34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Prediction of Extubation Failure for Neonates with Respiratory Distress Syndrome Using the MIMIC-II Clinical Database

Arthur Mikhno, MS, *Member, IEEE*, Colleen M. Ennett, PhD, *Member, IEEE*

Abstract— **Extubation failure (EF) is an ongoing problem in the neonatal intensive care unit (NICU). Nearly 25% of neonates fail their first extubation attempt, requiring re-intubations that are associated with risk factors and financial costs. We identified 179 mechanically ventilated neonatal patients that were intubated within 24 hours of birth in the MIMIC-II intensive care database. We analyzed data from the patients 2 hours prior to their first extubation attempt, and developed a prediction algorithm to distinguish patients whose extubation attempt was successful from those that had EF. From an initial list of 57 candidate features, our machine learning approach narrowed down to six features useful for building an EF prediction model: monocyte cell count, rapid shallow breathing index, fraction of inspired oxygen (FiO_2), heart rate, PaO_2/FiO_2 ratio where PaO_2 is the partial pressure of oxygen in arterial blood, and work of breathing index. Algorithm performance had an area under the receiver operating characteristic curve (AUC) of 0.871 and sensitivity of 70.1% at 90% specificity.**

Keywords— Extubation failure, neonatal intensive care unit, outcomes estimation, respiratory distress syndrome

I. Introduction

More than 80% of preterm infants born at <31 weeks gestational age will develop respiratory distress syndrome (RDS) requiring endotracheal intubation and mechanical ventilation support in the neonatal intensive care unit (NICU) [1]. If improvement is noted within a few days or weeks, a clinical decision is made to extubate, and place the patient on a nasal continuous positive airway pressure (CPAP) trial. Across studies, 15%-40% of infants fail the trial, and must be re-intubated [2-4]. The failed trial exposes the infant to respiratory shock requiring re-intubation, increased ventilator support due to alveolar collapse (or atelectasis), and puts them at additional risks associated with the intubation procedure [5]. While it is important to identify the ideal time point for extubation, it is equally important to minimize the infant's time on artificial ventilation to reduce the risk of ventilator incidents and health complications, such as bronchopulmonary dysplasia or airway trauma [6].

A clinical decision to extubate is based on a large amount of observation, physiological measurement, and clinical experience that is weighed by the neonatologist, and yet remains a difficult task despite advances in technology. An automated prediction system that helps the clinician make a more informed decision could reduce the risks and prevalence of extubation failure (EF). A number of prospective and retrospective studies were done to identify risk factors and indices that may predict EF. The risk factors are from a variety of sources such as demographics (e.g., gestational age, birth weight) [2], ventilator settings (e.g., breath rate, fraction of inspired oxygen, tidal volume), pulmonary mechanics measurements (e.g., work of breathing, minute volume, airway resistance) [3],[4], and blood gas measurements (e.g., base excess, partial pressure of arterial oxygen) [2],[4]. Validation of these metrics is ongoing. Other indices, such as diagrammatic pressure-time index and noninvasive respiratory muscle pressure-time index have been designed specifically for predicting extubation outcome in preterm neonates [4]. However, measurement of these parameters requires an intervention like taking the patient off the ventilator, thus making it impractical for automated prediction systems.

There have been attempts to build EF prediction models using artificial neural networks and multivariate logistic regression [7]. Potential predictors were selected through a literature review, and subsequently rated by clinicians prior to model building. While this approach is effective at leveraging the knowledge of clinicians, it may also miss variables overlooked in clinical practice.

A rule-based machine learning approach was applied to the 12,000+ adult records in the MIMIC-II (Multi-Parameter Intelligent Monitoring of Intensive Care) clinical database to identify variables, and build a model for predicting respiratory instability in the adult intensive care unit (ICU) with good performance [8]. The initial 16 candidate features were pruned to four features based on having high discriminatory ability, and incorporated into the final model.

In the MIMIC-II clinical database, Version 2.6 [9], laboratory LOINC codes allow use of digital laboratory results, instead of the less reliable manually-entered data. By applying machine learning tools, feature pre-selection was unnecessary, facilitating discovery of novel features that were previously ignored. This paper presents the machine-learning techniques used to locate features relevant to EF, and to develop a model for predicting extubation failure or success 2 hours prior to the extubation attempt.

II. Methods

A. Data Collection

Retrospective data from the MIMIC-II database (released August 2011) was used for this study. This version contained data collected from over 7800 neonates during their stay in the NICU. Recorded variables included patient demographic data, manually-validated patient monitoring data (i.e., heart rate, blood pressure), chart data, laboratory results, ventilator settings and values, ICD-9 codes, LOINC codes, and free-text nursing progress notes. Neonatal data were imported into

A. Mikhno is with the Biomedical Engineering Department, Columbia University, New York, NY, USA (email: am2679@columbia.edu).

C. M. Ennett is with the Clinical Decision Support Solutions Department, Philips Research North America, Briarcliff Manor, NY, USA.

a PostgreSQL database, and subsequently selected variables were exported into a Matlab (Mathworks, 2009b) framework. Using the LOINC codes, the manually-recorded chart data were compared to the lab table. If copying from the lab results was evident, we used the lab results instead. Where appropriate, data from multiple chart or lab variables were merged, and replaced by a single variable. In general, lab results were often more complete and extensive than the data recorded in charts.

B. Ventilation Times

Direct information of whether the patient is on or off the ventilator at any given time is not available in MIMIC-II. To infer ventilation status, we used a heuristic approach, empirically utilizing information from multiple chart variables that were informative of ventilation status: Airway, Ventilator Mode, Respiratory Support, Breath Rate, and Oxygen Delivery Device. Ventilation status was coded as "intubated" or "not-intubated". An "extubation" event was defined as the time point where "intubated" status changed to "not-intubated". This rule-based method was used as a first-pass method for patient selection. We manually verified the results against nursing notes to ensure the correct status was assigned to each event of each patient.

C. Patient Selection

Two groups of patients were selected for our reference data set. Inclusion criteria (Table I) were based on previous studies of extubation failure in neonates. Patients were 23 to 31 weeks of gestational age, and intubated within 24 hours of birth in the same hospital. The duration of intubation was at least one day, and all patients were assigned ICD-9 code 769, indicating a diagnosis of Respiratory Distress Syndrome. The patients were subsequently segregated into extubation failure (EF) and no extubation failure (noEF) groups. Any patient that was re-intubated within 48 hours of the first extubation attempt was placed in the EF group, otherwise they were designated noEF.

From the initial set of 7800 patient records, there were 242 patients that met our inclusion criteria. Sixty-three of these patients had to be excluded from the study based on information in the nursing notes. Fifteen extubation events could not be confirmed, and forty-eight extubation events were actually self-extubations by the infant meaning the extubation was not initiated by a clinician. The final data set consisted of a total of 179 patients, 24 EF (who were re-intubated within 48 hours of being extubated) and 155 noEF. Descriptive statistics are summarized in Table II.

D. Attribute Selection

With the goal of finding features that can be used to distinguish between EF and noEF patients, we first evaluated a preliminary set of features for correlation and discrimination ability. These features were individual variables available in MIMIC-II, and indices calculated from them. The preliminary feature list included 100+ routinely monitored vitals, ventilator settings, laboratory results, and calculated indices. Due to limitations of the MIMIC-II data, some of the calculated indices were approximations of traditional indices used in literature. For example, the Rapid Shallow Breathing Index (RSBI) is usually defined as RSBI = RR/V_T, where RR is the respiratory rate and V_T refers to tidal volume. By the original definition of RSBI, V_T and RR are measured without ventilator support [10]. In MIMIC-II, V_T and RR are always measured with active ventilator support (at least for our patient population of intubated neonates), so it should be noted that our calculation is not as accurate, but is still a metric of patient respiratory ability. We applied similar approximations to ten other calculated indices, where relevant ones are shown in the Appendix.

TABLE I. INCLUSION CRITERIA

	Intubated within 24 hours of birth
	Intubated for at least 24 hours
	Gestational age 23-31 weeks
	ICD9 code 769 present
EF	Re-intubated within 48 hours extubation
noEF	Remains extubated 48 hours after extubation

TABLE II. DESCRIPTIVE STATISTICS

	noEF (n=155) (M=89, F=66)			EF (n=24) (M=17, F=7)		
	Min	Max	Mean (std)	Min	Max	Mean (std)
Gestational age	23	29	26.7 (1.9)	23	29	25.9 (2.1)
Birth weight (g)	485	2240	1104 (340)	580	1395	942 (224)
Time from birth to intubation (h)	0	24	12 (7)	2	23	14 (7)
Time from birth to extubation (h)	31	1282	176 (256)	36	1292	256 (353)
Duration of intubation (h)	24	1259	164 (255)	24	1277	242 (353)
Length of stay (d)	7	279	68 (39)	11	173	96 (39)

g = grams, h = hours, d = days

The preliminary feature list (100+ features) was subsequently pruned. Pearson's correlation coefficient was calculated for each feature against every other feature. Features with correlations of $R^2 > 0.7$ and p-value > 0.05 were excluded from analysis. We then assessed the remaining features' ability to differentiate between EF and noEF classes based on the ss1 statistic using minimum and maximum values calculated over a 2-hour interval prior to the extubation attempt. The ss1=sensitivity+specificity-1, also known as Youden's index, represents the trade-off between sensitivity and specificity. Higher ss1 values signify better discrimination between patient populations. Although other intervals (e.g., 6, 12, and 24 hours) could have been chosen, we felt the 2-hour interval reflects most accurately on the patient's condition at the time of extubation. In our analysis, features with ss1<0.20 were noted, and if they were not previously mentioned in the literature as possible predictors in weaning trials, they were excluded. Finally, we excluded features that were recorded in <15 patients of the EF group to maintain a reasonable EF cohort for statistical analysis.

Our final feature set consisted of 57 features, where 19 features had an ss1 ≥ 0.2, and 38 features had an ss1 < 0.2. A representative list of feature's ss1 values is presented in Table III. From this table it is evident that some features with ss1 < 0.2 are sensible features to keep in the analysis: Positive Inspiratory Pressure (PIP) is a ventilator setting that is adjusted based on the patient's need; Respiratory Rate (RR)

TABLE III. SS1 OF SELECTED FEATURES

Feature	# noEF	# EF	Min	Max
Monocytes	153	22	0.36	0.36
PaO_2	154	22	0.26	0.32
Total Minute Volume	111	17	0.12	0.31
Birth Weight	155	24	0.3	0.3
SaO_2	155	24	0.15	0.12
PF Ratio	154	22	0.17	0.21
Respiratory Rate	155	24	0.14	0.11
PIP	154	24	0.11	0.11

SaO_2=arterial blood oxygen saturation, PF Ratio=PaO_2/FiO_2 ratio, PIP=peak inspiratory pressure

indicates the patient's respiratory drive. Other features such as Monocytes have not been reported in weaning trials but the ss1 indicated that it may be useful. Not all patients had every feature recorded, so the noEF and EF group sizes were not constant. For instance, for Total Minute Volume EF=17, while for PIP, EF=24. This means that in statistical analysis some models would inevitably contain fewer subjects than others.

E. Statistical Analysis

A total of 57 features were derived for statistical analysis. On deciding whether to use the minimum or maximum values calculated over the 2-hour interval prior to extubation, the calculation with highest ss1 was used. A logistic regression model building approach was used to generate candidate models for EF prediction. All combinations of 3 features, with and without interaction terms, were bootstrapped 100 times to obtain robust estimates of area under the receiver operating characteristic (ROC) curve (AUC) and sensitivity at 90% specificity ($SENS_{SP90}$). In a clinical setting, higher specificities are preferred to reduce false alarms. Candidate models with AUC<0.8 or $SENS_{SP90}$< 0.5 were excluded from subsequent analysis. The two models with the highest AUC and $SENS_{SP90}$, and that did not share any features, were combined into a final 6 feature (+/- interaction terms) EF prediction logistic regression model. Individual and combined models were bootstrapped 100 times to obtain mean ROC curves, estimates of AUC, and estimates of model parameters.

III. RESULTS

A. Candidate Models

Logistic regression was used to screen all combinations of 3 features from a pool of 57 features. This yielded a total of 58,520 candidate models (29,260 with and 29,260 without interaction terms). Overall, 43 models had an AUC>0.8 and 93 models had a $SENS_{SP90}$>0.5, however, only 13 models met criteria for both. The top performing model in terms of $SENS_{SP90}$ was [FiO_2, Monocytes, Total RSBI, *] with an AUC and $SENS_{SP90}$ of 0.839 and 0.614, respectively, where the * indicates an interaction term. The rest of the 13 models had common features (i.e. Monocytes), therefore they could not be combined with the top model, so we relaxed our criteria to AUC>0.7. Of the 92 models that had both AUC>0.7 and $SENS_{SP90}$>0.5, only 6 had features unique from the top model. The highest performing of these, in terms of $SENS_{SP90}$, was [Heart Rate, PF ratio, Work of Breathing, *] with an AUC and $SENS_{SP90}$ of 0.730 and 0.559, respectively. Combining the two models above yielded the highest performance with an AUC and $SENS_{SP90}$ of 0.872 and 0.701, respectively. Table IV presents candidate models and the combined model, sorted by $SENS_{SP90}$.

TABLE IV. MODELS WITH AUC > 0.70 AND $SENS_{SP90}$ > 0.50

Candidate Model	AUC	$SENS_{SP90}$
FiO_2, Monocytes, Total RSBI,*	0.839	0.614
Monocytes, Neutrophils, Total RSBI,*	0.847	0.522
V_T [Ventilator], Monocytes, Neutrophils,*	0.825	0.522
Lymphocytes, Monocytes, Total RSBI,*	0.823	0.501
Hematocrit, Monocytes, Spontaneous RSBI	0.817	0.500
...73 more models that include a Monocytes term…		
Heart Rate, PF ratio, Work of Breathing,*	0.730	0.559
PO_2,PF ratio, Work of Breathing,*	0.731	0.548
HGB,PF ratio, Work of Breathing,*	0.727	0.507
Total Bili, aA ratio, Work of Breathing,*	0.704	0.505
V_T [Spontaneous], Hematocrit, aA ratio,*	0.710	0.505
FiO_2, Monocytes, Total RSBI,* + Heart Rate, PF ratio, Work of Breathing,*	0.871	0.701

*interaction term

We generated bootstrapped ROC curves to gauge the performance of the combined model relative to the two individual models. Visually, the ROC curve followed the AUC results above. The (HR,PFratio,WOB,*) model had the worst performing ROC curve followed by the (Monocytes,RSBI,FiO_2,*) model. Combining the two models yielded a better ROC, shown in Figure 1. We obtained bootstrapped estimates of the logistic regression model coefficients for the combined model given by (1).

$$z = -13.4 + 0.16(Monocytes) + 0.14(RSBI) + 0.41(FiO_2) - 0.04(HR) + 0.04(PFRatio) + 2x10^{-2}(WOB) - 2x10^{-2}(*) - 7.9x10^{-4}(**) \quad (1)$$

*=*Monocytes*RSBI*FiO*$_2$
**=*HR*PFratio*WOB*

where FiO_2 = fraction of inspired oxygen, HR = heart rate, PFratio = ratio of PaO_2 to FiO_2, RSBI = rapid shallow breathing index, and WOB = work of breathing.

The maximum FiO_2 had the highest coefficient, 0.41, followed by Monocytes, 0.16, and RSBI, 0.14. All other coefficients were <0.1.

IV. CONCLUSION

We introduced a new model for EF prediction developed with logistic regression, and six variables were discovered through machine learning techniques applied to patient records in the MIMIC-II clinical database. All variables in the model are routinely recorded in the NICU, allowing for the possibility of developing a real-time clinical decision support system. Performance of the model (AUC~0.87) is similar to models developed with other techniques and data sets [7]. This is encouraging because our method did not require a complex prospective trial, and uses routinely recorded variables. These features should be validated by a clinician, but were taken from previous studies.

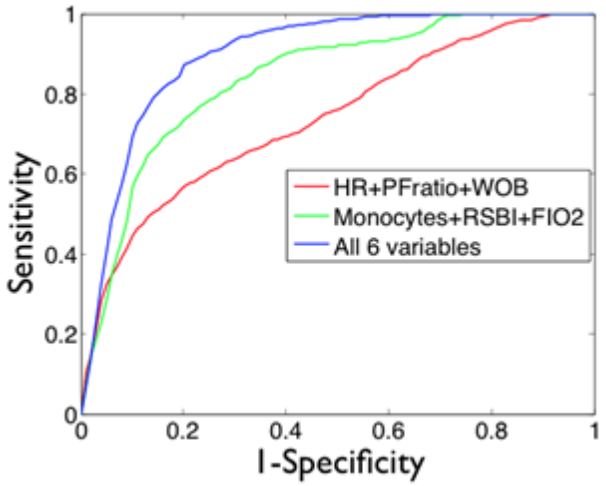

Figure 1. ROC curves for individual and combined models

Our model also contains a unique combination of variables, including PFratio, previously reported as an insignificant risk factor [4], and a novel variable, Monocyte cell count, not mentioned as a predictor or risk factor in any previously published reports. There is evidence that preterm neonates with RDS have altered levels of monocytes compared to healthy preterm neonates [10]. Monocyte alterations are also implicated in neonatal sepsis and other perinatal complications [11]. It is therefore plausible that diagnosis and treatment of basic underlying illness may improve extubation outcome. Finally, the predominant coefficients in our model correspond to Monocytes, PFratio and FiO_2, indicating that a combination of underlying illness respiratory insufficiency may influence the likelihood of EF.

By considering clinical judgment of when to extubate as the test and noEF status as a positive outcome, we can estimate the positive predictive value (PPV) of both the clinician and the algorithm. Assuming a noEF prevalence of 85%, clinical PPV is 86.5% while algorithm PPV is 97.5%. We see the algorithm yields an improvement in the probability that a patient selected for extubation will successfully complete it without extubation failure. The tradeoff is that, by fixing specificity at 90%, using this algorithm in a clinical setting would result in 10% of noEF patients staying on the ventilator longer. Although modestly longer ventilation times may not pose a large risk to patients, longitudinal analysis should be done to see how much longer the patient would stay ventilated before this algorithm deems them ready for extubation.

There are a number of limitations that need to be considered when interpreting results from this study. While we verified the status of all patients in the study using the nursing notes, it is possible that our rule-based patient selection method could have missed patients in the MIMIC-II database. The number of subjects included in the final model, n=116 (101 noEF, 15 EF), is considerably less than the number of subjects that met inclusion criteria, n=179 (155 noEF, 24 EF), because patients that did not have all 6 features recorded had to be excluded from analysis. In general, the more variables included in any model, the more likely patients were dropped. We chose not to impute missing data to avoid biasing the mean. The interval chosen for data collection was 2 hours prior to the extubation events. This is an arbitrary choice, other intervals (e.g., 6, 12 and 24 hours), or combinations thereof, should be evaluated. Finally, we used bootstrap only for model building. Ideally, feature selection and model building should be performed together to reduce the risk of overtraining.

APPENDIX

RSBI = RR/V_T, where RSBI = rapid shallow breathing index, RR = respiratory rate, V_T = tidal volume.

WOB = $MAP*RR*V_T$, where WOB = work of breathing, MAP = mean invasive arterial blood pressure.

REFERENCES

[1] J. A. Lemons, C. R. Bauer, W. Oh, et al., "Very Low Birth Weight Outcomes of the National Institute of Child Health and Human Development Neonatal Research Network, January 1995 Through December 1996," *Pediatrics*, vol. 107, no. 1, pp. e1–e1, Jan. 2001.

[2] F. Hermeto, B. M. R. Martins, J. R. M. Ramos, et al., "Incidence and main risk factors associated with extubation failure in newborns with birth weight < 1,250 grams," *Jornal de Pediatria*, vol. 0, no. 0, Aug. 2009.

[3] M. Szymankiewicz, D. Vidyasagar, and J. Gadzinowski, "Predictors of successful extubation of preterm low-birth-weight infants with respiratory distress syndrome," *Pediatr. Crit. Care Med.*, vol. 6, no. 1, pp. 44–49, Jan. 2005.

[4] G. Dimitriou, S. Fouzas, A. Vervenioti, et al., "Prediction of extubation outcome in preterm infants by composite extubation indices," *Pediatr. Cri.t Care Med.*, vol. 12, no. 6, pp. e242–249, Nov. 2011.

[5] O. da Silva and D. Stevens, "Complications of airway management in very-low-birth-weight infants," *Biol. Neonate*, vol. 75, no. 1, pp. 40–45, 1999.

[6] M. C. Walsh, B. H. Morris, L. A. Wrage, et al., "Extremely low birthweight neonates with protracted ventilation: mortality and 18-month neurodevelopmental outcomes," *J. Pediatr.*, vol. 146, no. 6, pp. 798–804, Jun. 2005.

[7] M. Mueller, C. L. Wagner, D. J. Annibale, et al., "Predicting extubation outcome in preterm newborns: a comparison of neural networks with clinical expertise and statistical modeling," *Pediatr. Res.*, vol. 56, no. 1, pp. 11–18, Jul. 2004.

[8] C. M. Ennett, K. Lee, L. J. Eshelman, et al., "Predicting respiratory instability in the ICU," in *Proc. 30th Annual IEEE EMBS,* Vancouver, BC, Canada, 2008, pp. 2848–2851.

[9] M. Saeed, M. Villarroel, A. T. Reisner, et al., "Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database," *Crit. Care Med.*, vol. 39, no. 5, pp. 952–960, May 2011.

[10] R. R. Thiagarajan, S. L. Bratton, L. D. Martin, et al., "Predictors of successful extubation in children," *Am. J. Resp.Crit. Care Med.*, vol. 160, no. 5, pp. 1562–1566, 1999.

[11] F. Kanakoudi-Tsakalidou, F. Debonera, V. Drossou-Agakidou, et al., "Flow cytometric measurement of HLA-DR expression on circulating monocytes in healthy and sick neonates using monocyte negative selection," *Clin. Exp. Immunol.*, vol. 123, no. 3, pp. 402–407, Mar. 2001.

[12] A. G. Weinberg, C. R. Rosenfeld, B. L. Manroe, and R. Browne, "Neonatal blood cell count in health and disease. II. Values for lymphocytes, monocytes, and eosinophils," *J. Pediatr.*, vol. 106, no. 3, pp. 462–466, Mar. 1985.