Prediction of Tracheostomy or Death in Neonates with Severe Bronchopulmonary Dysplasia: Based on Respiratory Parameters at 36-week Postmenstrual Age

Project 2: Regression Analysis*

Haiyue Song

2023-12-15

Abstract

Background: Tracheostomy placement in infants suffering from severe bronchopulmonary dysplasia presents a clinical challenge due to the absence of clear indication criteria and optimal timing. Previous studies predicting the likelihood of tracheostomy placement or mortality have utilized only baseline demographic information and clinical diagnoses without considering comprehensive respiratory parameters at various postmenstrual ages. This study aims to address this research gap by incorporating respiratory parameters at 36-week postmenstrual ages to predict tracheostomy or death in infants with severe bronchopulmonary dysplasia (Paul 2023b).

Methods: The data is a national data set of birth, demographic, diagnostic, and respiratory parameters of infants with severe bronchopulmonary dysplasia admitted to collaborative NICUs and with known respiratory support parameters at 36-week postmenstrual ages. Variables with a high proportion of missing data were excluded, and the remaining were imputed using multiple imputation techniques. Fixed-effect and mixed -effect logistic regression models were developed to predict the likelihood of tracheostomy or death, applying two variable selection methods: best subset selection and Lasso regularization.

Results: A total of 985 records were included in the final analysis, with 80% (N=794) for training and 20% (N=191) for testing. The final models, including one fixed-effect model for usage within existing facilities and another mixed-effect model for broader application, demonstrated excellent performance. The inclusion of diagnostic and respiratory parameters at 36-week PMA proved valuable for predicting the likelihood of tracheostomy or mortality. The fixed-effect model achieved an area under the curve of 0.9106, with sensitivity of 0.8485, specificity of 0.8582, accuracy of 0.8565. The mixed-effect model achieved an area under the curve of 0.9079, , with sensitivity of 0.8303, specificity of 0.8620, accuracy of 0.8565.

Conclusions: The models incorporating respiratory parameters at 36-week postmenstrual ages provide an earlier and more accurate prediction of the need for tracheostomy in infants with severe BPD at early postmenstrual ages. This has potential implications for counseling families, optimizing tracheostomy placement, and improving infant growth outcomes.

Keywords: Tracheostomy, Regression Analysis, Respiratory Parameters

1. Introduction

Tracheostomy placement in infants suffering from severe bronchopulmonary dysplasia (sBPD) presents a clinical challenge, compounded by the absence of clear indication criteria and optimal timing. Research indicates potential growth benefits from performing tracheostomies earlier in infants (Zhang et al. 2018). Prior studies

*This project is a collaboration with Dr. Chris Schmid in the Department of Biostatistics, Brown University School of Public Health. The instructor is Dr. Alice Paul from Department of Biostatistics, Brown University School of Public Health.

conducted on large databases have demonstrated that it is possible to make accurate predictions regarding the probability of tracheostomy placement or mortality by utilizing baseline demographic information and clinical diagnoses. However, these analyses have not incorporated comprehensive respiratory parameters and have not provided predictions for various postmenstrual ages (PMA) (Paul 2023b).

This gap indicates the need for a refined predictive model that can incorporate comprehensive respiratory parameters for prediction of tracheostomy at an early PMA. Our approach involves developing a logistic regression model with proper variable selection methods utilizing a national dataset of demographic, diagnostic, and respiratory parameters of infants admitted to NICUs. The success of our final models, including one fixed-effect model for usage within existing facilities and another mixed-effect model for broader application, marks a significant advancement in predicting tracheostomy among families with sBPD infants.

2. Methods

2.1. Study Setting and Population

The data is a national data set of birth, demographic, diagnostic, and respiratory parameters of infants with sBPD admitted to collaborative NICUs and with known respiratory support parameters at 36-week and 44-week PMA (McKinney and Levin 2023; Paul 2023b). The data set includes infants born at most 32 weeks PMA and was collected from retrospective case-control study across 10 centers of the sBPD collaborative. The outcome of interest is tracheostomy placing prior to discharge or death; as the number of cases for each is very small, we decided to consider the combined negative outcomes.

Data are collected at four time points: birth, 36-week PMA, 44-week PMA and discharge. At the time of birth, the birth weight, the gestational age, the birth length and head circumference, the delivery method, the maternal race, the maternal ethnicity, the infant gender, then use of prenetal steriods, the use of maternal chorioamnionitis, if the infant was small for gestational age, if the infant received any surfactant within 72 hours after birth were collected. At 36-week PMA and 44-week PMA, the weight, ventilation support level, fraction of inspired oxygen needed, peak inspiratory pressure, positive and exploratory pressure, medication for pulmonary hypertension were collected separately. Prior to the time of discharge, outcomes of tracheostomy placing (trach) or death were collected; the gestational age was collected at discharge as well.

From an initial 999 observations, duplicates were removed, leaving 996 entries. Two records with missing death data from different centers were excluded as it is reasonable to assume they are missing at random based on exploratory analysis. Missingness of center was addressed by complete record ID data consisting center information. A binary outcome variable was created to denote trach or death occurrences. Maternal race was omitted from the analysis due to coding clarity issues, and some complete prenatal steroid data gaps were filled based on negative corticosteroid responses. Surfactant data within the first 72 hours post-birth, missing in 43.46% of cases, was excluded from model fitting as it showed no significant association with the outcomes ($p = 0.123$ from Pearson's chi-square test), indicating less likely contribution to outcome prediction. Parameters from 44-week PMA were also excluded due to their high missingness and the NHLBI's (2018) BPD definition reliance on 36-week PMA parameters (McKinney and Levin 2023). Centers (i.e., center 20 and center 21) with insufficient data were dropped due to limitation of statistical power with few sample size. Records with implausible gestational ages at discharge (i.e., greater than 300 weeks, which are inpatients for 6 years) were considered outliers and removed. The other missingness is kept and will be address by

Table 1: Summary Statistics for Data Collected at Birth by Outcome

Variable	Overall, N = 985	No Trach or Death, N = 804	Trach or Death, N = 181	p-value
Maternal Ethnicity				0.3
Hispanic or Latino	70 / 929 (7.5%)	61 / 761 (8.0%)	9 / 168 (5.4%)	
Not Hispanic or Latino	859 / 929 (92%)	700 / 761 (92%)	159 / 168 (95%)	
(Missing)	56	43	13	
Birth Weight (g)				<0.001
Mean (SD)	805.78 (296.33)	816.21 (284.68)	759.47 (340.35)	
Gestational Age				0.5
Mean (SD)	25.78 (2.14)	25.76 (2.14)	25.87 (2.14)	
Birth Length (cm)				0.003
Mean (SD)	32.50 (3.80)	32.66 (3.72)	31.68 (4.13)	
(Missing)	77	44	33	
Birth Head Circumference (cm)				0.018
Mean (SD)	23.20 (2.75)	23.25 (2.65)	22.95 (3.21)	
(Missing)	76	43	33	
Delivery Method				0.031
Vaginal delivery	280 / 982 (29%)	241 / 802 (30%)	39 / 180 (22%)	
Cesarean section	702 / 982 (71%)	561 / 802 (70%)	141 / 180 (78%)	
(Missing)	3	2	1	
Prenatal Corticosteroids	828 / 952 (87%)	675 / 787 (86%)	153 / 165 (93%)	0.022
(Missing)	33	17	16	
Complete Prenatal Steroids	606 / 917 (66%)	497 / 764 (65%)	109 / 153 (71%)	0.2
(Missing)	68	40	28	
Maternal Chorioamnionitis	158 / 925 (17%)	130 / 760 (17%)	28 / 165 (17%)	>0.9
(Missing)	60	44	16	
Gender				0.6
Female	404 / 981 (41%)	333 / 800 (42%)	71 / 181 (39%)	
Male	577 / 981 (59%)	467 / 800 (58%)	110 / 181 (61%)	
(Missing)	4	4	0	
Small for Gestational Age	199 / 970 (21%)	139 / 793 (18%)	60 / 177 (34%)	<0.001
(Missing)	15	11	4	
Hospital Discharge Gestational Age				<0.001
Mean (SD)	51.91 (17.75)	47.82 (10.16)	73.74 (29.77)	
(Missing)	123	78	45	

[1] n / N (%)

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

multiple imputation in Section 2.3. We created dummy variables for categorical data and use the continuous variables themselves.

Table 1 using the *gtsummary* package (Baraldi and Enders 2010) provides a comparative summary of the baseline characteristics collected at birth from these 985 neonates, grouped by the outcome of interest - trach or death (N = 181) versus neither (N = 804). There is no signficant difference of maternal ethnicity ($p = 0.3$), gestational age ($p = 0.5$), complete prenatal steroids ($p = 0.2$), maternal chorioamnionitis ($p > 0.9$), gender between the groups of two outcomes ($p = 0.6$). There is a statistically significant difference in birth weight that infants with neither trach nor death having more average weight (816.21g) than those with either trach/death (759.47g, $p < 0.001$). This indicates that lower birth weight may be a risk factor for adverse outcomes. Similarly, birth length showed a significant difference ($p = 0.003$), with longer infants less likely to have trach/death. Head circumference also followed this trend (p = 0.018), further supporting the importance of infant growth metrics as predictors of the outcome. Delivery method displayed significant variation, with vaginal deliveries being more prevalent in the group with no trach or death ($p = 0.031$). The usage of prenatal corticosteroids was significantly higher in the group that experienced trach or death

Table 2: Summary Statistics for Data Collected at 36-week PMA by Outcome

Variable	Overall, N = 985	No Trach or Death, N = 804	Trach or Death, N = 181	p-value
Weight at 36 weeks (g)				<0.001
Mean (SD)	2,120.85 (413.14)	2,141.49 (392.92)	1,989.12 (506.55)	
Median (Range)	2,130.00 (710.00, 3,710.00)	2,150.00 (710.00, 3,710.00)	2,000.00 (915.00, 3,500.00)	
(Missing)	92	32	60	
Ventilation Support Level at 36 weeks				<0.001
No respiratory support or supplemental oxygen	115 / 955 (12%)	108 / 795 (14%)	7 / 160 (4.4%)	
Non-invasive positive pressure	584 / 955 (61%)	549 / 795 (69%)	35 / 160 (22%)	
Invasive positive pressure	256 / 955 (27%)	138 / 795 (17%)	118 / 160 (74%)	
(Missing)	30	9	21	
Fraction of Inspired Oxygen at 36 weeks				<0.001
Mean (SD)	0.34 (0.15)	0.31 (0.12)	0.49 (0.21)	
Median (Range)	0.30 (0.21, 1.00)	0.29 (0.21, 1.00)	0.45 (0.21, 1.00)	
(Missing)	89	30	59	
Peak Inspiratory Pressure (cmH2O) at 36 weeks				<0.001
Mean (SD)	5.26 (9.76)	3.85 (8.43)	15.50 (12.38)	
Median (Range)	0.00 (0.00, 46.00)	0.00 (0.00, 39.00)	14.00 (0.00, 46.00)	
(Missing)	125	48	77	
Positive and exploratory pressure (cm H2O) at 36 weeks				<0.001
Mean (SD)	6.34 (2.91)	6.18 (2.91)	7.42 (2.62)	
Median (Range)	7.00 (0.00, 18.00)	7.00 (0.00, 18.00)	8.00 (0.00, 14.00)	
(Missing)	115	47	68	
Medication for Pulmonary Hypertension at 36 weeks	64 / 955 (6.7%)	32 / 795 (4.0%)	32 / 160 (20%)	<0.001
(Missing)	30	9	21	

[1] n / N (%)

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

($p = 0.022$). There is significantly higher proportion of infants classified as small for gestational age (SGA) in the trach/death outcome group versus no trach/death group (34% vs. 18%, $p < 0.001$), highlighting SGA as a potential indicator of risk for trach/death. Moreover, the mean hospital discharge gestational age was substantially higher for the trach or death group (73.74 days), compared to those without such outcomes (47.82 days, $p < 0.001$), implying a longer hospital stay and possibly more complex postnatal course for those with adverse outcomes.

Table 2 shows the diagnostic and respiratory parameters of infants at 36-week PMA. All p-values are less than 0.001, showing these characteristics at 36-week PMA are all signficantly different between the trach/death group and the no trach/death group. Similar as the weight measured at birth, the weight at 36 weeks for infants in the no trach or death group had a higher mean weight (2141.49g) than infants in the trach or death group(1989.12g, $p < 0.001$). Ventilation support levels at 36 weeks PMA shows obvious differences. Infants who did not experience trach or death has higher proportion of no respiratory support or supplemental oxygen than infants with trach/death (14% vs. 4.4%, $p < 0.001$). In contrast, a larger fraction of infants with trach/death outcomes need invasive positive pressure (74% vs. 17%, $p < 0.001$), indicating the level of respiratory support may contribute to the outcome signficantly. In terms of respiratory parameters, the fraction of inspired oxygen at 36 weeks was significantly higher in the trach or death group ($p < 0.001$); the peak inspiratory pressure and the positive and exploratory pressure at 36 weeks also show this trend, with the trach or death group requiring higher pressures ($p < 0.001, p < 0.001$). Additionally, the use of medication for pulmonary hypertension at 36 weeks was significantly more prevalent among those with trach or death (20% vs. 4.0%, $p < 0.001$), emphasizing the potential influence of pulmonary vascular disease in the outcome of trach or death. These findings emphasize these diagnostic and respiratory parameters of infants at 36-week PMA may be strong predictors of the outcome.

Table 3: Summary Statistics by Center

Variable Center	1, N = 65	2, N = 629	3, N = 54	4, N = 59	5, N = 40	7, N = 32	12, N = 68	16, N = 38	p-value
Maternal Ethnicity									<0.001
Hispanic or Latino	6 / 40 (15%)	24 / 629 (3.8%)	13 / 52 (25%)	5 / 57 (8.8%)	8 / 40 (20%)	1 / 5 (20%)	7 / 68 (10%)	6 / 38 (16%)	
Not Hispanic or Latino	34 / 40 (85%)	605 / 629 (96%)	39 / 52 (75%)	52 / 57 (91%)	32 / 40 (80%)	4 / 5 (80%)	61 / 68 (90%)	32 / 38 (84%)	
(Missing)	25	0	2	2	0	27	0	0	
Birth Weight (g)									<0.001
Mean (SD)	689.88 (215.67)	832.23 (312.86)	773.59 (256.19)	837.20 (259.17)	605.35 (107.03)	724.88 (225.73)	779.54 (255.14)	889.29 (354.17)	
Gestational Age									<0.001
Mean (SD)	25.66 (1.84)	25.88 (2.19)	25.74 (2.09)	25.78 (2.03)	24.08 (1.54)	25.09 (1.86)	26.07 (1.93)	26.29 (2.36)	
Birth Length (cm)									<0.001
Mean (SD)	30.84 (3.95)	32.74 (3.85)	32.35 (3.56)	33.26 (3.47)	29.46 (2.01)	32.08 (3.33)	32.41 (3.01)	33.71 (3.96)	
(Missing)	9	24	0	1	1	6	36	0	
Birth Head Circumference (cm)									<0.001
Mean (SD)	22.58 (2.24)	23.31 (2.73)	23.68 (3.21)	23.76 (2.89)	21.05 (1.51)	22.27 (2.22)	23.23 (2.79)	23.76 (2.92)	
(Missing)	9	29	0	2	0	6	30	0	
Delivery Method									0.9
Vaginal delivery	16 / 64 (25%)	176 / 629 (28%)	15 / 54 (28%)	18 / 59 (31%)	14 / 40 (35%)	10 / 32 (31%)	17 / 66 (26%)	14 / 38 (37%)	
Cesarean section	48 / 64 (75%)	453 / 629 (72%)	39 / 54 (72%)	41 / 59 (69%)	26 / 40 (65%)	22 / 32 (69%)	49 / 66 (74%)	24 / 38 (63%)	
(Missing)	1	0	0	0	0	0	2	0	
Prenatal Corticosteroids	56 / 61 (92%)	543 / 628 (86%)	46 / 52 (88%)	46 / 58 (79%)	37 / 40 (93%)	26 / 30 (87%)	41 / 46 (89%)	33 / 37 (89%)	0.5
(Missing)	4	1	2	1	0	2	22	1	
Complete Prenatal Steroids	34 / 54 (63%)	414 / 609 (68%)	40 / 52 (77%)	24 / 54 (44%)	27 / 40 (68%)	15 / 27 (56%)	26 / 44 (59%)	26 / 37 (70%)	0.013
(Missing)	11	20	2	5	0	5	24	1	
Maternal Chorioamnionitis	17 / 35 (49%)	105 / 629 (17%)	4 / 32 (13%)	8 / 58 (14%)	14 / 39 (36%)	3 / 31 (9.7%)	5 / 68 (7.4%)	2 / 33 (6.1%)	<0.001
(Missing)	30	0	22	1	1	1	0	5	
Gender									0.7
Female	26 / 64 (41%)	248 / 627 (40%)	21 / 53 (40%)	28 / 59 (47%)	17 / 40 (43%)	16 / 32 (50%)	28 / 68 (41%)	20 / 38 (53%)	
Male	38 / 64 (59%)	379 / 627 (60%)	32 / 53 (60%)	31 / 59 (53%)	23 / 40 (58%)	16 / 32 (50%)	40 / 68 (59%)	18 / 38 (47%)	
(Missing)	1	2	1	0	0	0	0	0	
Small for Gestational Age	26 / 64 (41%)	117 / 619 (19%)	11 / 51 (22%)	5 / 58 (8.6%)	8 / 40 (20%)	8 / 32 (25%)	17 / 68 (25%)	7 / 38 (18%)	0.002
(Missing)	1	10	3	1	0	0	0	0	
Weight at 36 weeks (g)									0.011
Mean (SD)	2,064.17 (454.02)	2,134.23 (408.36)	2,131.27 (423.27)	2,119.87 (339.47)	1,921.83 (401.40)	2,169.32 (408.61)	2,044.28 (483.95)	2,219.50 (408.51)	
(Missing)	17	36	3	6	0	1	29	0	
Ventilation Support Level at 36 weeks									<0.001
No respiratory support or supplemental oxygen	8 / 64 (13%)	49 / 620 (7.9%)	5 / 53 (9.4%)	8 / 59 (14%)	0 / 40 (0%)	22 / 32 (69%)	1 / 49 (2.0%)	22 / 38 (58%)	
Non-invasive positive pressure	22 / 64 (34%)	425 / 620 (69%)	34 / 53 (64%)	34 / 59 (58%)	31 / 40 (78%)	8 / 32 (25%)	16 / 49 (33%)	14 / 38 (37%)	
Invasive positive pressure	34 / 64 (53%)	146 / 620 (24%)	14 / 53 (26%)	17 / 59 (29%)	9 / 40 (23%)	2 / 32 (6.3%)	32 / 49 (65%)	2 / 38 (5.3%)	
(Missing)	1	9	1	0	0	0	19	0	
Fraction of Inspired Oxygen at 36 weeks									<0.001
Mean (SD)	0.42 (0.20)	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)	
(Missing)	18	36	2	3	0	1	29	0	
Peak Inspiratory Pressure (cmH2O) at 36 weeks									<0.001
Mean (SD)	7.28 (8.27)	5.31 (10.78)	6.88 (7.74)	5.20 (5.80)	4.06 (6.74)	0.14 (0.79)	9.36 (7.10)	1.29 (4.67)	
(Missing)	20	39	5	15	13	1	32	0	
Positive and exploratory pressure (cm H2O) at 36 weeks									<0.001
Mean (SD)	7.34 (4.42)	6.49 (2.36)	7.71 (3.18)	5.58 (2.51)	8.83 (1.57)	1.68 (2.75)	6.56 (2.02)	3.34 (4.12)	
(Missing)	24	41	9	6	0	1	34	0	
Medication for Pulmonary Hypertension at 36 weeks	13 / 64 (20%)	25 / 620 (4.0%)	3 / 53 (5.7%)	10 / 59 (17%)	3 / 40 (7.5%)	2 / 32 (6.3%)	4 / 49 (8.2%)	4 / 38 (11%)	<0.001
(Missing)	1	9	1	0	0	0	19	0	
Outcome									<0.001
No Trach or Death	31 / 65 (48%)	545 / 629 (87%)	53 / 54 (98%)	47 / 59 (80%)	33 / 40 (83%)	31 / 32 (97%)	27 / 68 (40%)	37 / 38 (97%)	
Trach or Death	34 / 65 (52%)	84 / 629 (13%)	1 / 54 (1.9%)	12 / 59 (20%)	7 / 40 (18%)	1 / 32 (3.1%)	41 / 68 (60%)	1 / 38 (2.6%)	

[1] n / N (%)

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

Table 3 shows the summary statistics stratified by center. We can see that most of variables show significant differences between at least a pair of centers, which indicates including center in our model is essential. We observed distinct patterns of missingness across different centers, showing center-specific data management practices or reporting standards. For maternal ethnicity, center 7 had 27 of 32 records missing, whereas other centers had significantly lower proportions; center 12 reported nearly half of the birth length and head circumference data missing, about 35% missing steroid data, and approximately 30% to 50% missing diagnostic and respiratory parameters at 36-week PMA. Centers 1 and 3 each had around 40% to 45% missing values for maternal chorioamnionitis. All hospital discharge gestational age records were missing for center 4, and nearly all (98.46%) were missing for center 1.

2.2. Train-Test Split

We conducted train-test split before multiple imputation. We use stratified sampling method to select 80% records (N=794) in the training data and 20% records (N=191) in the test data. The stratification is based on the interaction of outcome and centers to assess the balance of not only the positive and negative outcome but also the centers.

2.3. Missing Data and Multiple Imputation

Tables 1 and 2 show that several covariates have a non-negligible proportion of missing values, with the peak inspiratory pressure at 36 weeks PMA having the highest with 125 (12.69%) missing entries. From Table 3, we observed distinct patterns of missingness across different centers.

Instead of using a deletion method, which could introduce bias and reduce statistical power due to a smaller sample size (Baraldi and Enders 2010), we opted for multiple imputation to fill in gaps before proceeding with model derivation, assuming that the missing data is random (MAR) (Rubin 1976, 2018). This assumption is reasonable, as we see strong correlations between variables, making imputation conditioning on existing data promising.

As I mentioned previously, we observed distinct patterns of missingness across different centers. These variations in missing data underscore the necessity for multiple imputation including the center variable. Record ID was excluded from the imputation. We performed multiple imputation on the training data using the *mice* function (Buuren and Groothuis-Oudshoorn 2011), generating five complete datasets using the mice function. This imputation method was then consistently applied to the test data, resulting in five complete test datasets.

2.4. Model Derivation

After preprocessing, each imputed dataset contains 19 potential predictors for a binary outcome: trach or death versus neither trach nor death. We aim to build a logistic regression model. As there are 19 predictors and more will be generated after one-hot encoding and considering the interaction terms, variable selection methods are implemented for model simplification and reducing the possibility of overfitting. Two variable selection methods are considered as they are typical and efficient: best subset selection (Beale et al. 1967), a classical method of selecting the optimal subset of predictors for fitting the outcome, and the least absolute shrinkage and selection operator (Lasso) that utilizes L1 norm regularization in the penalized likelihood

function (Tibshirani 1996). Our goal is to use these methods to create efficient models for predicting the outcome.

As there are differences among centers seen in Table 3, we considered include center indicators in models. Fixed effect for centers are simplier on implementation and easier to intrepret to non-technical audience, while it only ensures prediction among these existing centers, making model without generalizability to extensive population. Random effect for centers in mixed-effect model addresses the multilevel structure based on centers and it is available to be generalized to the other centers' population.

2.4.1. Interaction Terms

The previous study from Truog et al. (2014) only considered the main effect model. However, we conduct multivariate exploratory analysis here and consider to include some interaction terms in the model. By ANOVA testing, we found that the level of gestational age, delivery method, prenatal corticosteroids, infant's gender, whether infant is small for gestational age, weight at 36 weeks, fraction of inspired oxygen at 36 weeks, peak inspiratory pressure at 36 weeks and positive and exploratory pressure at 36 weeks and the use of medication for pulmonary hypertension at 36 weeks may modify the effect of ventilation support level at 36 weeks to the outcome. Therefore, we consider to include relevant interaction terms in variable selection procedure; variable selection methods selects proper predictors in the final model.

2.4.2. Fixed-effect Model with Best Subset Selection

After obtaining five imputed complete training datasets, we apply best subset logistic regression with tenfold cross-validation to minimize the cross-validation loss function for each imputed dataset. We use the *L0Learn* package (Hazimeh et al. 2023), which is based on the mixed integer optimization formulation (Dedieu et al. 2021).

The variable selection process may yield different subsets of predictors for the five imputed datasets, and we need to combine the five best subsets in one final best subset. Wood et al. (2008) proposed three ad hoc rules for finalizing selection after identifying the best subset of predictors in each imputed dataset (Du et al. 2020): Method 1 is to take the union of these five subsets, meaning any predictor appearing in any model will be in the final subset; method 2 is to include predictors appearing in at least half (i.e., three) of the models; method 3 is to include only the predictors that appear in all models. We use these methods to consolidate five best subsets into three final subsets. We then refit the logistic regression model using only the final subset of predictors to the five imputed datasets. The final coefficients with their confidence intervals are aggregated across the five refitted models using Rubin's rule (Rubin 2018). We thus obtain three fixed-effect best subset logistic regression models.

2.4.3. Fixed-effect Model with Lasso Variable Selection

In contrast to the best subset selection, Lasso regression employs L1 norm regularization to shrink some coefficients to zero, thereby excluding variables from the model. This regularization strength is controlled by a tuning parameter λ. As λ increases, more coefficients are set to zero, leading to sparser models. We use tenfold cross-validation to determine the optimal tuning parameter λ for each imputed training dataset, and refit the logistic Lasso regression to each training dataset. We use the three methods from Wood et al. (2008) and Rubin's rule (Rubin 2018) for combining coefficients. Method 1 is simply taking the average of

all coefficients, which means a coefficients is non-zero in any model will be in the final Lasso model; method 2 is taking the average of coefficients that are non-zero in at least a half (i.e., three) of the models; method 3 is taking the average of coefficients that are non-zero across five imputed data sets. We finally have three fixed-effect logistic lasso regression models.

2.4.4. Mixed-effect Model with Selected Variables

We also considered mixed-effect model to address the multilevel structure clustered by center. As we have selected predictors from best subset and Lasso model in Section 2.3.2 and Section 2.3.3, and we will apply the variable selection results to the mixed-effect model.

2.5. Model Performance

Models are evaluated for their discrimination and calibration on the test data. We firstly optimized the threshold for classification by selecting the optimal threshold that makes the point (1-specificity, sensitivity) on the ROC curve have the closest Euclidean distance to $(0, 1)$, on training data. This method minimizes the distance to the "perfect classification" point $(0, 1)$ on the ROC curve on training data. Based on optimized threshold, we evaluated model discrimination on test data using the receiver operating characteristic curve (ROC curve), area under the ROC Curve (AUC), sensitivity, specificity, accuracy, and precision. Model calibration is assessed by brier score and the calibration plot on test data, which visualizes how close our estimated distribution and true distribution are to each other (Paul 2023a).

3. Results

3.1. Model Coefficients

Table 4 shows the times of variable selected by best subset and lasso regression fixed-effect models in five imputed training data and the coefficients under three combination methods defined in Section 2.4. We can see that the variables finally included in the best subset models are less than those from lasso regression models.

The most consistent predictors are the interaction of ventilation support level and gender, the interaction of ventilation support level and medication for pulmonary hypertension at 36 weeks. They have non-zero coefficients across all models, which indicating they essentially contribute to the outcome prediction. This also validate the importance of including possible interaction terms in the model. And this shows the importance of the diagnostic and respiratory parameters of infants at 36-week PMA to the prediction of trach/death. Most of coefficients for center indicators are non-zeros across imputed data sets, which also indicates potential variability in patient populations by different medical centers could influence outcomes.

Table 4 also shows the fixed-effect part of mixed-effect models with variables selected by M2 from best subset and Lasso. We can see the importance of the diagnostic and respiratory parameters of infants at 36-week PMA to the prediction of trach/death. This finding reinforced the utility of these diagnostic and respiratory parameters at 36-week PMA in predicting trach/death, thereby addressing the research gap mentioned in Section 1.

The random intercept for best subset follows $N(\mu_{bs}, 1.519)$ and the random intercept for Lasso follows

Table 4: Model Coefficients

	Best Subset (Fixed-Effect)				Best Subset	Lasso (Fixed-Effect)				Lasso
	Times	M1	M2	M3	Mixed-Effect	Times	M1	M2	M3	Mixed-Effect
(Intercept)	5	**0.54**	1.56	2.05	**0.02**	5	-2.31	-2.31	-2.31	-2.38
center2	5	**-2.07**	-1.92	-1.80		5	-1.43	-1.43	-1.43	
center3	5	**-4.07**	-3.80	-3.78		5	-2.97	-2.97	-2.97	
center4	3	**-2.15**	-1.91	-1.95		5	-1.18	-1.18	-1.18	
center5	3	**-1.66**	-1.63	-1.45		5	-0.79	-0.79	-0.79	
center7	3	**-2.83**	-2.63	-2.57		5	-1.81	-1.81	-1.81	
center12	2	**-0.01**	0.15	0.23		5	0.48	0.48	0.48	
center16	3	**-2.49**	-2.33	-2.34		5	-1.81	-1.81	-1.81	
mat_ethnNot Hispanic or Latino	3	**0.60**	0.63		**0.63**	5	0.29	0.29	0.29	0.51
bw	1	**0.00**				4	0.00	0.00		0.00
birth_hc	2	**0.01**				4	0.02	0.02		0.02
prenat_sterYes	4	**0.76**	0.81		**0.81**	5	0.44	0.44	0.44	-0.05
com_prenat_sterYes	2	**0.19**				5	0.19	0.19	0.19	0.20
mat_chorioYes	1	**-0.27**				3	-0.09	-0.09		-0.22
genderMale	1	**-0.51**	-0.51	-0.63	**-0.50**	2	-0.05			-0.36
weight_today.36	3	**0.00**	0.00		**0.00**	5	0.00	0.00	0.00	0.00
ventilation_support_level.36Non-invasive positive pressure	1	**-5.36**	-5.54	-5.83	**-5.65**	1	-0.02			-1.06
genderMale:ventilation_support_level.36Non-invasive positive pressure	1	**0.04**	0.11	0.15	**0.11**	4	-0.18	-0.18		-0.04
genderMale:ventilation_support_level.36Invasive positive pressure	5	**0.91**	1.00	1.15	**1.00**	5	0.48	0.48	0.48	0.80
sgaYes:ventilation_support_level.36Invasive positive pressure	1	**0.01**				3	0.12	0.12		0.38
ventilation_support_level.36Non-invasive positive pressure:inspired_oxygen.36	1	**14.68**	15.03	16.17	**15.65**	3	3.21	3.21		15.13
ventilation_support_level.36Invasive positive pressure:inspired_oxygen.36	5	**9.25**	10.16	10.88	**10.73**	4	1.72	1.72		9.99
ventilation_support_level.36Non-invasive positive pressure:p_delta.36	2					5	0.04	0.04	0.04	
ventilation_support_level.36Non-invasive positive pressure:med_ph.36Yes	5	**14.60**	13.93	14.02	**14.22**	5	1.88	1.88	1.88	15.86
sgaYes		**0.18**				3	0.08	0.08		-0.01
ventilation_support_level.36Invasive positive pressure		**-1.90**	-2.43	-2.76	**-2.52**	2	0.03			0.64
inspired_oxygen.36		**-5.93**	-6.72	-7.34	**-7.29**	3	1.27	1.27		-6.85
med_ph.36Yes		**-12.54**	-12.11	-12.25	**-12.45**	1	0.02			-13.85
sgaYes:ventilation_support_level.36Non-invasive positive pressure		**0.09**								0.28
ventilation_support_level.36Invasive positive pressure:med_ph.36Yes		**12.47**	12.17	12.19	**12.52**					13.77
del_methodCesarean section						3	0.07	0.07		0.50
ga										0.07
p_delta.36						5	0.01	0.01	0.01	0.01
ga:ventilation_support_level.36Non-invasive positive pressure						3	-0.03	-0.03		-0.19
ga:ventilation_support_level.36Invasive positive pressure						2	0.00			-0.16
del_methodCesarean section:ventilation_support_level.36Non-invasive positive pressure						2	-0.03			-0.63
del_methodCesarean section:ventilation_support_level.36Invasive positive pressure						4	0.08	0.08		-0.28
prenat_sterYes:ventilation_support_level.36Non-invasive positive pressure										1.35
prenat_sterYes:ventilation_support_level.36Invasive positive pressure						4	0.18	0.18		0.77
weight_today.36:ventilation_support_level.36Non-invasive positive pressure						5	0.00	0.00	0.00	0.00
weight_today.36:ventilation_support_level.36Invasive positive pressure						4	0.00	0.00		0.00
peep_cm_h2o_modified.36						1	-0.01			
ventilation_support_level.36Invasive positive pressure:peep_cm_h2o_modified.36						1	0.01			

Note:
Blank cells in the table indicate that the variables are never selected in that senario. Times: times of variable appeared in five models from five imputed training data. M1: Method 1. M2: Method 2. M3: Method 3. Ventilation Support Level at 36 weeks: 2 is the indicator of Non-invasive positive pressure compared to No respiratory support or supplemental oxygen; 3 is the indicator of Invasive positive pressure compared to No respiratory support or supplemental oxygen. The coefficients of the final models are bold. For best subset model, when we include the interaction terms, the relevant main effect terms are also included in the model.

Table 5: Evaluation Metrics for Best Subset and Lasso Regression

	Best Subset (Fixed-Effect)			Best Subset	Lasso (Fixed-Effect)			Lasso
	M1	M2	M3	Mixed-Effect	M1	M2	M3	Mixed-Effect
AUC	**0.9106**	0.9134	0.8948	**0.9079**	0.9059	0.9068	0.8873	0.9054
Sensitivity	**0.8485**	0.8061	0.8061	**0.8303**	0.7697	0.8000	0.8545	0.8121
Specificity	**0.8582**	0.8772	0.8772	**0.8620**	0.8785	0.8646	0.8190	0.8456
Accuracy	**0.8565**	0.8649	0.8649	**0.8565**	0.8597	0.8534	0.8251	0.8398
Precision	**0.5556**	0.5783	0.5783	**0.5569**	0.5695	0.5523	0.4965	0.5234
Brier Score	**0.0870**	0.0864	0.0899	**0.0865**	0.0901	0.0898	0.1496	0.0899
Threshold	**0.1873**	0.2048	0.1966	**0.1873**	0.2309	0.1990	0.0210	0.2063

$N(\mu_{lasso}, 2.125)$ pooled by Rubin's Rule (Rubin 2018). The details of μ_{bs} and μ_{lasso} are in Table 6 and Table 7 in Appendix.

3.2. Model Performance

Evaluation metrics for models shown in Table 4 are displayed in Table 5. We used the optimal threshold from training data for classification. We can see that, generally, the model from pooling method 1 and method 2 outperforms the model with same variable selection method under pooling method 3; while model from pooling method 1 performed nearly equally well as pooling method 2. Sometimes, the model under pooling method 2 outperforms pooling method 1, which indicating the over-fitting exists under pooling method 1. For fixed-effect models, the best subset pooled by method 2 outperforms the other models, while the best subset pooled by method 1 has a higher sensitivity, indicating clinicians can find more proportion of the infants with higher risk of trach/death in practice. For mixed-effect models, the model based on variables from best subset variable selection with pooling method 2 has better performance.

We finally decides the best subset pooled by method 1 with higher sensitivity as the final fixed-effect model. The AUC on test is 0.9106 (95% CI: 0.8882-0.9331), with sensitivity of 0.8485, specificity of 0.8582, accuracy of 0.8565, precision of 0.5556 and the Brier score of 0.0870. And the mixed-effect model based on variables from best subset variable selection with pooling method 2 is the final mixed-effect model. The AUC on test is 0.9079 (95% CI: 0.8845-0.9313), with sensitivity of 0.8303, specificity of 0.8620, accuracy of 0.8565, precision of 0.5569, and Brier score of 0.0865. The coefficients for the final models are bold in Table 4.

The ROC curves are shown in Figure 1. We can see that both the fixed-effect models and the mixed-effect models show great performance on ROC curves. The plot also validates the importance of threshold optimization.

The calibration plots for the best models are shown in Figure 2. It is created from the estimated standard error to create corresponding 95% confidence intervals plot for the observed vs expected proportions. Overall, the plot shows that our models could be better calibrated. The observed proportions have higher variances in the middle range of probabilities, indicating less certainty about the calibration in this region. The points with error bars show that for lower predicted probabilities, the observed frequencies are higher than expected, indicating underestimation of risk. On the contrary, for higher predicted probabilities, the points align lower than the diagonal, suggesting overestimation of risk of trach/death. For best subset fixed-effect model, the calibration in the range of probabilities around 0.35 to 0.4 is relevantly poor; for the best subset mixed-effect

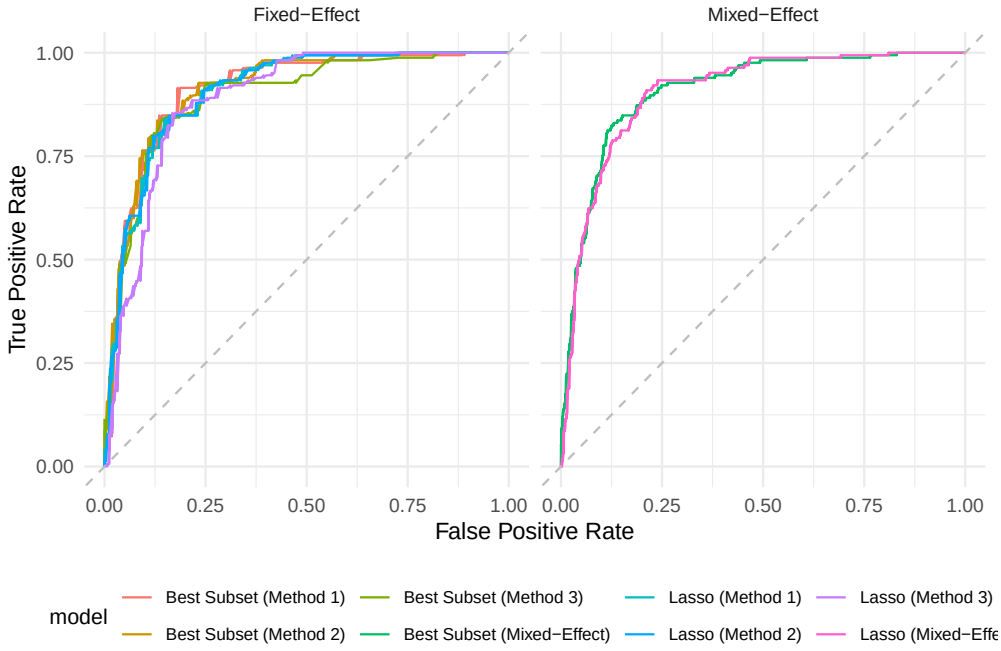

Figure 1: ROC curves for models excluding/including center

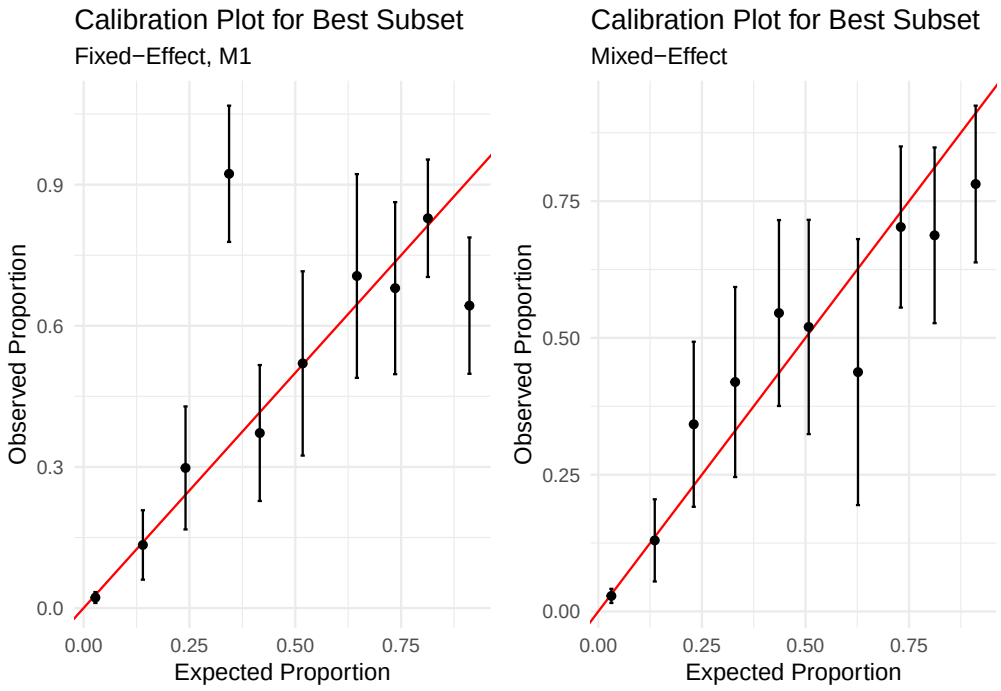

Figure 2: Calibration Plot for Final Models

model, the calibration in the range of probabilities around 0.6 to 0.65 is relevantly poor. The calibration plot shows the final models need better calibration.

4. Discussion

In our study, we construct logistic regression models with baseline characteristics, diagnostics, respiratory parameters at 36-week PMA with some possible interaction terms to predict tracheostomy or death in infants with sBPD. Initial exploratory analysis indicated certain covariates' strong unadjusted marginal association with outcomes. We then applied best subset selection and Lasso for robust variable selection in logistic regression modeling for fixed-effect models and then applied the variable selection results to mixed-effect model. Different methods are applied to each imputed training data and pooled by three methods. We then evaluate the model performance on the test data based on their discrimination and calibration.

We discussed final models in Section 3.2, and selected the best subset fixed-effect model and best subset mixed-effect model. The fixed-effect one outperforms the mixed-effect one slightly. Therefore, our final model selection strategy varies depending on the intended application: for new centers, we chose the best subset mixed-effect model for generalizability, while for existing centers, we selected best subset fixed-effect model. The final models are pooled by including all non-zero coefficients across five imputed training data. We found that both best subset fixed-effect model and best subset mixed-effect model performed well, with high sensitivity, specificity and accuracy, indicating clinicians leverage these models to find the infants with higher risk of trach/death in practice. The ROC curves with evaluation metrics show the model's great performance, while the calibration plots show the calibration around the middle range of probabilities is relevantly poor. Besides, we would also be interested in look at the model performance in different populations with transportability and generalizability analysis.

However, there are some limitations of our study. Firstly, we used multiple imputation assuming the data is MAR, which means the missing values can be predicted from the observed data. This assumption is untestable, and may be not hold if the data is missing not at random (MNAR), potentially affecting our estimates. Secondly, including interaction terms increased the complexity of the models. Simplier models might be considered for use in practice. Additionally, we do not conduct group k-fold cross-validation, remaining the generalizability invalidated. Future studies should include more analysis with group k-fold cross-validation, cluster bootstrap, and sensitivity analysis to validate the model's generalizability.

5. Conclusions

In this study, we used baseline characteristics, diagnostics, and respiratory parameters at 36-week PMA to predict tracheostomy or death in infants with sBPD. We applied best subset selection and Lasso for variable selection in fixed-effect and mixed-effect logistic regression modeling. The final models include the best subset fixed-effect model for prediction within known medical centers and best subset mixed-effect model for prediction generalized to new medical centers. Both final models show great performance for tracheostomy or death prediction. It provides accurate prediction of need for tracheostomy for infants with sBPD at early PMA, which would have implications for counseling families and tracheostomy placement.

References

Baraldi, A. N., and Enders, C. K. (2010), "An introduction to modern missing data analyses," *Journal of school psychology*, 48, 5–37.

Beale, E. M. L., Kendall, M. G., and Mann, D. W. (1967), "The Discarding of Variables in Multivariate Analysis," *Biometrika*, 54, 357–366. https://doi.org/10.2307/2335028.

Buuren, S. van, and Groothuis-Oudshoorn, K. (2011), "Mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software*, 45, 1–67. https://doi.org/10.18637/jss.v045.i03.

Dedieu, A., Hazimeh, H., and Mazumder, R. (2021), "Learning Sparse Classifiers: Continuous and Mixed Integer Optimization Perspectives," *Journal of Machine Learning Research*, 22, 1–47.

Du, J., Boss, J., Han, P., Beesley, L. J., Goutman, S. A., Batterman, S., Feldman, E. L., and Mukherjee, B. (2020), "Variable selection with multiply-imputed datasets: Choosing between stacked and grouped methods," arXiv. https://doi.org/10.48550/arXiv.2003.07398.

Hazimeh, H., Mazumder, R., and Nonet, T. (2023), "L0learn: A scalable package for sparse learning using l0 regularization," *Journal of Machine Learning Research*, 24, 1–8.

McKinney, R., and Levin, J. (2023), "Predicting the need for tracheostomy in infants with severe bronchopulmonary dysplasia."

Paul, A. (2023a), "Health Data Science in R."

Paul, A. (2023b), "Project 2: Regression Analysis."

Rubin, D. B. (1976), "Inference and missing data," *Biometrika*, 63, 581–592.

Rubin, D. B. (2018), "Multiple imputation," in *Flexible Imputation of Missing Data, Second Edition*, Chapman; Hall/CRC, pp. 29–62.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

Truog, W., Grover, T., Zhang, H., Asselin, J., Durand, D., and others (2014), "Predicting death or tracheostomy placement in infants with severe bronchopulmonary dysplasia," *Journal of Perinatology*, 34, 543–548.

Wood, A. M., White, I. R., and Royston, P. (2008), "How should variable selection be performed with multiply imputed data?" *Statistics in Medicine*, 27, 3227–3246. https://doi.org/10.1002/sim.3177.

Zhang, H., Luo, J., Shepard, S., Nilan, K., Harrington, A., Jensen, E., Maschhoff, K., and Kirpalani, H. (2018), "Improved Growth and Infant Participation in Developmental Activities After Tracheostomy Placement in Infants with Severe Bronchopulmonary Dysplasia," *Pediatrics*, 141, 530. https://doi.org/10.1542/peds.141.1MA6.530.

Appendix

Table 6: Mean of Random Effects of Mixed-Effect Lasso

	Imputed Data 1	Imputed Data 2	Imputed Data 3	Imputed Data 4	Imputed Data 5	Average
Center 1	1.78	1.57	1.70	1.64	1.65	1.67
Center 2	-0.19	-0.29	-0.36	-0.41	-0.37	-0.32
Center 3	-1.59	-1.54	-1.37	-1.40	-1.46	-1.47
Center 4	-0.21	-0.38	-0.26	-0.21	-0.19	-0.25
Center 5	0.12	-0.03	-0.04	0.05	0.01	0.02
Center 7	-0.68	-0.48	-0.37	-0.53	-0.46	-0.50
Center 12	1.82	1.77	1.53	1.48	1.77	1.68
Center 16	-0.59	-0.22	-0.46	-0.27	-0.54	-0.42

Table 7: Mean of Random Effects of Mixed-Effect Best Subset

	Imputed Data 1	Imputed Data 2	Imputed Data 3	Imputed Data 4	Imputed Data 5	Average
Center 1	1.49	1.42	1.61	1.48	1.48	1.50
Center 2	-0.18	-0.26	-0.31	-0.30	-0.33	-0.28
Center 3	-1.34	-1.39	-1.30	-1.32	-1.34	-1.34
Center 4	-0.22	-0.32	-0.23	-0.23	-0.16	-0.23
Center 5	0.01	-0.07	0.00	0.02	0.04	0.00
Center 7	-0.61	-0.46	-0.54	-0.51	-0.55	-0.53
Center 12	1.71	1.70	1.55	1.47	1.76	1.64
Center 16	-0.46	-0.25	-0.39	-0.26	-0.48	-0.37

Code Appendix

All code is compatible with R version 4.1.2, and can be found in the *Project 2 Regression Analysis* folder at Github.

Code for generating this report is called `Project-2_updated.Rmd`; the pre-processing code can be found in a separate file called `pre_processing_EDA_updated.R`; the code for best subset selection can be found in a separate file called `best_subset_updated.R`; the code for Lasso regression can be found in a separate file called `lasso_updated.R`; the code for multilevel models can be found in a separate file called `multilevel.R`.