Abstract

Aim—To predict the individual neonatal mortality risk of preterm infants using an artificial neural network “trained” on admission data.

Methods—A total of 890 preterm neonates (<32 weeks gestational age and/or <1500 g birthweight) were enrolled in our retrospective study. The neural network trained on infants born between 1990 and 1993. The predictive value was tested on infants born in the successive three years.

Results—The artificial neural network performed significantly better than a logistic regression model (area under the receiver operator curve 0.95 vs 0.92). Survival was associated with high morbidity if the predicted mortality risk was greater than 0.50. There were no preterm infants with a predicted mortality risk of greater than 0.80. The mortality risks of two non-survivors with birthweights >2000 g and severe congenital disease had largely been underestimated.

Conclusion—An artificial neural network trained on admission data can accurately predict the mortality risk for most preterm infants. However, the significant number of prediction failures renders it unsuitable for individual treatment decisions.

Keywords: artificial neural network; mortality; prediction

Mortality has traditionally been used to compare the therapeutic performances of neonatal intensive care units. As individual mortality risk depends on numerous confounding factors, crude mortality can be adjusted for prematurity, birthweight, initial illness severity, presence of congenital malformations, or a combination of these factors, for the purposes of quality assessment.

The Score for Neonatal Acute Physiology (SNAP), Perinatal Extension (SNAP-PE), Clinical Risk Index for Babes (CRIB) score and alternative models are excellent predictors of overall inpatient mortality of preterm neonates if accuracy is judged by the area under the receiver operating characteristic (ROC) curve. Despite such excellent results a significant number of individual prediction failures still occur. Case analysis of non-survivors with a low predicted individual mortality risk, or of survivors with a very high predicted individual mortality risk might yield further insight into the pitfalls of scoring systems.

As the quality of early therapeutic interventions might significantly influence the degree of illness severity, data on initial illness severity are compromised if collected later than a few hours after birth. Thus several variables contributing to CRIB or SNAP scores are regarded as input as well as output data with respect to intensive care unit performance. Shortening the data collection period is desirable. Artificial neural networks (ANN) are software tools with the capacity to learn. An ANN behaves like a child learning to differentiate between cats and dogs, by means of examples, under the supervision of his/her parents. An ANN can learn the relation between input variables (size, fleece, voice, behaviour patterns) and the output variable (cat or dog) by presenting multiple input–output pairs to it (supervised learning). After this learning or “training” period the ANN can “predict” the output (cat or dog) on inputs of further unknown examples. This capability is called generalisation. ANNs have shown excellent predictive accuracy in medicine even on inaccurate or incomplete input data.
reviews describe the many clinical applications of ANN. 19–20

Using an ANN for admission data on preterm neonates we aimed to: predict the individual neonatal mortality risk of preterm neonates from their admission data; compare the net’s performance with that of a logistic regression model; and characterise survivors despite a high predicted individual mortality risk, and non-survivors despite a low predicted individual mortality risk.

Methods

Our investigation was based on a pre-existing data set of 890 preterm neonates (gestational age < 32 completed weeks gestation and/or birthweight < 1500 g) born between 4 April 1990 to 24 October 1996. Admission and outcome data were available on computer files (tables 1 and 2). Outcome was defined with respect to neonatal death (death within the first 28 days of life).

The patient sample was split into a training and test (n = 435) and a validation set (n = 455) according to date of birth. The cutoff date was arbitrarily set to 5 November 1993. The training and test set data were subjected to a forward stepwise logistic regression analysis (cutoff p value to include a variable Pin = 0.05, cutoff p value to exclude a variable Pout = 0.1) (SPSS/PC release 6.1.2, SPSS Inc., Chicago, IL, USA), and were used to train the ANN (Predict, NeuralWare Inc., Pittsburgh, USA). Predict helped to develop a feed–forward fully connected three layer perceptron. The descent gradient learning rule was applied for error propagation. 21 22 The ANN input variable selection mode implemented in Predict is based on genetic algorithms and helps to find the optimal combination of inputs. 21 The training process was stopped as soon as the prediction error on a subset of the training data set, namely the test set, stopped declining further. After the training the ANN had settled to 13 input and three hidden processing elements. There was one output processing element.

For validation the inputs were derived from the independent validation data set (6 November 1993 to 24 October 1996). The logistic regression model and the ANN delivered the individual mortality risk of each single preterm neonate. Prediction and actual outcome were compared to assess performance.

For the purpose of performance assessment and method comparison ROC (receiver operating characteristic) curves were used. The outcome predictions as delivered by both models were continuous variables (individual mortality risk). The cutoff point used to transform them into dichotomous mortality predictions was varied to obtain ROC curves. 23 24 The maximum likelihood estimation of the binormal ROC curves, and the area under the fitted ROC curves were calculated using LABROC1 and CLABROC software (IBM-PC version 1.2.1; Metz CE et al. 1993, Department of Radiology, Chicago Medical Center, Chicago, IL, USA). CLABROC was used to test corresponding ROC curves for any significant difference (univariate Z-score test of the difference between areas under the two paired ROC curves, and univariate Z-score test of the difference between sensitivity values on the two paired ROC curves at a selected specificity level 25).

Surviving patients with an ANN based individual mortality risk of more than 0.50 were matched with infants identical in gestational age and birthweight (within ±5%), but with a predicted individual mortality risk of less than 0.50. Morbidity of the two groups was compared with respect to the variables listed in table 3. Non-survivors with very low predicted individual mortality risk (< 0.10) were analysed individually.

To test for associations between the variables in the training or validation set univariate tests were performed (χ2 test for categorical variables; t test, or Mann-Whitney U test for continuous variables). A p value of < 0.05 was considered significant.

Results

The training data set overall neonatal mortality was 8.7%, while in the validation set it was 7.9%. The difference is not significant (table 1). The validation dataset differed in several respects from the training set (table 1 and 2).
Table 6 Comparison of prediction accuracy of ANN and logistic regression models

<table>
<thead>
<tr>
<th>Specificity (%)</th>
<th>Sensitivity (95% confidence interval) (%)</th>
<th>ANN</th>
<th>Logistic regression model</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td>98.8 (93.1 to 99.9)</td>
<td></td>
<td>94.7 (85.1 to 98.6)</td>
<td>0.061</td>
</tr>
<tr>
<td>80</td>
<td>97.5 (89.8 to 99.6)</td>
<td></td>
<td>90.6 (78.8 to 96.7)</td>
<td>0.029</td>
</tr>
<tr>
<td>85</td>
<td>94.6 (94.2 to 98.7)</td>
<td>81.5 (69.5 to 92.5)</td>
<td>0.010</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>87.8 (74.2 to 95.4)</td>
<td>70.6 (54.4 to 83.5)</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>95</td>
<td>69.6 (51.6 to 83.7)</td>
<td>46.1 (28.9 to 64.0)</td>
<td>0.001</td>
<td></td>
</tr>
</tbody>
</table>

ROC curves reflecting each ANN’s performances are given in table 5.

When adjusted for a specificity level of 80, 85, 90, or 95%, in terms of sensitivity, the predictions of the ANN were significantly better than the logistic regression model (table 6). The ANN area under the ROC curve was 0.954 (SD 0.011), which is larger than the 0.917 (SD 0.017) of the logistic regression model area (p = 0.002, area test) (fig 1).

No preterm neonate had an ANN predicted mortality risk of >0.80. Thirteen (68.4%) of the 19 preterm neonates with an individual predicted mortality risk of >0.70 died in the neonatal period. Of the 67 preterm babies with a predicted mortality risk of >0.5, 36 (53.7%) died before 29 days of life and three thereafter (fig 2).

When morbidity was graded according to grossly abnormal brain ultrasound scans, or to the number of postnatally acquired diseases requiring one or more surgical interventions, survivors with an ANN based mortality risk >0.50 had a significant higher morbidity than matched survivors with an ANN based mortality risk < 0.50 (p=0.02) (table 3).

The clinical characteristics, cause of death, and the individual ANN derived mortality risk of the three low risk non-survivors (predicted mortality risk < 0.10) are summarised in table 7.

When morbidity was graded according to grossly abnormal brain ultrasound scans, or to the number of postnatally acquired diseases requiring one or more surgical interventions, survivors with an ANN predicted mortality risk >0.5, 36 (53.7%) died before 29 days of life and three thereafter (fig 2).

Case 1, a preterm infant of 27 weeks gestation, in his first hours of life, developed respiratory distress syndrome that was unresponsive to surfactant rescue treatment. Cultures on blood and tracheal fluid grew Escherichia coli. The infant died on his third day of life, presenting with septic shock syndrome.

Case 2 had a positive history with respect to neonatal death in three maternal siblings. Polyhydramnios had been present in each case, and all the mother’s siblings had died of respiratory insufficiency. The mother turned out to have myotonic dystrophia. Her infant was in need of aggressive stabilisation immediately after birth, including adrenalin and sodium bicarbonate. After stabilisation the infant was transported to neonatal intensive care. On admission her capillary pH was within an acceptable range (table 7).

Case 3 had complex cardiac malformation (double outlet right ventricle, hypoplastic pulmonary artery, pulmonary valve stenosis).
and respiratory distress syndrome grade IV. Blood pressure and blood pH could be stabilised for only a short period.

In the latter two cases the individual predicted mortality risks increased when predictions were obtained from an ANN trained without information on gestational age and birthweight (table 7).

Discussion

Artificial neural networks have been widely used for outcome prediction. Mortality has been predicted with great accuracy for patients on intensive care units, and for patients who received cardiopulmonary resuscitation in hospital. (Doig GS, et al. Proceedings of annual symposium on computers applied to medical care, 1993: 361-65).

Solely based on routine data for the first minutes of life, our ANN was capable of accurately predicting individual neonatal mortality risk in preterm neonates despite changes in patient characteristics between the training and validation time period. The ANN performed better than a logistic regression model.

While the logistic regression model comprised six items, the ANN used 13 items (table 5). The input variable selection of the ANN seems reasonable as most selected variables were reportedly associated with mortality and as all predictive variables isolated by the stepwise logistic regression analysis were also included in the ANN input variable set. Hence subtle and/or non-linear associations between input variables and neonatal mortality not detected by logistic regression analysis might account for the superior performance of the ANN.

Gestational age, birthweight, and condition on admission are the only items that must not be omitted during the training process without severely compromising performance (table 5). The special impact of the first two of those items on neonatal mortality risk prediction is illustrated by the fact that the area under the ROC curve of the ANN trained with just those two items (table 5) is larger than comparable reports suggest. We cannot exclude that our neonatologists on duty underestimated the gestational age of some of the neonates, including more mature and "healthier" infants (who are at almost no risk for neonatal mortality and where mortality risk is easier to predict) into our study population. Any comparable studies focused on prediction of inpatient mortality (table 8). Thus neonatal mortality may be easier to predict than inpatient mortality.

The method to find the optimal ANN input variable set still awaits standardisation despite recent advances in the field. It seems reasonable to use pruning or genetic algorithms for that purpose. We relied on the intrinsic variable selection mode of Predict, but—while feature inherent in genetic algorithms—we cannot be absolutely sure that we have found the variable set with the maximum predictive power.

There is no easy way to assess the relative impact of the individual ANN input variables. We favoured a pragmatic approach by training several ANNs, subsequently omitting one input of the optimal input variable set, and comparing the different performances.

Twenty five survivors with a high (> 0.50) ANN derived individual mortality risk underwent a matched pairs analysis with counterparts of lower risk (< 0.50) (table 3). All 50 preterm neonates were very immature (median gestational age 26.3 weeks, median birthweight about 880 g), and most of them survived but at the expense of major sequelae. The ANN identified a subgroup of preterm neonates with significantly higher morbidity when they survived. Individuals in this subgroup presenting with a predicted individual neonatal mortality risk of > 0.50 had a higher rate of cerebral com-

Table 7 Case analysis of non-survivors with ANN predicted individual mortality risk < 0.10

<table>
<thead>
<tr>
<th>Variable</th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apgar 1/5/10 minutes</td>
<td>09/10/10</td>
<td>04/05/05</td>
<td>04/07/08</td>
</tr>
<tr>
<td>Birthweight (g)</td>
<td>1085</td>
<td>2005</td>
<td>2025</td>
</tr>
<tr>
<td>Capillary blood pH on admission</td>
<td>7.35</td>
<td>7.32</td>
<td>7.17</td>
</tr>
<tr>
<td>Condition on admission</td>
<td>Stable</td>
<td>Vitality endangered</td>
<td>Vitality endangered</td>
</tr>
<tr>
<td>Visible congenital malformation on admission</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Emergency delivery</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Gestational age (completed weeks)</td>
<td>27</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td>Inborn</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Body temperature on admission (°C)</td>
<td>36.5</td>
<td>36.2</td>
<td>36.2</td>
</tr>
<tr>
<td>Cause of death</td>
<td>E coll sepsis</td>
<td>Congenital myotonic dystrophy, lung hypoplasia, RDS IV° and pneumonia on postmortem examination</td>
<td>Complex congenital heart disease</td>
</tr>
<tr>
<td>ANN derived mortality risk</td>
<td>0.022</td>
<td>0.044</td>
<td>0.098</td>
</tr>
<tr>
<td>ANN-2i derived mortality risk</td>
<td>0.006</td>
<td>0.378</td>
<td>0.624</td>
</tr>
</tbody>
</table>

*ANN trained omitting the items gestational age and birth weight
Table 8  Comparison of different mortality prediction models

<table>
<thead>
<tr>
<th>Patient group</th>
<th>Institutions</th>
<th>Patient group size - validation set included</th>
<th>Study period</th>
<th>Data collection period</th>
<th>End point</th>
<th>Sensitivity at 0.74 specificity</th>
<th>Area under ROC curve for validation sample</th>
<th>Name of score (reference)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inborn 501-1500 g</td>
<td>7</td>
<td>3603</td>
<td>1987-9</td>
<td>Pre- and perinatal</td>
<td>Death</td>
<td>0.77 0.82</td>
<td>Horbar 1993</td>
<td>Tarnow 1990</td>
</tr>
<tr>
<td>&lt; 1500 g and ventilated; patients who died in the first 12 h or had severe congenital malformations were excluded</td>
<td>2</td>
<td>262</td>
<td>1983-9</td>
<td>Pre-, peri-, and neonatal (12 h)</td>
<td>Death &gt; 12 h</td>
<td>0.81</td>
<td>This article</td>
<td></td>
</tr>
<tr>
<td>Newborns; late readmitted and moribund patients excluded</td>
<td>3</td>
<td>1621</td>
<td>1989-90</td>
<td>Neonatal (24 h)</td>
<td>Death</td>
<td>0.88 0.90 (0.82†)</td>
<td>SNAP (Richardson 1993)</td>
<td>SNAP-PE (Richards 1993)</td>
</tr>
<tr>
<td>Newborns; late readmitted and moribund patients excluded</td>
<td>3</td>
<td>1621</td>
<td>1989-90</td>
<td>Peri- and neonatal (24 h)</td>
<td>Death</td>
<td>0.91 0.93 (0.79†)</td>
<td>CRIB (Richards 1993)</td>
<td>CRIB (INN 1993)</td>
</tr>
<tr>
<td>&lt; 1500 g or &lt; 31 weeks gestation; lethal congenital malformations excluded</td>
<td>4</td>
<td>1201</td>
<td>1988-90</td>
<td>Pre-, peri-, and neonatal (12 h)</td>
<td>Death</td>
<td>0.87 0.90</td>
<td>This article</td>
<td></td>
</tr>
<tr>
<td>&lt; 1500 g or &lt; 32 weeks gestation</td>
<td>1</td>
<td>890</td>
<td>1990-6</td>
<td>Pre- and perinatal</td>
<td>Neonatal death</td>
<td>0.99 0.95</td>
<td>This article</td>
<td></td>
</tr>
<tr>
<td>Pre- and perinatal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-, peri-, and neonatal (12 h)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-, peri-, and neonatal (24 h)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-, peri-, and neonatal (12 h)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neonatal (24 h)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neonatal (24 h)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*According to Rantonen et al (1994)†† Prenatal: input data collection from conception until labour
Perinatal: input data collection from labour until admission (admission data included)
Neonatal: input data collection after admission (time span as indicated)

Applications, and developed more conditions that eventually required surgical intervention. Those findings agree with those of recent studies on illness severity scores (SNAP, CRIB) where a positive association between high scoring and the presence of a major cerebral abnormality on ultrasound brain scan was found.1

Although overall prediction of neonatal mortality is accurate enough for the purpose of intrahospital quality management and risk stratification, for the purpose of individual non-treatment decisions, scoring systems as well as the ANN still fail to differentiate between high risk preterm neonates who will eventually die and those who will survive for at least 28 days. Important factors contributing to preterm mortality risk were probably not included in the developed of established scoring systems and in the training data of our ANN. And perhaps those contributing factors (for example, antioxidant status, etc) have not been measured yet.

One probable reason for the prediction failures in cases 2 and 3 of low risk non-survivors is the tendency of the ANN to underestimate the mortality risk the heavier and more mature the babies are. Our database lacked important information on physiological measures such as cord blood pH value, oxygen requirement in the first hours of life, or the presence of signs of infection. Hence the ANN overestimated the impact of immaturity on mortality. In case 1 sepsis started during day 1 of life and the baby died on day 3. The data collecting period required to detect physiological derangements would have been 12 to 24 hours—the same as in the CRIB or SNAP scores. Our admission data did not adequately reflect individual mortality risk in this case.

Several models that can predict mortality in preterm neonates have been developed.1 2 3 17 18 A performance comparison of these models proved difficult as the studies differ in respect to authors’ intention, study population, number of institutions enrolled, data collecting period, items used to characterise the neonate, primary endpoint, and general applicability (table 8).

Although high impact items were missing in our database, and data collecting was restricted to the prenatal and immediate perinatal periods, the ANN’s performance is equivalent to established scoring systems. In contrast to most other mortality prediction models, we have opted for neonatal mortality instead of inpatient mortality as the endpoint. The reason for this is that individual diseases contributing substantially to mortality after 28 days (late necrotising enterocolitis, central catheter sepsis, respiratory syncytial virus infection, etc.) are unlikely to be associated with admission data. It should be mentioned that in contrast to published reports,2 13 surfactant treatment and high frequency oscillation, both of which can influence mortality,16 17 had already been introduced into routine care during the study period.

We relied on pre-existing admission data. We feel that information on important topics was missing in our database—for example, illness severity in the first hours of life. One of the items used was a very subjective measure of illness severity—the assessment of the overall condition on admission, as performed by the physician on duty. Stevens et al31 and Richardson et al32 reported that the physician’s and nurse’s estimation of neonatal mortality risk was highly correlated with the actual mortality risk, or with the SNAP score. The assessment of the overall condition on admission had not been intended to serve as an estimate of mortality risk at the time our data were recorded. Notably, an ANN trained with those subjective assessments made by 30 different physicians delivered reliable results when fed with independent data of the subsequent validation time period, when overall condition on admission was assessed by another 30 physicians. The use of this kind of subjective input variable is a major limitation of the present study. Efforts are under way to substitute those variables with physiological measures, and to broaden the database. The aim is to train an ANN exclusively on objectively measurable physiological data, to render it broadly applicable to other neonatal intensive care units.

A feature intrinsic to an ANN is its retraining potential in case of advances in medical care, or simply in case of additional data becoming available. In contrast to multiple regression analysis, expanding the database allows the ANN to improve its knowledge using non-
linear relations between input and output data, whereas multiple regression at some point reaches its intrinsic limitations.22 With multiple regression, the limit of accuracy is determined by two factors— namely, “noise” (missing, inaccurate, or false data), and the degree of non-linearity between the input and outcome variables. With the ANN approach it is noise alone that determines the limit of accuracy. As Richardson and Tarnow-Mordi have pointed out,4,44 as sophisticated systems for automatic acquisition of large quantities of routine clinical information become available, emphasis may shift in favour of more complex clinical scoring systems to maximise predictive accuracy; the ability of an ANN to use a very large number of variables from different sources (parameters, settings, and monitoring data from mechanical ventilators, ECG signals or clinical chemistry data) is a distinct advantage. Neurol networks are able to include both quantitative and qualitative data into the same model. There are no limitations with respect to ordinal scaled data as there are in logistic regression analysis.

With the recent advances in hardware and software any clinician is able to develop ANNs to depersonalise his/her experience, and to make it accessible to junior colleagues. This can even happen under circumstances where rules are difficult to formulate and situations are too complex to be analysed by classic statistics.18

Trained ANNs predicting morbidity and mortality might help to assure high quality of care through comparison of predicted and actual outcome; to assess the time course of intrahospital advances in treatment; and to warn physicians and nurses, when individual infants are at high risk and deserve intensified attention. But prediction failures tell us that neonatal intensive care medicine is too complex to devise an individual no-treatment policy on a prediction that might turn out to be very inaccurate for individual patients.

The help of Miss Ulla Ollesch and the staff of the neonatology department at the Vestische Kinderklinik for data entry is gratefully acknowledged. The expert linguistic advice of Heidi Walsh and Neil McIntosh is also very much appreciated. Special thanks to Charles E Metz who kindly provided the software for the ROC analyses. Part of the data in this study are an integral part of Antje Westermann’s thesis.

22 Jacoba RA. Increased rates of convergence through learning rate adaption. Neural Networks 1981;1:299-308.
24 Hanlon JA, McNel BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. Radiology 1982;143:29-36.