

ARTICLES

Predicting Extubation Outcome in Preterm Newborns: A Comparison of Neural Networks with Clinical Expertise and Statistical Modeling

MARTINA MUELLER, CAROL L. WAGNER, DAVID J. ANNIBALE, THOMAS C. HULSEY, REBECCA G. KNAPP, AND JONAS S. ALMEIDA

Department of Biometry & Epidemiology [M.M., R.G.K., J.S.A.] and Department of Pediatrics [C.L.W., D.J.A., T.C.H.], Medical University of South Carolina, Charleston, SC.

ABSTRACT

Even though ventilator technology and monitoring of premature infants has improved immensely over the past decades, there are still no standards for weaning and determining optimal extubation time for those infants. Approximately 30% of intubated preterm infants will fail attempted extubation, requiring reintubation and resuming of mechanical ventilation. A machine-learning approach using artificial neural networks (ANNs) to aid in extubation decision making is hereby proposed. Using expert opinion, 51 variables were identified as being relevant for the decision of whether to extubate an infant who is on mechanical ventilation. The data on 183 premature infants, born between 1999 and 2002, were collected by review of medical charts. The ANN extubation model was compared with alternative statistical modeling using multivariate logistic regression and also with the clinician's own predictive insight using sensitivity analysis and receiver operating characteristic curves. The optimal ANN model used 13 parameters and achieved an area under the receiver operating characteristic curve of 0.87 (out-of-sample validation), comparing favorably with multivariate logistic regression. It also compared well with the clinician's expertise, which raises the possibility of being useful as an automated alert tool. Because an

ANN learns directly from previous data obtained in the institution where it is to be used, this makes it particularly amenable for application to evidence-based medicine. Given the variety of practices and equipment being used in different hospitals, this may be particularly relevant in the context of caring for preterm newborns who are on mechanical ventilation. (*Pediatr Res* 56: 11–18, 2004)

Abbreviations

ANN, artificial neural network
AUC, area under the ROC curve
CE, clinical expert
I:E ratio, ratio of inspiratory to expiratory time
INSP, inspiratory time
MAP, mean airway pressure
MLR, multivariate logistic regression
PEEP, positive end-expiratory pressure
ROC, receiver operating characteristic
Sao₂, saturation of arterial oxygen
V_T, tidal volume

Despite many technological advances over the past decade, predicting the ideal time point for extubation in premature infants who are on mechanical ventilation requires excellent diagnostic skills and remains a difficult task. A recently published study of the accuracy of clinical assessment of mortality risk in the neonatal intensive care unit found that the clinical predictions, in general, were unreliable and became increasingly unreliable with

increasing number of days of life (1). Determining the optimal time point for extubation is crucial to minimize infants' time on artificial ventilation, thus minimizing their risk of developing barotrauma (caused by high pressures in the lung), retinopathy (caused by high arterial oxygen), and subsequent bronchopulmonary dysplasia or chronic lung disease (2). Complicating the decision to extubate, however, is the smaller risk associated with having to re-intubate—subjecting the infants to subsequent increases of ventilatory support as a result of alveolar collapse or atelectasis. These risks could be reduced by use of an automated prediction system that could alert the neonatal intensive care unit staff to potential extubations.

Over the past decade, data mining by machine-learning tools to aid in clinical decision making has come of age. In partic-

Received August 13, 2003; accepted January 27, 2004.

Correspondence: Martina Mueller, Ph.D., Medical University of South Carolina, Department of Biometry & Epidemiology, 135 Cannon Street, Suite 301, Charleston, SC 29425, U.S.A.; e-mail: muellem@muscc.edu

Presented in part at the GNPI (Society of Neonatology and Pediatrics Intensive Medicine) meeting, Cologne, Germany, July 2003.

DOI: 10.1203/01.PDR.0000129658.55746.3C

ular, a significant number of clinical applications using neural networks have been developed (3–7). This growing interest can be explained by the ability of neural networks to learn directly from data, thus accommodating the intrinsic nonlinear nature of biomedical data to the detriment of rule-based systems. It is interesting that these developments unfolded in parallel to a growing emphasis in evidence-based medicine (8). Despite the expected advantage of artificial neural networks (ANNs) over more conventional statistical modeling, namely multivariate logistic regression (MLR), for classification problems, their development is seldom accompanied by comparative assessment of the clinicians' own decision making. As such, the three classification methods, ANN, MLR, and that by clinical practitioners, were evaluated extensively in the study reported here.

ANNs are a machine-learning technique modeled after natural nervous systems. In accordance with the structure of the nervous system, the units in an ANN discriminant function are called neurons, or nodes, and are arranged in layers. Nodes are fully interconnected between layers but, typically, not within them. All connections between nodes carry weights. The process of repeatedly adjusting these weights is considered learning, which is stopped when the resulting error function is minimized (9). Cross-validation for early stopping of the ANN regression combined with bootstrapping of the number of hidden nodes enables ANN predictors to identify signal from noisy and unbalanced sampled data (10,11). For a recent review of ANN for the analysis of complex biomedical data, see (12).

MLR is a well-established statistical modeling tool (13) that uses multiparametric sigmoidal discriminant functions. Because ANN identification is analogous to two consecutive logistic regressions, the multilayered ANN topology in this study uses sigmoid transfer functions. That ANNs generalize MLR to accommodate nonlinear dependencies makes the expectation of better performance of ANNs over MLR mostly an issue of proper implementation. Nevertheless, the relative novelty of neural computing argues for a comparative study of ANNs and MLR. In particular, forward variable addition to prevent overdetermination in small data sets is not a mature procedure in ANN identification, unlike MLR. In this report, the procedure was implemented and compared for both techniques.

The aim of this study was to develop a neural network to predict extubation outcome and to compare the performance of the ANN with both the performance of the MLR model and clinical expertise. A comparison among the three methods was performed by describing the dependency between false-positive and false-negative ranges using receiver operating characteristic (ROC) curves (14). The ROC technique provides a comprehensive assessment of accuracy of a predictor by screening the range of threshold values for the decision making. In addition, it provides a convenient display to characterize the properties of actual decisions, such as the decisions of individual clinical practitioners. Documenting automated prediction tools with ROC curves, as part of an automated system, will optimize the decision making surrounding ventilator weaning strategies by allowing for the simultaneous visualization of likelihood and type of error associated with the prediction.

METHODS

Study population. Data collection was based on an Institutional Review Board–approved retrospective chart review of medical records of infants who were born between January 1999 and October 2002. To ensure a satisfactory sample size, we included very low birth weight infants (1000–1500 g) and infants who weighed between 900 and 1000 g. Because birth weight may vary strongly for infants with comparable gestational age, selecting a sample based on gestational age could result in a wide spread of birth weights. Conversely, infants with birth weights <900 g are extremely immature and, therefore, clinically distinct from infants with larger weights (15). Similarly, infants with birth weights >1500 g are generally less premature and, therefore, less likely to develop respiratory distress syndrome. Thus, to be included in the study, infants had to have a birth weight between 900 and 1500 g and a diagnosis of respiratory distress syndrome. The infants had to be intubated and managed on mechanical ventilation within 6 h after birth. Data were collected from time of intubation to the first attempt of extubation. Extubation was considered a failure when an infant needed to be reintubated within 48 h after extubation.

The records of a total of 183 infants were selected for this study. An additional 27 infants were excluded from the study for the following reasons: pulmonary hypertension ($n = 1$), time between delivery and intubation >6 h ($n = 12$), extubation from ventilators other than assist-control or synchronized intermittent mandatory ventilation ($n = 5$), life support withdrawn without any previous attempt to extubate ($n = 4$), and no respiratory data retrievable from medical records ($n = 6$). The 183 infants who fulfilled the eligibility criteria were randomly assigned to two independent data sets, one set using ~70% of the total sample for development and training of the neural network ($n = 130$) and the second set using ~30% of the total sample for testing (out-of-sample validation) purposes ($n = 53$).

Descriptive characteristics were similar in the two data sets (Table 1). Mean (\pm SD) gestational age for the training set was 29.3 wk (\pm 2.0) compared with 29.1 wk (\pm 2.0) for the out-of-sample validation set. The mean birth weight was 1170 g (\pm 168) for the training set and 1145 g (\pm 182) for the

Table 1. Demographic characteristics of subject population

	Training set ($n = 130$)	Validation set ($n = 53$)	p value
Gestational age (wk; mean \pm SD)	29.3 \pm 2.0	29.1 \pm 2.0	0.47*
Birth weight (g; mean \pm SD)	1170 \pm 168	1145 \pm 182	0.39*
Time on ventilator (d; median)	2	2	–
Male sex	52.3% (68/130)	52.8% (28/53)	0.95†
Ethnicity			0.23†
White	62.3% (81/130)	54.7% (29/53)	
Black	33.1% (43/130)	45.3% (24/53)	
Hispanic	3.8% (5/130)	0	
Asian	0.8% (1/130)	0	
Outcome			0.24†
Extubation failure	16.9% (22/130)	24.5% (13/53)	

* Pooled t test.

† χ^2 test.

validation set. Median time on ventilator was 2 d for both the training set and the validation set.

Potential predictive variables. Fifty-one variables were identified by clinicians as having potential relevance for the prediction model. These included demographic variables (e.g. gestational age, gender, ethnicity, birth weight), Apgar scores, ventilator settings, capillary and arterial blood gas analysis results, and vital signs (for a complete list of variables, see Tables 2 and 3). To be included, variables were required to be part of routine evaluation of the infant. Consequently, only variables easily available at the bedside were considered. For variables recorded at routine evaluations, such as ventilator settings, blood gas analysis results, and vital signs, data from the two time points immediately before extubation were included.

Imputation of missing data. Missing data were found only in the ventilator setting recordings. Deleting all cases with

Table 3. Frequency for total sample and explanations of variables used

Variable	Frequency (n = 183)	Proportion
AB (arterial blood gas)	125	68.3%
AB (cap. blood gas)	58	31.7%
Balanced pattern (no)	115	62.8%
Balanced pattern (yes)	68	37.2%
Extubation failure	35	19.1%
Extubation success	148	80.9%
Mode (AC)	38	20.8%
Mode (SIMV)	145	79.2%
Overventilated (no)	88	48.1%
Overventilated (yes)	95	51.9%
Ethnicity (black)	67	36.6%
Ethnicity (Other)	6	3.3%
Ethnicity (white)	110	60.1%
Sex (female)	87	47.5%
Sex (male)	96	52.5%

Table 2. Range for total sample and explanations of variables used

Variable	Minimum	Maximum	Explanation
Age_D	1	31	Age in days (~ = time on ventilator)
APGAR_1	0	9	Apgar score at 1 min
APGAR5_1	-2	7	Difference between Apgar at 1 and 5 min
BE	-7.5	6	Base excess content in blood
BP	28.5	67	Blood pressure
CurrWeight	808	1600	Current weight
dBE	-4.5	1.7	Change in base excess since last measurement
dBP	-13.7	27.0	Change in blood pressure since last measurement
dFIO ₂	-30.0	21.8	Change in FIO ₂ since last measurement
dHCO ₃	-3.0	4.0	Change in HCO ₃ since last measurement
dIErat	-3.5	2.8	Change in I:E ratio since last measurement
dINSP	0.0	1.0	Change in INSP since last measurement
dMAP	-2.4	1.1	Change in MAP since last measurement
dPaCO ₂	-6.4	22.0	Change in PaCO ₂ since last measurement
dPaO ₂	-63.3	50.0	Change in PaO ₂ since last measurement
dPEEP	-3.0	1.3	Change in PEEP since last measurement
dPH	-0.2	0.1	Change in pH since last measurement
dPIP	-3.0	2.4	Change in PIP since last measurement
dPulse	-27.0	42.0	Change in heart rate since last measurement
dRATE	-15.0	10.6	Change in ventilatory breathing rate since last measurement
dRRatio	-3.3	3.4	Change in rate ratio (spont. vs vent.) since last measurement
dSaO ₂	-4.0	15.0	Change in SaO ₂ since last measurement
dTIME	0.3	51.6	Time between last two blood gas analyses
dV _T	-1.8	2.5	Change in V _T
FiO ₂	20	60	Oxygen content in air delivered by ventilator
Gst_age	25	35	Gestational age
HCO ₃	16	35	Bicarbonate content in blood
IEratio	1	17.2	I:E ratio (inspiratory to expiratory time)
INSP	0.23	0.5	Inspiratory time
Lag	0.25	17.5	Lag time between last blood gas result and extubation
MAP	4.2	10.9	Mean airway pressure
Paco ₂	23	65	Partial pressure of carbon dioxide in blood
Pao ₂	23	228	Partial pressure of oxygen in blood
PEEP	3	5	Positive end-expiratory pressure
pH	7.24	7.58	Acidity or alkalinity of blood
PIP	10	21	Peak inspiratory pressure
Pulse	112	187	Heart rate
R_ratio	0.4	6.0	Ratio: spontaneous breathing rate/ventilatory rate
Rate	10	60	Ventilatory breathing rate
Saline	0	26	Saline bolus within 24 h before extubation
SaO ₂	79	100	Oxygen saturation in blood
Theoph	0	8.3	Theophylline bolus within 24 h before extubation
V _T	3	15.7	Tidal volume
TXBETAME	0	1	Maternal betamethasone

FiO₂, fraction of inspired oxygen; PIP, peak inspiratory pressure.

missing data would not only significantly reduce the sample size but also lead to a potentially biased data set and result in the loss of potentially valuable information (16). A total of 74 infants had missing values for at least one of the following variables: mean airway pressure (MAP), inspiratory time (INSP), ratio of inspiratory to expiratory time (I:E ratio), and tidal volume (V_T). For infants ($n = 72$) for whom the MAP value closest to extubation was missing, we estimated the MAP from previously measured values of MAP, peak inspiratory pressure, positive end-expiratory pressure (PEEP), INSP, and expiratory time using the equation described by (17). In 52 cases for which INSP, I:E ratio, and V_T values were missing closest to extubation, last observation carried forward was used to impute the missing values. For two infants, no previous values were available for any of the four variables; therefore, overall means were used to impute missing values.

Neural network. For the development of the ANN, we used a web-based neural computing tool (bioinformatics.musc.edu/webnn) created and maintained at the Medical University of South Carolina, which follows guidelines suggested in (12) to automate extraction of signal from noise using multilayer, fully connected feed-forward networks. This report is referred to for further information regarding implementation of cross-validation and topology design criteria. In addition to the predictions and its corresponding residuals, that tool generates ROC curves for the median ANN predictor and analyzes the

sensitivity of predictions to the values of input variables. By selecting the number of hidden nodes by the minimum cross-validated predictive error and the number of inputs by the minimum of externally validated predictive error, this procedure is decoupling two components typically put together as a single pruning step. It should be noted that the sensitivity to an input is a function of all connection weights that emanate from the input and subsequently modulated by the connection between the hidden nodes and the output parameter, extubation success. In that sense, it is still the low value of connection weights, just as in conventional pruning, that will ultimately drive an input parameter to be excluded from the final ANN model.

Initially developed was a fully cross-validated neural network with optimized topology that included all 51 variables previously identified by clinicians as potentially predictive. The results of the corresponding sensitivity analysis were subsequently used in a forward addition scheme to determine the optimal number of input variables for the final ANN. Starting with the variable that had the largest sensitivity, we developed a series of five cross-validated ANNs and determined the area under the ROC curves for the training and out-of-sample validation sets and the median absolute difference between the areas for each ANN. The variable with the next largest sensitivity was added, and the process was repeated until all remaining variables had sensitivities close to

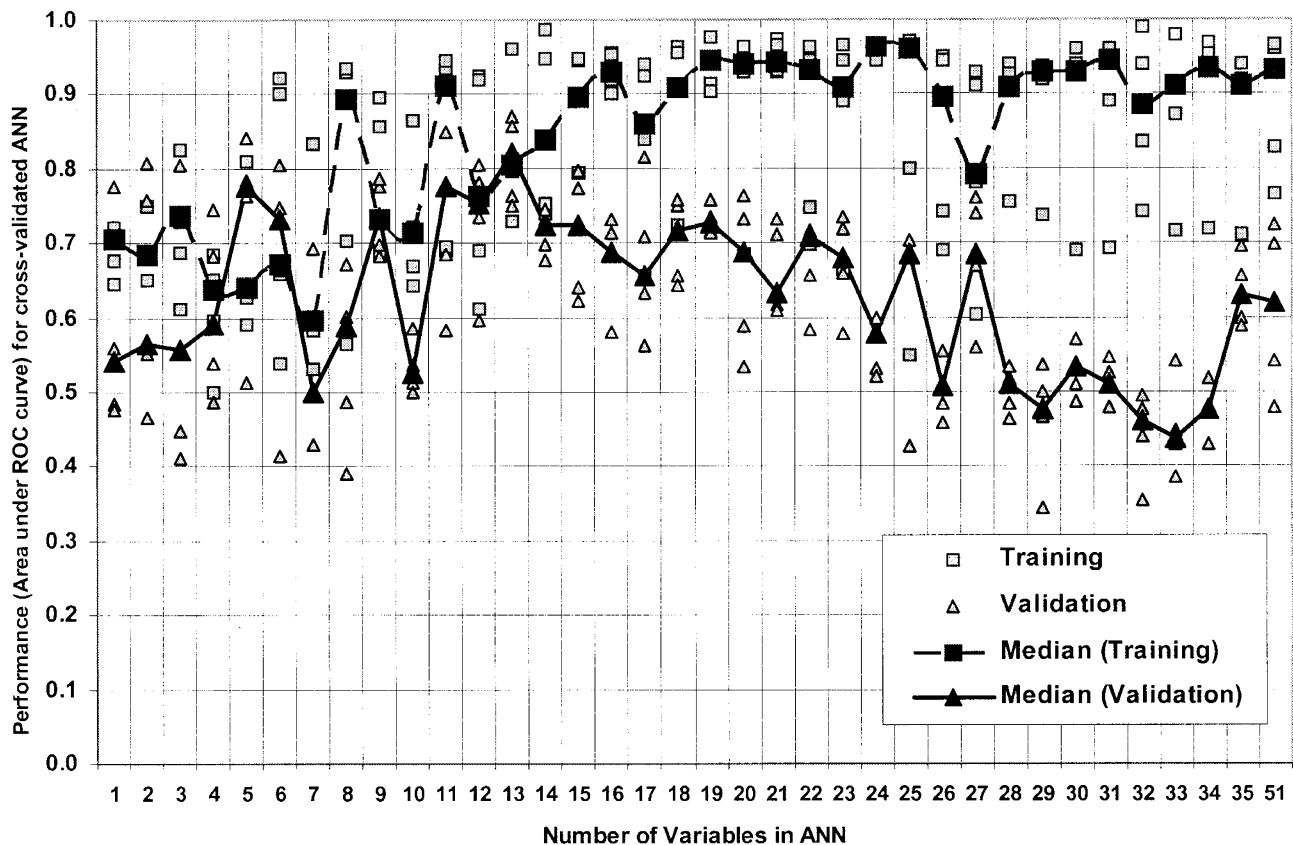


Figure 1. AUCs of five ANNs for training and validation sets using increasing numbers of variables for the ANN development. The performance of the ANN with the median absolute difference between the AUC (for training and validation) crosses over for small numbers of variables, whereas the median performance of the validation set increases until a maximum at which the curves diverge.

zero and the median performance of the ANNs did not change by adding additional variables. All median-performing ANNs were compared; the ANN with the smallest absolute difference between the areas under the ROC curves was selected as optimal (see "ROC curves"). The areas under the ROC for all of the 180 ANNs trained are provided in Figure 1. Variable addition proceeds until divergence occurs between cross-validated (training) and test (independent validation) results (for 13 variables in Fig. 1).

Logistic regression. All 51 clinician-selected variables also were considered in the multivariate logistic regression analysis. In the development of the MLR model, possible correlations among variables were determined in a first analysis step using SAS software (18). For variables having correlations $|\rho| > 0.6$, interaction terms were created and entered along with the variables themselves. Furthermore, variables that were known or suspected to be correlated were included as interaction terms of first and second order. All 51 variables and their interaction terms were subjected to forward and stepwise selection procedures. The most parsimonious model was used for comparison with the ANN and clinical experts (CEs).

CEs. Four neonatologists were provided with the data sets that contained only the information used to develop the mathematical prediction models. The clinicians were then asked to predict extubation failure or success for the infants in the data sets. The results were used to plot sensitivity and specificity of the predictions.

ROC curves. ROC curves (14) and the area under the curves (AUC) were used to compare ANN, MLR, and expert models for predicting extubation outcome. The accuracy of a series of diagnostic decisions depends on the amount of overlap between two distributions: the distribution of patients with a disease (signal) and the distribution of patients without the disease (no signal). The ability to distinguish between diagnoses will increase as the amount of overlap of the two distributions decreases. The sensitivity of the diagnostic decision is the conditional probability of the true-positive response (patients who had the disease and correctly received a diagnosis of having the disease, or, in our case, infants who succeeding extubation and had been predicted to succeed). The specificity of the decision is the conditional probability of the true-negative response (an infant who fails extubation and had been predicted to fail, with failure being defined as reintubation within 48 h of extubation). The true-positive rate (sensitivity) is plotted *versus* the false-positive rate (1 minus the specificity). The resulting ROC curve depends on the degree of overlap of the two patient distributions. By moving along the curve, sensitivity and specificity can be changed and the threshold for the best performance can be determined.

Both the ANN and the MLR return prediction scores between 0 and 1. Depending on the chosen threshold, these scores are translated into extubation success (1) or failure (0). If the threshold is increased, then the number of predicted successes will decrease, whereas the number of predicted failures increases, and therefore, the number of false-negative and false-positive predictions will change. The threshold can be moved along the ROC curve (19) to determine the best result for both sensitivity and specificity

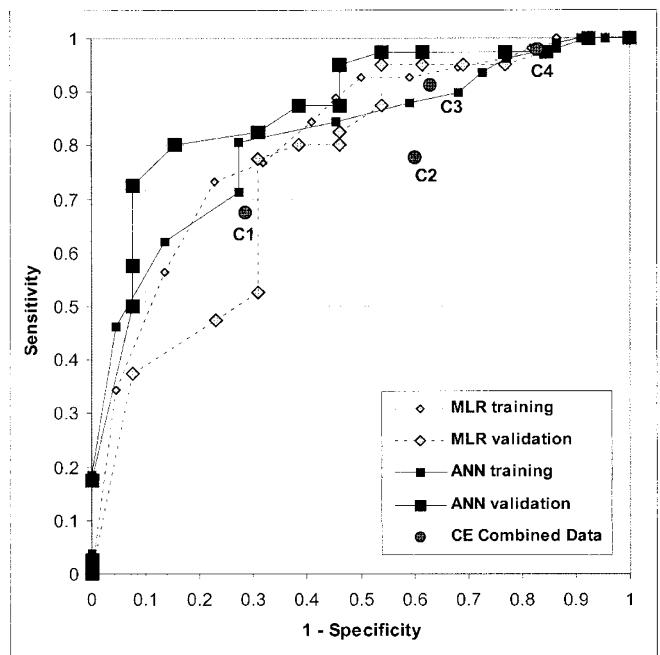


Figure 2. Comparison of the performance of ANN, MLR, and CE using the AUCs for training and (independent) validation sets. The ANN achieved an AUC of 0.81 for training and 0.87 for validation compared with 0.81 (training) and 0.75 (validation) for the MLR.

of the predictive tool. The AUC corresponds to the likelihood that a randomly selected infant who is extubated successfully will be considered more likely to be extubated successfully than a randomly selected infant who fails extubation. The area represents a measure for the performance of a predictive tool and allows the comparison of performance among different models, such as ANN and MLR.

RESULTS

The ANN that achieved the smallest absolute difference between AUCs for the training and out-of-sample validation sets included 13 input variables (sensitivity $> 1.1\%$). It is noteworthy that the validation corresponds to predictions for an independent data set; it does not correspond to the cross-validated data set, which is subsampled from the training set. This approach reduces the potential for validation bias by pseudo-replicates, *i.e.* the repeated use of a single sample. The final network consisted of an optimal architecture of three layers with 13 input nodes in the first layer¹; 7 hidden nodes in the second layer; and a single output node, corresponding to the prediction of extubation outcome. The four input variables that had the largest sensitivities and, therefore, were considered most important were pH, saturation of arterial oxygen (Sao_2), gestational age, and PEEP (a ventilatory parameter). The ANN predicted 82% correctly for the training set and 85% for the validation set using a threshold of 0.5. The AUC for the training set was 0.81 compared with 0.87 for the validation set (Fig. 2).

¹ For the 13 variables selected, see methods and text further below in this section and Fig. 1 for individual results.

The most parsimonious logistic regression model obtained from stepwise selection contains the four variables that retained statistical significance: V_T , partial arterial pressure of CO_2 , ventilation mode, and ethnicity. The Akaike information criterion for the model was 102.0, the lowest value for all models tested. The Hosmer and Lemeshow goodness-of-fit test returned $p = 0.9$, suggesting a good fit of the model. The maximum likelihood estimates of the variables included in the model were largest for ethnicity (0.54) and smallest for ventilation mode (-0.66), as shown in Table 4.

The logistic regression model achieved a correct prediction rate of 84% in the training set and 81% in the validation set. The AUC for the logistic regression model was 0.81 for the training set and 0.75 for the validation set (Fig. 2).

The neonatologists (CEs) predicted on average 78% ($\pm 6.4\%$; range, 72–84%) correctly in the training data set and 70% (± 9.4 ; range, 59–79%) in the validation data set. Because the clinicians predicted success (1) or failure (0), as a dichotomous outcome instead of continuous prediction scores between 0 and 1 as returned by the ANN and the MLR, the procedure resulted in a single pair of sensitivity/specificity per clinician. We used an approximation based on geometric rules proposed by Cantor and Kattan (20) to determine the ROC curves and the corresponding AUCs. However, the results do not reflect the diverse decision-making approaches of the four clinicians, such as the preference to err on the side of extubating an infant “too early” rather than “too late” compared with a more conservative approach that influences the false-positive and false-negative responses, respectively. This lack of distinction between the different approaches questions the applicability of the Cantor and Kattan procedure for this comparative analysis. Therefore, only the single sensitivity/specificity pairs for each clinician were considered.

Furthermore, the clinicians were provided only with the combined data set, because for them, no training/validation phase was involved in predicting extubation outcome. The results for the sensitivity/ $1 -$ specificity pairs achieved by the CEs ranged from 0.68/0.29 to 0.98/0.83 (Fig. 2). These numbers reflect the different preferences of the clinicians with clinician 1 preferring to extubate infants later, resulting in increased numbers of false negatives, whereas clinicians 2–4 tend to extubate earlier, leading to increased numbers of false positives.

Figure 1 shows the change in performance of a series of ANNs for training and validation data with increasing number of input variables used to develop the ANN (see “Methods”). The two curves [the median performing ANNs are chosen for each set of iterations, as recommended by (12)] approximated the performance of the ANNs for the training and validation

set. With the addition of variables to the models, the performance for both sets increased until a maximum for the validation set was reached after which the curves started to diverge. The upper curve, depicting the performance for the training set, further increased, indicating overfitting of the ANN with increasing number of variables, thus reducing the generalizability of the models to new data. The lower curve, showing the performance of the ANNs for the validation set, declined at first and stabilized around a performance close to 0.5, which is equal to chance, suggesting loss of performance from noise (*i.e.* inclusion of many variables with very small sensitivities), therefore also negatively affecting the generalizability of the models.

Divergence of the two curves—cross-validated training and out-of-sample validation—started (Fig. 1) after the inclusion of 13 variables in the ANN development. The most consistent performance for training and validation data also was observed for the series of ANNs with 13 input variables.

A comparison of the sensitivity analysis for 22 ANNs, each developed using the 13 input variables, is shown in Figure 3. The majority of ANNs assigned the largest sensitivity to pH, the second largest to SaO_2 , and the third largest to gestational age. Ranking of the variables according to their importance across all ANNs with 13 input variables and across all ANNs with the original 51 input variables showed that 12 of the 13 variables used for the final ANN were found among the first 13 variables of the original variable set. The variables pH and SaO_2 were ranked first and second, respectively, in both sets of ANN. Gestational age was ranked third in the ANNs with 13 variables and fourth for 51 variables. PEEP, MAP, and V_T were ranked the same way in both sets (Table 5).

DISCUSSION

In this article, we compared the performance of ANN, MLR, and CE models as predictors of extubation outcome to explore both ANN and MLR with the most advanced methods to determine the superior performer. Our data showed that the ANN model outperformed MLR and CE in accurately predict-

Table 4. Maximum likelihood estimates for variables included in final logistic regression model

Variable	Estimate	p value
Intercept	3.38	0.047
V_T	0.36	0.042
$Paco_2$	-0.11	0.001
Ethnicity (black)	0.54	0.043
Ventilation mode	-0.66	0.033

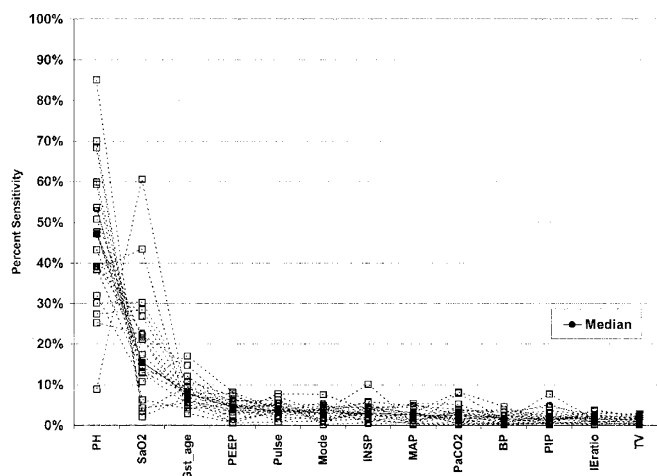


Figure 3. Comparison of assignment of sensitivities to 13 input variables used for development of 22 ANNs showing the majority of largest sensitivities assigned to pH; the smallest sensitivities were assigned to tidal volume (TV).

Table 5. Comparison of mean rank for ANNs developed with 13 vs 51 input variables

51 variables	Mean \pm SD	13 variables	Mean \pm SD
pH	1.0 \pm 0.0	pH	1.2 \pm 0.4
Sao ₂	2.4 \pm 0.7	Sao ₂	2.1 \pm 0.8
INSP	4.9 \pm 4.1	Gestational age	3.2 \pm 1.3
Gestational age	5.9 \pm 4.1	Pulse	6.0 \pm 2.4
PEEP	7.2 \pm 3.4	PEEP	6.1 \pm 2.5
Sex	8.8 \pm 5.1	INSP	7.0 \pm 2.5
Pulse	9.0 \pm 5.3	Mode	7.7 \pm 3.2
Mode	11.8 \pm 4.1	Paco ₂	8.7 \pm 2.7
Paco ₂	12.9 \pm 7.8	PIP	9.2 \pm 2.5
MAP	13.5 \pm 8.2	MAP	9.2 \pm 3.0
PIP	13.6 \pm 6.9	I:E ratio	10.1 \pm 2.0
I:E ratio	13.8 \pm 4.2	BP	10.1 \pm 2.4
V _T	13.8 \pm 5.0	V _T	10.4 \pm 1.9

ing extubation outcome in premature infants who were on mechanical ventilators. The ANN achieved the largest AUC of all models (0.87). The ANN predicted 86% of the extubation successes and 86% of the failures correctly compared with 84% and 56%, respectively, for the MLR and on average 84% and 41%, respectively, for the CE. These results build on the work of Sargent (21), who reported that ANNs may outperform or match logistic regression. He cautioned that because ANNs may occasionally underperform statistical models, a comparison is necessary for each new proposed application.

Logistic regression has long been viewed as the “gold standard” for clinical prediction problems (3,22,23). This method has the advantage of providing coefficient estimates that can be easily interpreted, with the coefficients of the MLR model related to the mathematical odds of failing extubation. Therefore, the relative importance of each variable included in the logistic regression model can be assessed directly, *i.e.* it is explicit in the model formulation. Conversely, ANNs have been considered “black boxes” because the weights of the connections between the nodes of the different layers have no clear interpretation, *i.e.* relative importance is implicit in the combination of connection weights (24). However, this is more of a numerical than a fundamental issue. By resorting to sensitivity analysis, light is shed on the “black box,” providing a quantitative assessment of the relative importance of each variable. Comparing with explicit model formulations such as MLR, the main difference is that the relative importance of each variable will depend on the value of the variables themselves, rather than being fixed values. Nevertheless, because the decision process by the ANN is not easily understood without the additional effort of performing sensitivity analysis, clinicians’ acceptance of prediction systems that used ANNs historically has been low. With the use of sensitivity analysis, however, we are now able to explain the discriminating dependency of ANNs and to validate those against the expertise of clinicians and traditional statistical methods, such as MLR. By assessing the effect of change in input variables on the output, sensitivity analysis allows quantification of variables used to develop ANN models. This, in turn, allows us to discard variables with negligible effect on the output across the range of values observed for the independent variables and to de-

velop and select ANN models that are more robust and more general.

Simpler models with fewer variables are often better understood than more complex ones and, therefore, preferred for use in clinical applications. However, the MLR model, which contains only four variables, is difficult to explain in the face of a problem as complex as the decision of whether to extubate a premature infant—involving a large amount of information of which only part is explicitly processed in the human decision-making process and recognized as such by clinicians. Furthermore, two of the four variables included in the MLR model were “ethnicity African-American *versus* Caucasian” and “ventilation mode.” While ventilation mode seems to be important in the decision of some clinicians, for other clinicians it is not. Clinicians who deem it important typically do not extubate infants who are on “assist-control” but rather change the ventilation mode to synchronized intermittent mandatory ventilation before an extubation attempt. Being of African-American descent did not seem to have any or only a very small impact on the decision of whether to extubate. For these reasons, acceptance of a decision-support system based on the MLR model by our clinicians was understandably low. In comparison, the ANN with the best performance used 13 variables to predict extubation outcome, most of which can be easily recognized by the clinicians as being relevant. Among the 13 selected variables were blood gas parameters (pH, partial arterial pressure of CO₂, and Sao₂), ventilator settings (peak inspiratory pressure, PEEP, MAP, INSP, I:E ratio, V_T, and mode), blood pressure, heart rate, and gestational age of the infant, all of which were considered by clinicians during their decision making. This similarity between ANN machine-learning and human decision-making processes reflects that both share the capacity to accommodate complex nonlinear relationships and both are experiential, *i.e.* learning from experience is not conditioned by model formulation.

Finally, ANNs that are developed with data obtained at a certain institution are tailored to the needs of each specific institution—a seeming limitation for their generalizability. However, because ANNs learn from the data provided, they can be seamlessly extended to adapt to the unique challenges of different institutions by simply incorporating the new data in the training and validation of the model.

CONCLUSION

Our experience in modeling the prediction of extubation outcome in mechanically ventilated premature infants showed that ANN models performed best compared with MLR models and clinical expertise, particularly with regard to less experienced practitioners, which raises the additional prospect of using ANN models for education and training. We were able to identify 13 relevant input variables for the ANN model using sensitivity analysis. A system for prediction of extubation outcome based on our ANN model provides a powerful tool to assist clinicians in their decision making.

Acknowledgments. We thank Dr. William M. Southgate and Dr. John B. Cahill, Jr. (Department of Pediatrics, Medical

University of South Carolina) for ready participation in the assessment of clinical expertise in the prediction of extubation outcome. We also thank Tom G. Smith (Center for Academic Excellence, Medical University of South Carolina) for helpful comments and suggestions in the preparation of this manuscript.

REFERENCES

1. Meadow W, Lantos J 2003 Ethics at the limit of viability: a premie's progress. *Neoreviews* 4:E157–E162
2. Spitzer AR 1996 *Intensive Care of the Fetus and Neonate*. Mosby, St. Louis, pp 553–570
3. Snowden S, Brownlee KG, Smye SW, Dear PRF 1993 An advisory system for artificial ventilation of the newborn utilizing a neural network. *Med Inform* 18:367–376
4. Zernikow B, Holtmannspoetter K, Michel E, Pielemeier W, Hornschuh F, Westermann A, Hennecke KH 1998 Artificial neural network for risk assessment in preterm neonates. *Arch Dis Child Fetal Neonatal* Ed 79:F129–F134
5. Ambalavanan N, Carlo WA 2001 Comparison of the prediction of extremely low birth weight neonatal mortality by regression analysis and by neural networks. *Early Hum Dev* 65:123–137
6. Braithwaite EA, Dripps J, Lyon AJ, Murray A 2001 Neural networks for neonatal intensive care. In: Dybowski R, Grant V (eds) *Clinical Applications of Artificial Neural Networks*. Cambridge University Press, Cambridge, pp 102–119
7. Schumacher M, Rossner R, Vach W 1996 Neural networks and logistic regression: part I. *Comput Stat Data Anal* 21:661–682
8. Landry MD, Sibbald WJ 2001 From data to evidence: evaluative methods in evidence-based medicine. *Respir Care* 46:1226–1235
9. Weigend AS, Rumelhart DE, Huberman BA 1991 Generalization by weight elimination with application to forecasting. In: Lippmann RP, Moody J, Touretzky DS (eds) *Advances in Neural Information Processing Systems 3*. Morgan Kaufman, San Francisco, pp 875–882
10. Bishop CM 1995 *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, pp 116–377
11. Reed RD, Marks RJ 2nd 1999 *Neural Smoothing—Supervised Learning in Feedforward Artificial Neural Networks*. A Bradford Book, MIT Press, Cambridge, pp 239–290
12. Almeida J 2002 Predictive non-linear modeling of complex data by artificial neural networks. *Curr Opin Biotechnol* 13:72–76
13. Myers RH 1990 *Classical and Modern Regression with Applications*. 2nd Ed. Duxbury Press, Belmont, pp 317–325
14. Goodenough DJ, Rossman K, Lusted LB 1974 Radiographic applications of receiver operating characteristic (ROC) curves. *Radiology* 110:89–95
15. York J, DeVoe M 2002 Health issues in survivors of prematurity. *South Med J* 95:969–976
16. Ennett CM, Frize M, Walker CR 2001 Influence of missing values on artificial neural network performance. *Medinfo* 10:449–453
17. Carlo WA, Martin RJ 1986 Principles of neonatal assisted ventilation. *Pediatr Clin North Am* 33:221–37
18. SAS Software, Version 8.2 of the SAS System for Windows 95/98. SAS Institute Inc., Cary, NC
19. Bradley AP 1997 The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pat Rec* 30:1145–1159
20. Cantor SB, Kattan MW 2000 Determining the area under the ROC curve for a binary diagnostic test. *Med Decis Making* 20:468–470
21. Sargent DJ 2001 Comparison of artificial neural networks with other statistical approaches—results from medical data sets. *Cancer* 91:1636–1642
22. Clermont G, Angus DC, DiRusso SM, Griffin M, Linde-Zwirble WT 2001 Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. *Crit Care Med* 29:291–296
23. Nguyen T, Malley R, Inkelis S, Kuppermann N 2002 Comparison of prediction models of adverse outcome in pediatric disease using artificial neural networks and logistic regression analyses. *J Clin Epidemiol* 55:687–695
24. Lisboa PJ 2002 A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural Netw* 15:11–39