Prediction Model for Tracheostomy Needs or Mortality in Neonates with Severe Bronchopulmonary Dysplasia

Jialin Liu

2023-12-15

Abstract

Background: Current prediction models, utilizing extensive databases, can estimate the probability of tracheostomy placement or mortality based on baseline demographics and clinical diagnoses. However, these analyses have yet to leverage detailed respiratory parameters and do not offer predictions at various postmenstrual ages (PMA). This study introduces a novel predictive model designed to address this gap.

Methods: We have developed and internally validated a logistic regression mixed-effects model aimed at predicting the need for tracheostomy in infants afflicted with severe bronchopulmonary dysplasia. Variable selection techniques, including the lasso and forward stepwise methods, were employed to identify the most pertinent variables for inclusion in the logistic model. Both models underwent rigorous internal validation, and their performance matrices were assessed.

Results: In the 36-week models, birth-related variables including birth weight, height, and head circumference showed significance in predicting adverse outcomes based on small p-values for estimated coefficients. Respiratory-related variables about ventilation support levels, fraction of inspired oxygen, peak inspiratory pressure, and exploratory pressure played significant roles in prediction. The 44-week models included critical variables of prenatal corticosteroids, complete prenatal steroids, and respiratory parameters measured at 36 weeks. Furthermore, parameters measured at 44 weeks, such as medication for pulmonary hypertension, peak inspiratory pressure, positive end exploratory pressure, ventilation support levels, and fraction of inspired oxygen, emerged as crucial factors in predicting the need for Tracheostomy placement or mortality.

Conclusion: These conclusions stem from a medical study evaluating the necessity for Tracheostomy or adverse outcomes, with the 44-week models generally exhibiting superior predictive performance compared to the 36-week models with a higher AUC of 91.5%.

Introduction

Bronchopulmonary dysplasia (BPD), also known as chronic lung disease, causes long-term breathing problems in newborn babies especially for those who are born prematurely. As the most common complication of prematurity, this disease affects estimated 10,000-15,000 infants each year in the United States, which is caused many multifactorial individual characteristics both from genetic and epigenetic aspects and substantial impact infant's susceptibility(Jensen and Schmidt 2014). Compared to the healthy lung tissue which can support normal breathing, the lungs with BPD have fewer and larger the tiny air sacs of the lung (alveoli), causing tissue destruction (fibrosis and metaplasia) within the lungs and usually showing signs of respiratory distress, such as breathing quickly and grunting(Sweet and Halliday 2005). This deficit in pulmonary vascular development has no cure, but it can be treated and most babies go on to live a long and healthy life. There are four levels of severity of BPD, in particular, 75% of babies with grade 3 BPD are always dependent on a ventilator at 36 weeks gestational age when they are discharged from the hospital. To allow babies to be hooked up to a ventilator for a long time, they needs a tracheostomy that is a surgical hole in the neck and tube inserted in the trachea to allow them in breath and out breath to lungs. Since some studies show that tracheostomy associated with improved outcomes within 4 months of age and a list of benefits to performing a tracheostomy, up to 12% babies with severe grade 3 BPD are required to have a tracheostomy(Akangire and Manimtim 2023). However, risks associated with a tracheostomy are also existing, which include increased risk of death compared to no tracheostomy, accidentally cannula obstruction or abscission, and increased rates of infection on skin, trachea and lungs.

Existing prediction models based on large databases can accurately estimate the likelihood of tracheostomy placement or death given baseline demographics and clinical diagnoses. However, these analyses have not used detailed respiratory parameters and have not provided prediction at different post-menstrual age (PMA). Accurate prediction the need for tracheostomy at early PMA would have implications for counseling of families and appropriate timing og tracheostomy placement, which is an active area of debate in severe BPD (sBPD). Motivated by deficiency in the previous work, models are designed to determine who really needs a tracheostomy, and when is an ideal time frame to refer a patient for tracheostomy. We will be using clinical data collected from multicenter, retrospective case-control study and recorded infants who are born at ≤ 32 weeks and their respiratory support at 36 and 44 weeks PMA. Outcomes of interest (tracheostomy or death) are recorded at the time point when they were discharged from hospitals. We developed and validated two prediction models and compared the performance of their predictability with respect to eventual needs for tracheostomy or death prior to discharge.

Methods

Study Population

This study analyzed data from a national data set of demographic, diagnostic, and respiratory parameters of infants with sBPD admitted to collaborative Neonatal intensive care units (NICUs) across multiple centers. The data consists of 999 participants who are born at ≤ 32 weeks and their corresponding 30 factors and outcomes of eventually healthy status at(or before) discharge. The rest of this section is devoted to describing summary statistics, procedures for preprocessing data, exploring any potential relationships between variables as well as missing values before building models with variable selection. We will conduct brief exploratory data analysis in terms of three aspects: birth and demographic variables, respiratory support variables, weight and tracheostomy placement at 36 and 44 weeks.

Of those 999 patients' records, some duplicated patient ID with information have been detected and removed for further analysis. To make sure completeness of outcome variables, we found two places of missingness in `Death` outcome variable. One of them should be corrected as "No" death since this patient was discharged from hospital at 43 weeks without tracheostomy placement, thus it is reasonable to replace this missing value with known outcome based on our basic speculation. The another missing value in `Death` cannot be deduced as missing value appeared in hospital discharge gestational age `hosp_dc_ga`, without this supporting information, we couldn't make assumption upon these bunch lack of data so that we removed this particular patient from our observational data. For conciseness and fewer number of models needed to construct, we combined two outcomes of interest, `Trach` and `Death`, into one final outcome about healthy status that refers to "Yes(1)" when babies neither had tracheostomy placement nor died at(or before) discharge, and conversely, "No(0)" represents adverse outcome of health, equally saying, babies either had tracheostomy or died.

For hospital discharge `hosp_dc_ga` variable, there exists some values that lie far much away from the main body of observations and may distort summaries of the distribution. For example, some cases showed hospital discharge gestational ages are greater than 300 weeks, which seems to be irrational in this study, since, based on boxplot and interquartile range, most often patients were discharged around 40-50 weeks. Therefore, we simply removed those few cases. Additionally, the data has been collected from multiple centers, we wanted to check if number of observations are evenly distributed and balanced. We found that center 20 and 21 only consisted with a total of 5 patients, which sample size are too small to conduct further analysis on these two centers. So we decided to remove `center` from further models due to usability of predictive models across different populations. In addition, levels in maternal race variable did not align with specified categories shown in the code book. Without explanation for this error, we removed this variable. For model simplicity, we considered respiratory and diagnostic related variables measured at 36 weeks only, and further analyses will be considered to add later time points and give a more comprehensive insight into an ideal time point to refer patients for tracheostomy.

Table 1 describes summary statistics for a part of participant birth and demographic variables, stratified by outcome of interest `adverse_outcome`, with the sample size for healthy group being 806 (noted as N = 806) and for group who either had tracheostomy placement or died prior to discharge (N = 182). Birth variables, which include birth weight (in g) `bw`, birth length (in centimeters) `blength`, and head circumference at birth `birth_hc`, show statistically significant differences between healthy and non-healthy groups since all p-values are greatly less than significance level ($\alpha = 0.05$). In particular, we could observe that those babies who had adverse outcome are high likely to have lower birth weights (mean of 757g) and smaller hear circumference (mean of 22.89cm), probably due to prematurity, than those who hadn't tracheostomy placement or died at discharge birth weights (mean of 816g) and larger head circumference (mean of 23.25cm). Delivery Method `del_method` was reported as categorical data with two methods: 1 represents for vaginal delivery and 2 represents for cesarean section. Higher percentage of babies who were delivered by cesarean section experienced adverse outcome than those who were delivered by vaginal method, with a significance difference between two groups based on a small p-value from Chi-squared test. One notable thing is that `any_surf` to record if the infant receive surfactant in the first 72 hours consisted with 44% of missingness in the original data set, thus it is crucial to carefully criticize whether the assumptions of multiple imputation are likely to hold and this variable cannot be reasonably imputed from the other available data by multiple imputation method.

Table 1: Participants Baseline Demographics Variables

Characteristic	Missing	0, N = 810	1, N = 183	p-value
bw	0 (0%)			<0.001
Mean (Maximum, SD)		817 (2,725, 285)	756 (2,615, 340)	
blength	77 (7.8%)			0.002
Mean (Maximum, SD)		33 (48, 4)	32 (45, 4)	
birth_hc	76 (7.7%)			0.010
Mean (Maximum, SD)		23.25 (36.00, 2.65)	22.88 (38.30, 3.29)	
del_method	3 (0.3%)			0.018
1		244 (30%)	39 (21%)	
2		564 (70%)	143 (79%)	
prenat_ster	33 (3.3%)			0.025
No		113 (14%)	13 (7.8%)	
Yes		680 (86%)	154 (92%)	
sga	15 (1.5%)			<0.001
Not SGA		658 (82%)	118 (66%)	
SGA		141 (18%)	61 (34%)	
any_surf	433 (44%)			0.092
No		89 (19%)	12 (12%)	
Yes		372 (81%)	87 (88%)	

[1] n (%)

[2] Wilcoxon rank sum test; Pearson's Chi-squared test

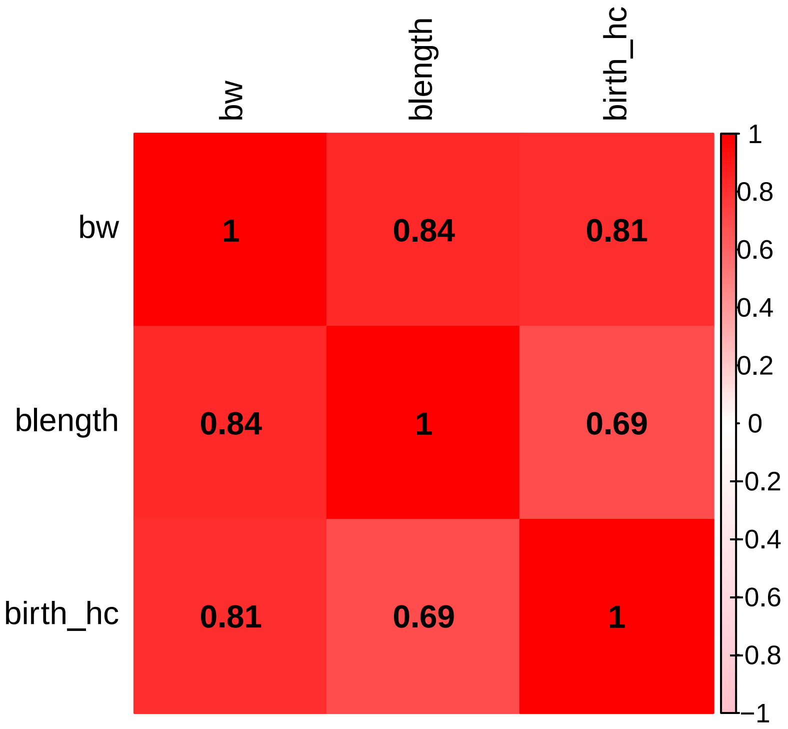
Table 2 provides data on respiratory-related variables measured at 36 weeks across different centers with an additional column for missing center name. For weight at 36 weeks `weight_today.36` variable, statistically significant p-value (0.021) is measured to assess a tendency of weights in at least one of the groups to be different than weights in at least one of the other group. The mean weight ranges from 2,073 grams in center 1 to 1,922 grams in center 5, with varying levels of missing data, potentially indicating variability in the mean weights at 36 weeks across multiple centers. For categorical ventilation support level at 36 weeks `ventilation_support_level.36` with three levels (0 means no respiratory support; 1 means non-invasive positive pressure; and 2 means invasive positive pressure) shows a significant difference across different centers ($p < 0.001$) by the Pearson's Chi-squared test. Even though Chi-squared test is the most commonly used test for assessing difference in distribution of a categorical variable between two or more independent groups (`center` here), the Chi-squared approximation to the distribution of the test statistic relies on the counts being roughly normally distributed. Because many of cell sizes are very small being less than 5 observations, the approximation may be poor. Instead, we run the Chi-squared test with computation of the simulated p-values, and resulted in a strong association between `center` and `ventilation_support_level.36` variables. The mean fraction of inspired oxygen at 36 weeks `inspired_oxygen.36` also varies significantly ($p < 0.001$) across centers. Some centers reported higher proportion of missingness, such as 43% in center 12. The peak inspiratory pressure (cmH_2O) at 36 weeks shows significant difference in at least one center ($p < 0.001$) compared to others, and a still high percentage of missingness observed in center 12, as well as in positive and exploratory pressure (cmH_2O) at 36 weeks `peep_cm_h2o.36` variable. For medication for pulmonary hypertension at 36 weeks `med_ph.36`, we could reject the null hypothesis and conclude that there is a strong association between center and medication for pulmonary hypertension given a significantly small p-value. Lastly, the mean hospital discharge gestational age `hosp_dc_ga` appears to have significant differences across centers, with a 100% missing rate in center 4 and 98% of missingness in center 1. There is notable variability of missingness in these respiratory variables at 36 weeks across centers, for example, center 12 consisted with almost half missing values in variables related to inspiratory and exploratory pressure and inspired oxygen, and some centers did not record hospital discharge gestational age at all. These missing data rates are concerning, which could impact the reliability of the statistical analyses.

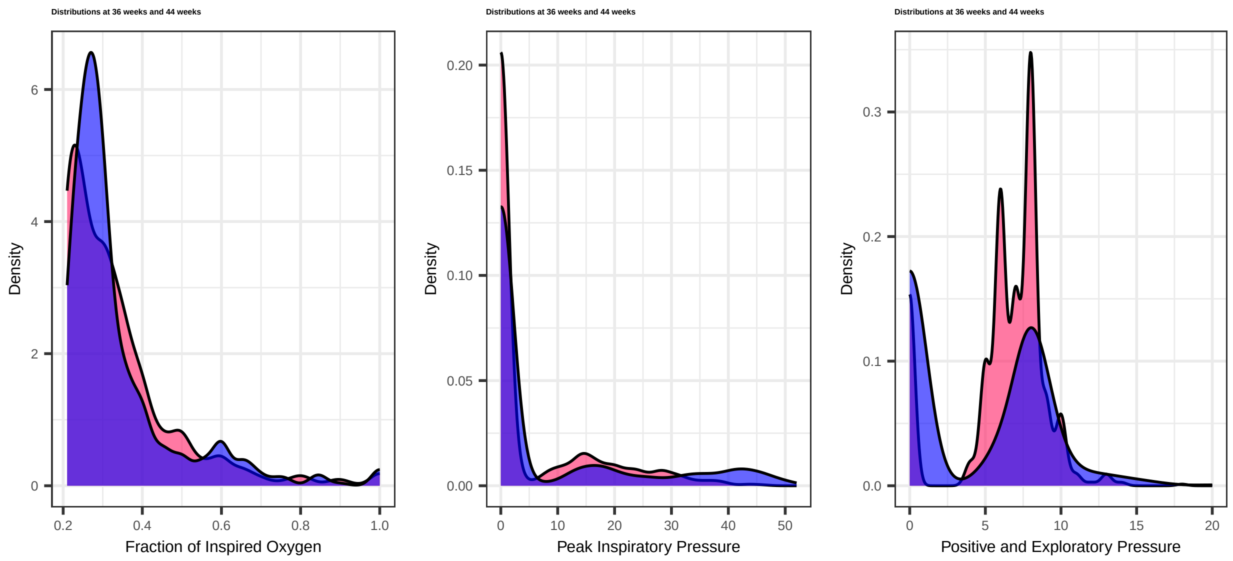
The below heatmap shows the correlation coefficients between three different birth variables: birth weight (`bw`), birth length (`blength`), and head circumference at birth (`birth_hc`). The colors range from red to white, where deep red indicates a stronger positive correlation, and white would indicate no correlation. The scale goes from -1 to 1, where 1 means a perfect positive correlation, whereas -1 would mean a perfect negative correlation. All the variables show strong positive correlations with each other (0.7 or higher), meaning as one variable increases, the others tend to increase as well. Below the heatmap are three distribution plots, which represent respiratory variables for newborns at 36 weeks (pink regions) and 44 weeks (blue regions). The first plot shows the distribution of fraction of inspired oxygen", which is typically a measure of the oxygen concentration in the air mixture being delivered to a patient. We can observe that the distribution of inspired oxygen at 44 weeks are generally higher than that of variable at 36 weeks, especially with a higher peak at 44 weeks. The second plot about the distribution of peak inspiratory pressure, which is a measure used in ventilator settings during mechanical ventilation, indicates a higher density in lower inspiratory pressure settings at 36 weeks. The third plot appears to be positive and exploratory pressure, but it looks to be fluctuated at 36 weeks with an overall higher density than the measure at 44 weeks.

Table 2: Summary Statistics of Respiratory Variables at 36 Weeks by Center

Characteristic	**1**, N = 55	**2**, N = 630	**3**, N = 56	**4**, N = 59	**5**, N = 40	**7**, N = 32	**12**, N = 68	**16**, N = 38	**20**, N = 0	**21**, N = 0	**(Missing)**, N = 10	**p-value**
weight__today.36												0.021
Mean (SD)	2,073 (440)	2,135 (408)	2,113 (426)	2,120 (339)	1,922 (401)	2,169 (409)	2,044 (484)	2,220 (409)	NA (NA)	NA (NA)	1,992 (618)	
N missing (% missing)	12 (22%)	36 (5.7%)	3 (5.4%)	6 (10%)	0 (0%)	1 (3.1%)	29 (43%)	0 (0%)	0 (NA%)	0 (NA%)	5 (50%)	
ventilation__support__level.36												<0.001
0	7 (13%)	50 (8.1%)	5 (9.1%)	8 (14%)	0 (0%)	22 (69%)	1 (2.0%)	22 (58%)	0 (NA%)	0 (NA%)	1 (11%)	
1	19 (35%)	425 (68%)	34 (62%)	34 (58%)	31 (78%)	8 (25%)	16 (33%)	14 (37%)	0 (NA%)	0 (NA%)	3 (33%)	
2	29 (53%)	146 (24%)	16 (29%)	17 (29%)	9 (22%)	2 (6.2%)	32 (65%)	2 (5.3%)	0 (NA%)	0 (NA%)	5 (56%)	
inspired__oxygen.36												<0.001
Mean (SD)	0.43 (0.21)	0.32 (0.14)	0.31 (0.09)	0.40 (0.12)	0.36 (0.13)	0.36 (0.10)	0.40 (0.19)	0.35 (0.11)	NA (NA)	NA (NA)	0.40 (0.06)	
N missing (% missing)	14 (25%)	36 (5.7%)	2 (3.6%)	3 (5.1%)	0 (0%)	1 (3.1%)	29 (43%)	0 (0%)	0 (NA%)	0 (NA%)	4 (40%)	
p__delta.36												<0.001
Mean (SD)	7 (8)	5 (11)	7 (8)	5 (6)	4 (7)	0 (1)	9 (7)	1 (5)	NA (NA)	NA (NA)	6 (8)	
N missing (% missing)	17 (31%)	39 (6.2%)	7 (12%)	15 (25%)	13 (32%)	1 (3.1%)	32 (47%)	0 (0%)	0 (NA%)	0 (NA%)	3 (30%)	
peep__cm__h2o__modified.36												<0.001
Mean (SD)	7 (4)	6 (2)	8 (3)	6 (3)	9 (2)	2 (3)	7 (2)	3 (4)	NA (NA)	NA (NA)	7 (5)	
N missing (% missing)	18 (33%)	41 (6.5%)	11 (20%)	6 (10%)	0 (0%)	1 (3.1%)	34 (50%)	0 (0%)	0 (NA%)	0 (NA%)	6 (60%)	
med__ph.36												<0.001
0	42 (76%)	596 (96%)	52 (95%)	49 (83%)	37 (92%)	30 (94%)	45 (92%)	34 (89%)	0 (NA%)	0 (NA%)	9 (100%)	
1	13 (24%)	25 (4.0%)	3 (5.5%)	10 (17%)	3 (7.5%)	2 (6.2%)	4 (8.2%)	4 (11%)	0 (NA%)	0 (NA%)	0 (0%)	
hosp__dc__ga												<0.001
Mean (SD)	60 (NA)	53 (18)	46 (21)	NA (NA)	54 (18)	45 (7)	54 (14)	41 (3)	NA (NA)	NA (NA)	NA (NA)	
N missing (% missing)	54 (98%)	0 (0%)	0 (0%)	59 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (NA%)	0 (NA%)	10 (100%)	

Correlation Matrix of Birth Variables





Multiple Imputation

Multiple imputation is applied to address missing values by creating 5 complete datasets. This method is trying to handle with each missing entry by estimating multiple reliable values such as regression models, running analysis across those completed dataset, aggregating all previous results, and analyzing how far they spread out in terms of standard deviations and confidence intervals. We constructed train-test sets and fitted the imputation model on the train data and applied the model to the test set. Given 5 complete train datasets, we will run lasso and forward stepwise regression for variables selection and use combined 5 test data as a validation dataset to assess performance of each model.

Variables Selection for the 36-week Model

Table 3: Coefficients of Lasso Regression Method for 36-week

	Lasso1	Lasso2
(Intercept)	-4.009	-4.009
ga	0.049	0.049
birth_hc	0.000	0.000
del_method2	0.000	0.000
prenat_sterYes	0.220	0.220
com_prenat_sterYes	0.115	0.115
sgaSGA	0.184	0.184
weight_today.36	-0.001	-0.001
ventilation_support_level.362	1.536	1.536
inspired_oxygen.36	4.313	4.313
p_delta.36	0.005	0.000
peep_cm_h2o_modified.36	0.020	0.020
med_ph.361	0.071	0.000

Table 4: Coefficients of Forward Stepwise Selection Method for 36-week

	Forward.1	Forward.2	Forward.3	Forward.4	Forward.5	Average
(Intercept)	-5.413	-2.767	-2.093	-2.425	-6.912	-3.922
bw	0.000	0.001	0.001	0.001	0.000	0.001
ga	0.000	0.000	0.000	0.000	0.110	0.022
blength	0.000	0.000	0.000	0.000	0.000	0.000
birth_hc	0.095	0.000	0.000	0.000	0.000	0.019
del_method2	0.000	0.000	0.000	0.000	0.000	0.000
prenat_sterYes	0.000	0.000	0.981	0.674	0.871	0.505
com_prenat_sterYes	0.404	0.448	0.000	0.000	0.000	0.170
mat_chorioYes	0.000	0.000	0.000	0.000	0.000	0.000
genderMale	0.000	0.000	0.000	0.000	0.000	0.000
sgaSGA	0.531	0.433	0.000	0.000	0.000	0.193
weight_today.36	-0.001	-0.001	-0.002	-0.001	-0.001	-0.001
ventilation_support_level.361	0.000	0.000	0.000	0.000	0.000	0.000
ventilation_support_level.362	1.706	2.522	1.245	1.916	2.139	1.906
inspired_oxygen.36	4.869	5.397	3.867	5.002	6.153	5.057
p_delta.36	0.000	-0.034	0.039	0.000	-0.030	-0.005
peep_cm_h2o_modified.36	0.073	0.000	0.000	0.000	0.083	0.031
med_ph.361	0.000	0.000	0.000	0.000	0.000	0.000

After multiple imputation method, we have 5 imputed datasets for which we do cross-validation in each imputed training dataset. Cross-validation helps with overfitting issues and with generalizability of the lasso model. Then since we will have for each fold in cross-validation results for different lambdas, we choose the lambda that has the lowest model error. This means that we will have five sets of estimated coefficients (one for each imputed data set). We considered two possible solutions to generate final lasso model: The first method **Lasso 1** is simply averaging over all 5 sets of coefficients to obtain the final lasso model. The second method **Lasso 2** is counting the number of times each variable being selected, if more than or equal to 3 times the variable was not selected, we will force to remove the variable from the final lasso model. In other words, if some variables occasionally were not selected, we wanted to neglect occasional situations and averaging over all estimated coefficients from 5 imputation datasets. In order to include transformations or interactions in the model, it is needed to have an idea of which interactions or transformations need to be included based on explanatory analysis or prior professional knowledge before applying the cross-validation to the imputed datasets. So first, after we have the imputed data, we fit the lasso model and evaluate variables' significance by running summary of model after cross-validation. Table 3 provides us with a list of final estimated coefficients under lasso models of 10-fold cross-validation stratified by outcome in the train imputation datasets.

Except for lasso regression models, to preserve generalizability and overfitting, we remove center, include main effects of covariates, and perform forward stepwise selection with cross-validation for each imputed data set. Starting with the empty model and sequentially adding predictors to the model, one at a time, and choosing the best predictor at each step based on a criterion like AIC and BIC, but it can be seen as a "locally optimal", instead of globally optimal in the sense of the best subset. Table 4 provides coefficients for each imputed data set and averaging over 5 coefficients together to obtain the final forward stepwise model. To compare the estimated coefficients using different variable selection or shrinkage methods we found that intercept values vary a lot across different methods, with lasso method having the most negative value (-4.009), which means the log odds of adverse outcome while all other variables are held at zero. For variables with non-zero coefficients, a positive coefficient indicates a positive association between the variable and the odds of outcome. In forward selection model, `pre_nat_sterYes` has a positive average coefficient (0.505), indicating that having a prenatal corticosteroids are 0.505 times the odds of adverse outcome compared to those who did not have prenatal corticosteroids. Conversely, weight at 36 weeks (-0.001) is negatively associated with the odds of adverse outcome, even though it is least likely to have adverse outcome compared to other covariates adjusting for remaining variables. For those variables with zero coefficients all 5 times represents the least importance to be include in predicting adverse outcome of patients.

Variables Selection for the 44-week Model

We explore important variables when data on 44 weeks are available and investigate mixed-effect models guided by Lasso and forward stepwise methods to estimate coefficients for important variables and random effects for center. Table 5 presents the coefficients for two Lasso regression models. Lasso (Least Absolute Shrinkage and Selection Operator) regression is a type of linear regression that uses shrinkage and is particularly useful when we want to automate certain parts of variables included in models. The intercept term (-5.871) represents the odds of adverse outcome when all predictors are zero. The coefficients represent the association between each predictor and the outcome variable. A coefficient of 0.000 would suggest no association under the Lasso regression constraints. The `inspired_oxygen` related variables have a coefficient of 2.433 at 36 weeks and a coefficient of 1.014 at 44 weeks, suggesting for each unit increase in fraction of inspired oxygen at 36 and 44 weeks is associated with 2.433 or 1.014 times the odds of adverse outcome adjusting for other covariates. The fact that some coefficients are exactly zero is a feature of Lasso regression, which help us determine variables in the mixed-effect model.

Table 5: Coefficients of Lasso Regression Method for 44-week

	Lasso1	Lasso2
(Intercept)	-5.871	-5.871
bw	0.001	0.001
blength	0.000	0.000
birth_hc	0.007	0.000
del_method2	0.348	0.348
prenat_sterYes	0.771	0.771
com_prenat_sterYes	0.073	0.000
mat_chorioYes	0.023	0.000
genderMale	0.023	0.000
sgaSGA	0.205	0.205
ventilation_support_level.362	1.184	1.184
inspired_oxygen.36	2.433	2.433
p_delta.36	0.007	0.007
peep_cm_h2o_modified.36	0.060	0.060
med_ph.361	0.038	0.000
weight_today.44	0.000	0.000
ventilation_support_level_modified.44	0.699	0.699
inspired_oxygen.44	1.014	1.014
p_delta.44	0.009	0.009
peep_cm_h2o_modified.44	0.030	0.030
med_ph.44	1.351	1.351

Forward Stepwise Selection is a type of model building that begins with no variables in the model, tests the addition of each variable using a chosen model fit criterion, adding the variable (if any) whose inclusion gives the most statistically significant improvement of the fit, and repeats this process until none improves the model to a significant extent. The process results in different coefficients across the models as variables are added one at a time based on their statistical significance. Table 6 shows the results of a forward stepwise selection regression analysis over five imputed training sets (Forward.1 to Forward.5) and their average coefficients. These two methods are trying to determine the factors that affect adverse outcome when records on both 36 weeks and 44 weeks are available.

Table 6: Coefficients of Forward Stepwise Selection Method for 44-week

	Forward.1	Forward.2	Forward.3	Forward.4	Forward.5	Average
(Intercept)	-2.118	-6.296	-6.329	-9.156	-8.012	-6.382
bw	0.002	0.001	0.001	0.001	0.001	0.001
ga	-0.215	0.000	0.000	0.000	0.000	-0.043
blength	0.000	0.000	0.000	0.000	0.000	0.000
birth_hc	0.000	0.000	0.000	0.000	0.000	0.000
del_method2	0.655	0.800	0.616	0.464	0.000	0.507
prenat_sterYes	1.599	1.297	1.154	1.257	1.366	1.335
com_prenat_sterYes	0.000	0.000	0.000	0.000	0.000	0.000
mat_chorioYes	0.000	0.000	0.000	0.000	0.000	0.000
genderMale	0.000	0.000	0.000	0.000	0.000	0.000
sgaSGA	0.656	0.000	0.000	0.998	0.722	0.475
weight_today.36	0.000	0.000	0.000	0.000	0.000	0.000
ventilation_support_level.361	0.000	0.000	0.000	0.000	0.000	0.000
ventilation_support_level.362	1.595	1.842	1.416	1.121	1.428	1.480
inspired_oxygen.36	1.325	3.123	3.541	4.326	2.940	3.051
p_delta.36	0.000	-0.033	0.000	0.000	0.000	-0.007
peep_cm_h2o_modified.36	0.107	0.121	0.121	0.131	0.094	0.115
med_ph.361	0.000	0.000	0.000	0.000	0.000	0.000
weight_today.44	-0.001	-0.001	-0.001	0.000	0.000	0.000
ventilation_support_level_modified.44	0.000	0.878	1.148	0.898	1.117	0.808
inspired_oxygen.44	1.697	0.000	0.000	1.446	0.000	0.628
p_delta.44	0.029	0.025	0.000	0.000	0.000	0.011
peep_cm_h2o_modified.44	0.151	0.000	0.000	0.000	0.000	0.030
med_ph.44	1.564	2.057	1.572	1.362	1.753	1.662

Internal Validation

Mixed-effects Models for 36-week data

Given important fixed-effect variables selected from Lasso 2 and forward stepwise methods, we fit the mixed-effects models for 36-weeks, which is considered `center` as a random effect. We choose significant variables and center and fit the model using the `glmer()` function in the combined 5 imputed training set for 36-weeks only and finally derive new estimated coefficients of variables in the mixed-effect models under Lasso 2 and forward stepwise methods, respectively. The below table depicts estimated center-specific intercepts in two mix-effect models guided by Lasso and forward stepwise selections. We then apply new mix-effect models to predict outcome in the internal validation set derived from a combination of 5 imputed test sets.

Table 7: Estimated Random Effects of Center in the Mix Model for 36-weeks by Lasso and Forward Stepwise Methods

	Lasso	Forward

1	1.1488697	1.1905300
2	-0.2019520	-0.1817007
3	-1.4882200	-1.4650921
4	-0.3530425	-0.3168509
5	0.0450642	0.0181617
7	-0.9872873	-0.9198745
12	1.5746558	1.6055910
16	-0.7930766	-0.7523810
20	-0.6010226	-0.5686936
21	1.9191463	1.6357369

We validated the mixed-effect models for 36-weeks by assessing their performance in the internal validation dataset obtained from multiple imputation method. Then we choose the best thresholds obtained from ROC curves to cutoff the binary predicted adverse outcome. Then we use AUC, specificity, sensitivity, and Brier scores to evaluate model performance in validation set. According to scores and graphical means, we can look at how well our model differentiates between the two classes. By comparison, the lasso model on the test data (AUC = 88.7%) explains the outcome of adverse outcome is slightly better than forward stepwise methods. For both models, the AUC is quite high (0.887 for Lasso and 0.882 for forward), indicating good model performance. Sensitivity, also known as the true positive rate, measures the proportion of actual positives that are correctly identified. The Lasso model has a sensitivity of 0.830, and the Forward model has a slightly lower sensitivity of 0.826. Specificity, also known as the true negative rate, measures the proportion of actual negatives that are correctly identified. The Lasso model shows a specificity of 0.827, and the forward model has a slightly lower specificity of 0.810. The Brier score measures the accuracy of probabilistic predictions. It is a score of 0 for a perfect model and 1 for a model that performs no better than random chance. Here, the Lasso model has a Brier score of 0.140 and the forward model has a slightly higher Brier score of 0.144, indicating the Lasso is slightly more accurate in its predictions when birth and 36-weeks data are considered in the model.

Table 8: Performance Matrices of Mix-effects Models for 36 weeks in Test Set

	Sensitivity	Specificity	AUC	Brier
Lasso	0.830	0.827	0.887	0.140
Forward	0.826	0.810	0.882	0.144

Mixed-effects Models for 44-week data

Table 9: Estimated Random Effects of Center in the Mix Model for 44-weeks by Lasso and Forward Stepwise Methods

	Lasso	Forward
1	0.9473799	0.9360729
2	-0.2628819	-0.3160842
3	-0.8853526	-0.9719838
4	0.6042586	0.7364965
5	0.7408230	0.8271650
7	-0.6134596	-0.6162574
12	1.8698833	1.7600271
16	-0.4354158	-0.4590019
20	-1.3866205	-1.3159962
21	-0.2537093	-0.2570907

Table 10 shows performance matrices of mix-effects models for 44 weeks in the combined test set. As for sensitivity, the Lasso model maintains the same sensitivity as at 36 weeks (0.830), but the forward model sees a significant increase to 0.889. Also, there is a notable increase in specificity for the Lasso model to 0.866, indicating better performance in correctly identifying true negatives at 44 weeks compared to 36 weeks. However, the forward model sees a decrease in specificity to 0.823. The AUC values have increased for both models, with the Lasso model at 0.913 and the forward model at 0.915. This suggests that both models are better at distinguishing between the classes at 44 weeks than at 36 weeks. However, the Brier scores have slightly increased, indicating a reduction in predictive accuracy, with Lasso model's Brier score of 0.151 and the forward model's score of 0.145.

Table 10: Performance Matrices of Mix-effects Models for 44 weeks in Test Set

	Sensitivity	Specificity	AUC	Brier
Lasso	0.830	0.866	0.913	0.151
Forward	0.889	0.823	0.915	0.145

Conclusions and Limitations

In summary, the forward model exhibited a notable enhancement in sensitivity from 36 to 44 weeks, signifying its improved ability to accurately identify true positives in the 44-week mixed-effect model. Conversely, the Lasso model displayed an improved level of specificity during this transition, indicating its enhanced capability to correctly identify true negatives. Both mixed-effects models demonstrated a higher AUC at 44 weeks, reflecting an overall improved classification performance when incorporating covariates measured at this stage. However, it's important to note that the Brier score increased for both models, indicating a slight reduction in the accuracy of probabilistic predictions as patients aged from 36 to 44 weeks.

These conclusions stem from a medical study evaluating the necessity for Tracheostomy or adverse outcomes, with the 44-week models generally exhibiting superior predictive performance compared to the 36-week models. Specifically, in the 36-week models, birth-related variables such as birth weight, height, and head circumference showed significance in predicting adverse outcomes based on small p-values ($\leq$ 0.05) for estimated coefficients. Additionally, respiratory-related variables such as ventilation support levels, fraction of inspired oxygen, peak inspiratory pressure, and exploratory pressure played significant roles in prediction. In contrast, the 44-week models included critical variables such as prenatal corticosteroids, complete prenatal steroids, and respiratory parameters measured at 36 weeks, including inspired oxygen and ventilation support levels. Furthermore, parameters measured at 44 weeks, such as medication for pulmonary hypertension, peak inspiratory pressure, positive end exploratory pressure, ventilation support levels, and fraction of inspired oxygen, emerged as crucial factors in predicting the need for Tracheostomy placement or mortality.

However, it's important to acknowledge the limitations of our findings. First, the multiple imputation method, while providing more unbiased estimates than single imputation, may not be highly reliable due to potential violations of assumptions, especially in the presence of missing not-at-random (MNAR) data and limitations in sample size. Secondly, to ensure against overfitting and enhance generalizability, we should consider the myriad potential interaction terms and transformations. Nevertheless, it's worth noting that the forward stepwise method has its own limitations, as it can only be regarded as "locally optimal" rather than globally optimal in terms of selecting the best subset of variables. Therefore, future improvement analyses should incorporate the best subset method and assess predictive accuracy among these three variable selection methods for a more comprehensive evaluation.

References

Akangire, Gangaram, and Winston Manimtim. 2023. “Tracheostomy in Infants with Severe Bronchopulmonary Dysplasia: A Review.” *Frontiers in Pediatrics* 10 (January). https://doi.org/10.3389/fped.2022.1066367.

Jensen, Erik A., and Barbara Schmidt. 2014. “Epidemiology of Bronchopulmonary Dysplasia.” *Birth Defects Research Part A: Clinical and Molecular Teratology* 100 (3): 145–57. https://doi.org/10.1002/bdra.23235.

Sweet, David G, and Henry L Halliday. 2005. “Modeling and Remodeling of the Lung in Neonatal Chronic Lung Disease: Implications for Therapy.” *Treatments in Respiratory Medicine* 4 (5): 347–59. https://doi.org/10.2165/00151829-200504050-00006.