaxis	0.39 (0.22 to 0.71)	0.002
Vaginal delivery	2.46 (1.24 to 4.90)	0.010
Multiple birth	2.02 (1.28 to 3.17)	0.002
5 min Apgar < 7	1.72 (1.01 to 2.91)	0.045
Being SGA	3.20 (1.54 to 6.66)	0.002
Congenital anomalies	4.92 (1.55 to 15.55)	0.007

Odds ratio for SNAPPE-II (per point change): 1.036 (1.02 to 1.05), p < 0.001.
SNAPPE, score for neonatal acute physiology—perinatal extension; SGA, small for gestational age.

Table 4 Odds ratios for CRIB-II based logistic regression model of mortality

	Odds ratio (95% CI)	p Value
Gestational age (per week)	1.20 (1.00 to 1.45)	0.052
Steroid prophylaxis	0.42 (0.22 to 0.80)	0.009
Vaginal delivery	2.23 (0.96 to 5.17)	0.062
Multiple birth	2.16 (1.26 to 3.71)	0.005
5 min Apgar < 7	1.54 (0.95 to 2.48)	0.077
Being SGA	1.94 (0.86 to 4.36)	0.108
Congenital anomalies	5.01 (1.72 to 14.61)	0.003

Odds ratio for CRIB-II (per point change): 1.853 (1.64 to 2.09), p < 0.001.
CRIB, Clinical risk index for babies; SGA, small for gestational age.

obliterate this difference, and thus reduce CRIB performance. In this sample, 48.4% of the newborns received surfactant at any time, and only 12.5% of these in the delivery room. We cannot therefore comment on the performance of CRIB when a prophylactic surfactant strategy is used. However, the problem of using variables that can be arbitrarily determined by doctors, such as FIO_2,[21] was one of the main reasons for developing CRIB-II, which only uses information available within one hour of admission and independent of the care provided. Our data show that the new score has a discrimination that is comparable to that of the "old" CRIB, but potentially less confounded by treatment given by the team.

A surprising finding with CRIB-II is that GA is almost significant (table 4), with death risk increasing with increasing GA. This suggests that the BW–GA–sex matrix used in CRIB-II is not well calibrated for our sample. This interpretation is strengthened by the observation that risk adjustment using all scores is imperfect, and even after accounting for "severity of illness", other important prenatal and perinatal factors influence VLBWI survival, including antenatal steroid prophylaxis, Apgar score at five minutes, type of delivery, multiple birth, and congenital anomalies. These factors are known to influence survival in VLBWI, and a logistic model including only these variables and BW, but without CRIB, CRIB-II, or SNAPPE-II, showed excellent discrimination. CRIB-II and SNAPPE-II do not take into account congenital anomalies, but the large estimated effect on mortality (an odds ratio of 5.01 for CRIB-II and 4.92 for SNAPPE-II) when added to the model suggests that this information should be included in a severity score.[22]

The imperfect adjustment has been reported before.[6 23 24] For antenatal steroid prophylaxis, Richardson et al[15] argue against including it in the score, because this would excuse poor obstetrical management, and "it may obscure the ill effects of improper treatment". This line of reasoning cannot, however, be applied to other factors which do not depend on medical decisions, such as multiple birth, congenital anomalies, or being SGA. The important contribution to risk of these factors means that the use of CRIB, SNAPPE-II, or, to a lesser degree, CRIB-II alone cannot guarantee an unbiased risk adjustment, thus partly losing the reason for using these scores.

In conclusion, this study shows that CRIB and CRIB-II were superior to SNAPPE-II in VLBWI, but their goodness of fit was not good. CRIB-II appeared to be as accurate as CRIB in this sample, and was less confounded by other factors. All scores do not fully estimate the risk of death—that is, in addition to CRIB, CRIB-II, or SNAPPE-II, other perinatal factors influence survival and should be included in any analysis to improve risk adjustment.

. . . .
Authors' affiliations
L Gagliardi, A Cavazza, A Brunelli, M Battaglioli, D Merazzi, F Tandoi, D Cella, G F Perotti, G Pelti, I Stucchi, F Frisone, A Avanzini, R Bellù, Neonatal Intensive Care Units of the following hospitals: Mangiagalli (Milan), Spedali Civili (Brescia), Niguarda (Milan), V Buzzi (Milan), S Anna (Como), Varese, S Raffaele (Milan), S Matteo (Pavia), Salvini (Rho), Valduce (Como), Fornaroli (Magenta), A Manzoni (Lecco)

REFERENCES

1 **Tarnow-Mordi W**, Ogston S, Wilkinson AR, et al. Predicting death from initial disease severity in very low birthweight infants: a method for comparing the performance of neonatal units. BMJ 1990;**300**:1611–14.
2 **Richardson DK**, Tarnow-Mordi WO, Escobar GJ. Neonatal risk scoring systems. Can they predict mortality and morbidity? Clin Perinatol 1998;**25**:591–611.
3 **Richardson DK**, Tarnow-Mordi WO, Lee SK. Risk adjustment for quality improvement. Pediatrics 1999;**103**:255–65.
4 **Perlman M**. Neonatal Intensive Care Units report cards. Pediatrics 1998;**102**:975–6.
5 **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.
6 **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.
7 **The 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.
8 **Richardson DK**, Phibbs CS, Gray JE, et al. Birth weight and illness severity: independent predictors of neonatal mortality. Pediatrics 1993;**91**:969–75.
9 **Escobar GJ**, Fisher A, De Kun Li, et al. Score for Neonatal Acute Physiology: validation in three Kaiser Permanente neonatal intensive care units. Peditrics 1995;**96**:918–22.
10 **Baumer JH**, Wright D, Mill T. Illness severity measured by CRIB score: a product of changes in perinatal care? Arch Dis Child Fetal Neonatal Ed 1997;**77**:F211–15.
11 **Petridou E**, Richardson DK, Dessypris N, et al. Outcome prediction in Greek neonatal intensive care units using a Score for Neonatal Acute Physiology (SNAP). Pediatrics 1998;**101**:1037–44.
12 **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.
13 **Rautonen J**, Makela A, Boyd H, et al. CRIB and SNAP: assessing the risk of death for preterm neonates. Lancet 1994;**343**:1272–3.
14 **Eriksson M**, Bodin L, Finnstroem O, et al. Can severity-of-illness indices for neonatal intensive care predict outcome at 4 years of age? Acta Paediatr 2002;**91**:1093–100.
15 **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.
16 **Parry G**, Tucker J, Tarnow-Mordi WO. CRIB-II: an update of the clinical risk index for babies score. Lancet 2003;**361**:1789–91.
17 **Marlow N**. Illness severity measures in neonatal intensive care. Acta Paediatr 2002;**91**:367–8.
18 **Hanley JA**, McNeil BJ. The meaning and the use of the area under a receiver operating characteristics (ROC) curve. Radiology 1982;**143**:29–36.
19 **Hosmer DW**, Lemeshow S, eds. Applied logistic regression. New York: J Wiley and Sons, 1989:135–175.
20 **Sutton L**, Bajuk B, Berry G, et al. Reliability of SNAP (score for neonatal acute physiology) data collection in mechanically ventilated term babies in New South Wales, Australia. Acta Paediatr 2002;**91**:424–9.
21 **Gagliardi L**, Barbarini M, Pugni L, et al. Effect of changes in inspired oxygen tension on indexes of oxygenation in ventilated neonates. Pediatr Crit Care Med 2002;**3**:34–8.
22 **Thompson MP**, Richardson DK, Stark AR, et al. Mortality risk due to congenital anomalies in VLBW infants. A simple scoring system. Pediatr Res 1997;**41**:211.
23 **Di Lallo D**, Miceli M, Orzalesi M, et al. Neonatal illness severity scores: the search of a simple solution. Helsinki: Proceedings of the European Society for Pediatric Research Annual Meeting, 2001:96.
24 **International Neonatal Network, Scottish Neonatal Consultants, Nurses Collaborative Study Group**. 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.