Predicting the Need for Tracheostomy in Infants with Severe Bronchopulmonary Dysplasia

Kerry Ye

2023-11-06

Abstract

Objective: In collaboration with Dr. Chris Schmid, this project aims to develop statistical models for predicting the composite outcome of tracheostomy or death, considering the indication criteria and timing of tracheostomy placement.

Methods: We proposed four models: 1) 36-Week Model Without Interactions; 2) 36-Week Model With Interactions; 3) 44-Week Model Without Interactions; 4) 44-Week Model With Interactions. The first step in model development involves constructing datasets for both the 36-week and 44-week models. We address missing data in each dataset using the multiple imputation technique under the MAR assumption. We then allocate 70% of each imputed dataset as training data and 30% as the testing set, ensuring the same proportion of the outcome in both sets through stratification. The variable selection process is conducted using a lasso model. The third step involves examining the frequency table of zero coefficients across the five imputed datasets and removing variables that are nearly zero. Finally, we evaluate the performance of the four proposed models using both discrimination and calibration metrics.

Results: Overall, all four models demonstrate similarly high performance in predicting the need for tracheostomy in infants with severe bronchopulmonary dysplasia. The 44-week models have a higher AUC score, indicating better performance in distinguishing between the two outcomes. They also have lower Brier scores, suggesting more accurate probabilistic predictions. The 44-week models are more precise in predicting positive cases, while the 36-week models excel in predicting negative cases. In healthcare contexts, where correctly identifying true positives is often crucial due to the potential consequences of missing a serious condition, the 44-week models may be more appropriate.

Introduction

In collaboration with Dr. Chris Schmid, this project aims to develop statistical models for predicting the composite outcome of tracheostomy or death, considering the indication criteria and timing of tracheostomy placement. Bronchopulmonary Dysplasia (sBPD), a common sequela of prematurity, severely impacts the health of infants[1]. It affects 10,000-15,000 infants annually and is largely influenced by individual susceptibility. Current research suggests that early tracheostomy placement may benefit those with severe Bronchopulmonary Dysplasia (sBPD). However, the indication criteria and optimal timing for tracheostomy placement in neonates with sBPD remain unclear. Tracheostomy provides a stable airway, improves ventilator synchrony, and is associated with improved sBPD outcomes when performed within the first 4 months of age. Conversely, tracheostomy also carries significant risks, including an increased likelihood of death and higher infection rates[2]. Therefore, accurately predicting the eventual need for tracheostomy placement before discharge is essential.

Data Preprocessing

The underlying recruitment population were drawn from the BPD Collaborative Registry, a multi-center consortium of BPD programs located in the United States and Sweden in order to enhance the care of children with severe forms of BPD. The registry includes infants whose gestational age is less than 32 weeks and who have severe sBPD. In the registry, standard demographic and clinical data are collected at four time points: birth, 36 weeks PMA, 44 weeks PMA and discharge. For this study, we collected the data from patients with BPD and complete growth data between January 1 and July 19, 2021.

The raw dataset initially contained 999 observations and 30 variables. After removing 3 duplicate entries, 996 individuals remained. We first converted various categorical variables into factors, reflecting their different levels. Notably, 10 individuals lacked corresponding center information; however, their **record__id** prefix (1xxxxx) indicated they were from **center** 1, so we manually imputed their center values as 1. One individual, with **record__id:2000824**, had four duplicate rows; we retained only one row and removed the duplicates. The **mat__race** column was removed from the data due to coding inconsistencies with the codebook. Additionally, one individual's **record__id** was coded as 21000001, and their **center** code as 21. We corrected their **record__id** to 1000001 and **center** to 1, as their original record_id and center were incorrectly recorded for center 21 instead of center 1. Finally, we created the composite outcome, **outcome**, where individuals who had either undergone a tracheostomy or died, as well as those who had undergone both, were coded as 1, while those who had neither undergone tracheostomy nor died were coded as 0. This composite outcome can provide a more comprehensive understanding of the overall burden and risk associated with tracheostomy and mortality, thereby reflecting the overall severity or progression of Bronchopulmonary Dysplasia.

Exploratory Data Analysis

We conducted exploratory data analysis, which includes examining summary statistics, analyzing missing data, and evaluating correlations among variables.

Summary Statistics

A summary table is important because it provides crucial insights into the data's structure, helps identify potential issues, and informs the choice of appropriate statistical models and methods for analysis. In this section, we first created the following summary statistics table for the dataset, focusing on the two composite outcomes: 1 and 0.

Characteristic	N	**0**, N = 811[1]	**1**, N = 183[1]
center	994		
1		31 (3.8%)	35 (19%)
2		545 (67%)	84 (46%)
3		55 (6.8%)	2 (1.1%)
4		47 (5.8%)	12 (6.6%)
5		33 (4.1%)	7 (3.8%)
7		31 (3.8%)	1 (0.5%)
12		28 (3.5%)	41 (22%)
16		37 (4.6%)	1 (0.5%)
20		4 (0.5%)	0 (0%)
mat__ethn	937		
1		64 (8.3%)	10 (5.9%)
2		703 (92%)	160 (94%)
bw	994		
Mean (SD)		817 (285)	756 (340)
ga	994		
Mean (SD)		26 (2)	26 (2)
blength	916		
Mean (SD)		33 (4)	32 (4)
birth__hc	917		
Mean (SD)		23.25 (2.65)	22.88 (3.29)
del__method	991		
1		245 (30%)	39 (21%)
2		564 (70%)	143 (79%)
prenat__ster	959	679 (86%)	154 (92%)

com__prenat__ster	801	499 (76%)	109 (76%)
mat_chorio	932	132 (17%)	28 (17%)
gender	990		
Female		334 (41%)	73 (40%)
Male		473 (59%)	110 (60%)
sga	979		
Not SGA		658 (82%)	118 (66%)
SGA		142 (18%)	61 (34%)
any__surf	562	374 (81%)	87 (88%)
weight__today.36	902		
Mean (SD)		2,142 (393)	1,981 (507)
ventilation_support_level.36	964		
0		109 (14%)	7 (4.3%)
1		553 (69%)	36 (22%)
2		140 (17%)	119 (73%)
inspired_oxygen.36	902		
Mean (SD)		0.31 (0.12)	0.49 (0.21)
p_delta.36	866		
Mean (SD)		4 (8)	16 (12)
peep__cm__h2o__modified.36	877		
Mean (SD)		6 (3)	7 (3)
med__ph.36	964		
0		770 (96%)	129 (80%)
1		32 (4.0%)	33 (20%)
weight__today.44	550		
Mean (SD)		3,695 (643)	3,473 (782)
ventilation_support_level_modified.44	572		
0		261 (60%)	8 (6.0%)
1		124 (28%)	22 (16%)
2		53 (12%)	104 (78%)
inspired_oxygen.44	548		
Mean (SD)		0.31 (0.11)	0.45 (0.20)
p_delta.44	548		
Mean (SD)		4 (11)	22 (16)
peep__cm__h2o__modified.44	550		
Mean (SD)		3 (4)	9 (3)
med__ph.44	572		
0		405 (92%)	68 (51%)
1		33 (7.5%)	66 (49%)
hosp_dc_ga	871		
Mean (SD)		49 (24)	73 (30)
Trach	994		
0		811 (100%)	37 (20%)
1		0 (0%)	146 (80%)
Death	994	0 (0%)	54 (30%)

[1]n (%)

From the summary table, we observe that there are a total of 811 infant observations where neither tracheotomy nor death occurred, while 183 observations involve either tracheotomy, death, or both. This imbalanced proportion of the outcome necessitates stratifying the outcome when conducting the train-test split. The study records data from nine centers, with Center 2 contributing the most observations, accounting for 630 individuals. Some centers have notably fewer observations; for example, Center 20 has only 4 recorded observations. Regarding birth variables, it is observed that mothers of non-Hispanic or Latino ethnicity

Table 2: Missingness in Tracheostomy Dataset

variable	n_miss	pct_miss
inspired_oxygen.44	448	44.98
p_delta.44	448	44.98
weight_today.44	446	44.78
peep_cm_h2o_modified.44	446	44.78
any_surf	433	43.47
ventilation_support_level_modified.44	424	42.57
med_ph.44	424	42.57
com_prenat_ster	193	19.38
p_delta.36	128	12.85
hosp_dc_ga	124	12.45
peep_cm_h2o_modified.36	117	11.75
weight_today.36	92	9.24
inspired_oxygen.36	92	9.24
blength	78	7.83
birth_hc	77	7.73
mat_chorio	62	6.22
mat_ethn	57	5.72
prenat_ster	35	3.51
ventilation_support_level.36	30	3.01
med_ph.36	30	3.01
sga	15	1.51
gender	4	0.40
del_method	3	0.30
Death	2	0.20
outcome	2	0.20

have a higher proportion of infants undergoing tracheotomy or experiencing death. Infants who underwent tracheotomy or died tend to have lower birth weights, weights at 36 weeks, and weights at 44 weeks compared to those who did not. Additionally, males are more likely to undergo tracheotomy or experience death compared to females. Obstetrical gestational age, birth length, and head circumference appear comparable between groups with composite outcomes of 1 and 0. It is also noted that more individuals in invasive positive pressure ventilation at 36 and 44 weeks underwent tracheotomy or died compared to those in non-invasive positive pressure or without respiratory support. This observation aligns with the expectation that more severe symptoms require more invasive respiratory support. Furthermore, more cases of Cesarean section deliveries, as opposed to vaginal births, are associated with tracheotomy or death. This is consistent with the understanding that earlier deliveries, often necessitating Cesarean sections, are more likely in severe cases of Bronchopulmonary Dysplasia, a common complication in premature births.

Missing Data

Missing data can diminish the statistical power of analyses and may potentially yield less precise and robust estimates in subsequent modeling. To address this, an initial examination of the number and percentage of missing values across the entire dataset is presented in the table below.

Examining the 'Missingness in Tracheostomy Dataset' table reveals that the five variables with the highest percentages of missing data are **inspired_oxygen.44**, **p_delta.44**, **weight_today.44**, **peep_cm_h2o_modified.44**, and **any_surf**, with missingness percentages ranging from 44.98% to 43.47%. Relying on these columns for analysis could skew results due to their substantial amount of

missing data. Notably, **inspired_oxygen.44** and **p_delta.44** have the highest missing data percentage, approximately 44.98%, each with 424 missing values. This table also indicates that variables recorded at 44 weeks exhibit a much higher percentage of missingness compared to those recorded at 36 weeks and at birth. The variables measured at 44 weeks contribute to the highest level of missingness, as many patients are discharged before reaching 44 weeks. Given that the discharge significantly impacts the missing data and the variables themselves are not inherently related to their own missingness, we assume that the data is Missing At Random (MAR).

We further examine the missing values in the data by recording the missingness percentage across nine different centers in order to account for the missingness difference across different centers.

Table 3: Missing Percentage Across Different Centers

|	Center 1	Center 2	Center 3	Center 4	Center 7	Center 12	Center 16	Center 20
mat_ethn	37.88	0.00	3.51	3.33	84.38	0.00	0.00	25
bw	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
ga	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
blength	13.64	3.81	0.00	1.67	18.75	53.62	0.00	0
birth_hc	13.64	4.60	0.00	3.33	18.75	44.93	0.00	0
del_method	1.52	0.00	0.00	0.00	0.00	2.90	0.00	0
prenat_ster	6.06	0.16	5.26	1.67	6.25	33.33	2.63	0
com_prenat_ster	24.24	16.67	17.54	28.33	28.12	37.68	13.16	50
mat_chorio	45.45	0.00	40.35	1.67	3.12	0.00	13.16	25
gender	1.52	0.32	1.75	0.00	0.00	0.00	0.00	0
sga	1.52	1.59	5.26	1.67	0.00	0.00	0.00	0
any_surf	39.39	46.83	1.75	73.33	81.25	17.39	57.89	50
weight_today.36	25.76	5.71	5.26	10.00	3.12	42.03	0.00	0
ventilation_support_level.36	1.52	1.43	1.75	0.00	0.00	27.54	0.00	0
inspired_oxygen.36	28.79	5.71	3.51	5.00	3.12	42.03	0.00	50
p_delta.36	30.30	6.19	12.28	25.00	3.12	46.38	0.00	25
peep_cm_h2o_modified.36	36.36	6.51	19.30	10.00	3.12	49.28	0.00	0
med_ph.36	1.52	1.43	1.75	0.00	0.00	27.54	0.00	0
weight_today.44	10.61	40.00	66.67	100.00	65.62	37.68	86.84	0
ventilation_support_level_modified.44	6.06	37.94	64.91	100.00	62.50	31.88	86.84	0
inspired_oxygen.44	13.64	40.16	66.67	100.00	65.62	36.23	86.84	0
p_delta.44	10.61	39.37	66.67	100.00	68.75	37.68	86.84	0
peep_cm_h2o_modified.44	10.61	39.68	68.42	100.00	62.50	40.58	86.84	0
med_ph.44	6.06	37.94	64.91	100.00	62.50	31.88	86.84	0
hosp_dc_ga	96.97	0.00	0.00	100.00	0.00	0.00	0.00	0
Trach	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
Death	0.00	0.16	0.00	1.67	0.00	0.00	0.00	0
outcome	0.00	0.16	0.00	1.67	0.00	0.00	0.00	0

From the table 'Missing Percentage Across Different Centers,' it is evident that overall, all centers have relatively low missingness in the mother and child information recorded at birth. Notably, no center has missing values for birth weight and obstetrical gestational age. However, Center 4 and Center 1 exhibit a significant proportion of missingness in recording maternal ethnicity, and Center 12 has a considerable proportion of missingness in infants' birth length and head circumference. Examining the variables associated with 36 weeks, Center 16 has no missingness, while Center 12 accounts for half of the missingness. In terms of the variables associated with 44 weeks, almost all the centers exhibit around or more than half of the data missing, except for Center 1 and Center 20, which have a smaller amount of missingness. This high level of missingness is due to many patients being discharged before 44 weeks. Given that Center 20 has only 4 observations, its missing data record at 44 weeks may not be reliable. Center 4 shows the highest percentage of missingness for variables associated with 44 weeks, with 100% missingness, indicating that this center has never recorded information for infants at 44 weeks. Center 16 has over 86% missingness in recording 44-week data. The missingness at 44 weeks is larger than at 36 weeks, and the missingness at 36 weeks is greater than for the variables at birth, primarily because many patients are discharged during 36 and 44 weeks. Researchers should be cautious of this variation in missingness across different centers when

developing related prediction models.

From the patient's perspective, the patient with the highest percentage of missing data is identified by record_id:2000081 from Center 2, with 48.27% missingness and 14 missing values. The patients with the second and third highest missingness, having record_ids 1000049 and 2000586, both exhibit 44.83% in missing values, each with 13 missing values.

Correlation

We identified the issue of multicollinearity by using a correlation heatmap on all continuous variables.
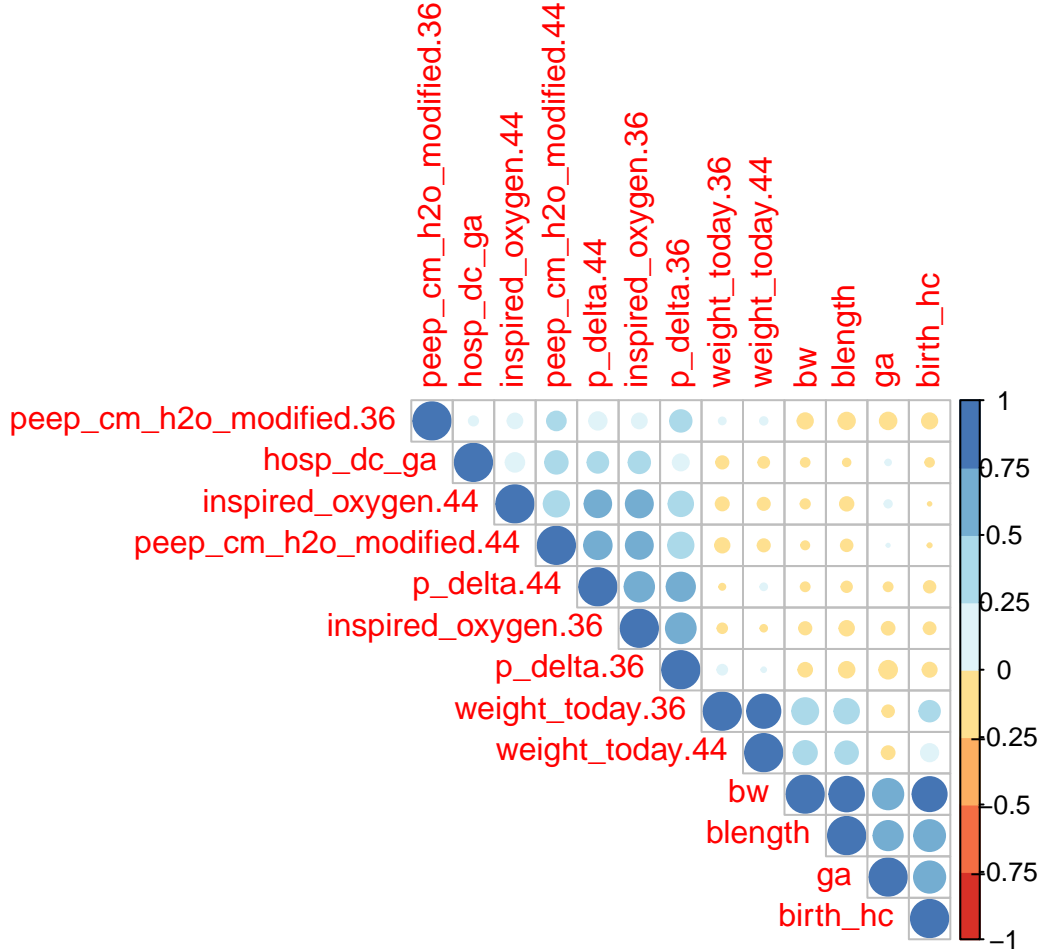
From this correlation heatmap, we can see that there are three pairs of variables that are highly positively correlated, with correlation coefficients above 0.75. These pairs are: weight at 36 weeks (weight_today.36) and weight at 44 weeks (weight_today.44) with a correlation of 0.79; birth weight (bw) and birth head circumference (birth_hc) with a correlation of 0.83; and birth weight (bw) with birth length (blength) with a correlation of 0.86. These highly correlated variables can introduce multicollinearity, which may lead to unreliable and unstable estimates of regression coefficients. However, given that the variable selection method we propose can automatically handle multicollinearity by shrinking the coefficients of correlated predictors to zero, it is acceptable to keep these variables in the variable selection process to retain their information.

Model Development

We propose four models:

1) **36 Week Model Without Interactions**: This model includes all variables measured at birth and up to 36 weeks, considering only main effects.

2) **36 Week Model With Interactions**: This includes all variables at birth and up to 36 weeks. In addition to the main effects, we consider two interaction terms: ventilation_support_level.36:inspired_oxygen.36 and ventilation_support_level.36:med_ph.36.

3) **44 Week Model Without Interactions**: This model includes all variables measured at birth, 36 weeks, and 44 weeks, again considering only main effects.

4) **44 Week Model With Interactions**: This includes all variables at birth, 36 weeks, and 44 weeks. Besides the main effects, four interaction terms are considered: ventilation_support_level.36:inspired_oxygen.36, ventilation_support_level.36:med_ph.36, ventilation_support_level_modified.44:inspired_oxygen.44, and ventilation_support_level_modified.44:med_ph.44.

We focus on these two interactions – inspired oxygen level with ventilation support level, and inspired oxygen level with pulmonary hypertension medication – in both proposed models with interactions. This is because patients with higher inspired oxygen levels might require more invasive ventilation support compared to those with normal oxygen levels, potentially affecting the outcomes of tracheostomy placement or mortality. Additionally, patients with higher inspired oxygen levels might be more likely to take pulmonary hypertension medications, which could also influence these outcomes.

We proposed a variable selection method, **Lasso**, to identify the appropriate variables for inclusion in the proposed prediction models. Lasso regression is an analysis method that performs both variable selection and regularization, enhancing the prediction accuracy and interpretability of the statistical model. It incorporates a penalty term, $\lambda \sum_{j=1}^{p} |\beta_j|$, applied to different parameters of the model, effectively shrinking some coefficients and setting others to zero.

First Step

The first step in model development involves constructing datasets for the 36-week and 44-week models. The 36-week models include variables from birth and at 36 weeks, while the 44-week models encompass variables from birth, at 36 weeks, and at 44 weeks.

For each dataset, we address the missing data using the mice() function in the 'mice' package to create five imputed datasets with filled-in missing values, under the assumption of Missing At Random (MAR). We then allocate 70% of each imputed dataset as the training data and 30% as the testing set, ensuring the same proportion of the outcome in both the training and testing sets through stratification. We use the training data to train the four proposed models and evaluate their performance using the testing sets.

Second Step

The second step in model development involves constructing the variable selection process for each imputed dataset. For each dataset, we fit a lasso regression model. Then, a 10-fold cross-validation procedure is implemented on the lasso regression to determine the best lambda value that minimizes both the sum of squared likelihoods and the penalty term. The coefficients obtained from fitting the lasso model with the best lambda value for each imputed dataset are averaged across five imputed sets to find the final model coefficients.

This process is repeated for the four proposed models. For models that include interaction terms, we specify these terms in the data for the variable selection process to choose from.

Third Step

The Third step of the variable selection process involves examining the frequency table of zero coefficients across the five imputed datasets and removing variables that were shrunk to zero coefficients by the lasso for all five imputed datasets. Additionally, we removed variables that had coefficients nearly zero. Although these variables did not shrink to exactly zero, their impact on the outcome is minimal. After the exclusion of these variables, we established our final model for the five imputed sets.

This process is repeated for the four proposed models.

The below table presents the final coefficients for 36 week models:

Table 4: 36 Week models Regression Coefficient Comparisons

	36 week model	36 week model with interactions
(Intercept)	-3.903	-4.258
Maternal Ethnicity	0.424	0.417
Birth weight (g)	-	-
Obstetrical gestational age	-	-
Birth length (cm)	-	-
Birth head circumference (cm)	-	-
Delivery Method	0.155	0.143
Prenatal Corticosteroids Yes	0.533	0.482
Complete Prenatal Steroids Yes	-	-
Maternal Chorioamnionitis Yes	0.145	0.101
Gender Male	-	-
Infant small for gestational age Yes	-	-
infant receive surfactant Yes	0.14	-
Weight 36wks	-	-
Non-invasive Respiratory Support 36wks	-	-
Invasive Respiratory Support 36wks	1.533	1.725
Inspired Oxygen 36wks	3.228	3.191
Peak Inspiratory Pressure 36wks	-	-
Positive and exploratory pressure 36wks	-	-
Pulmonary Hypertension Med 36wks	0.405	-
Non-invasiveSupport.36:Inspired Oxygen.36	-	0.253
InvasiveSupport.36:InspiredOxygen.36	-	-
Non-invasiveSupport.36:PulmHyperMed.36	-	1.224
InvasiveSupport.36:PulmHyperMed.36	-	-

These two 36 week models share selections for the main effect variables; they both eliminate some highly correlated birth variables such as Birth Weight (g), Obstetrical Gestational Age, Birth Length (cm), and Birth Head Circumference (cm). We can see that during 36 weeks, some factors have significant effects on the outcome of tracheotomy placement and death. For example, the coefficients of 0.533 and 0.482 for prenatal corticosteroids indicate that the use of these steroids is positively associated with the outcome. With coefficients of 3.228 and 3.191 for inspired oxygen, these large positive coefficients indicate a significant association with the outcome. Additionally, for the interaction term Non-invasive Support and Pulmonary Hypertension Medicine (1.224), suggesting that the combined effect of non-invasive support and pulmonary hypertension medication at 36 weeks considerably increases the log odds of the outcome.

The below table presents the final coefficients for 44 week models:

Table 5: 44 Week models Regression Coefficient Comparisons

	44 week model	44 week model with interactions
(Intercept)	-4.328	-4.729
Maternal Ethnicity	0.187	0.274
Birth weight (g)	-	-
Obstetrical gestational age	-	-
Birth length (cm)	-	-
Birth head circumference (cm)	-	-
Delivery Method	0.238	0.256
Prenatal Corticosteroids Yes	0.591	0.534
Complete Prenatal Steroids Yes	-	-
Maternal Chorioamnionitis Yes	-	-
Gender Male	-	-
Infant small for gestational age Yes	-	-
infant receive surfactant Yes	-	-
Weight 36wks	-	-
Non-invasive Respiratory Support 36wks	-	-
Invasive Respiratory Support 36wks	0.697	1.111
Inspired Oxygen 36wks	1.521	1.386
Peak Inspiratory Pressure 36wks	-	-
Positive and exploratory pressure 36wks	-	-
Pulmonary Hypertension Med 36wks	-	-
Weight 44wks	-	-
Non-invasive Respiratory Support 44wks	-	-
Invasive Respiratory Support 44wks	0.994	0.943
Inspired Oxygen 44wks	0.185	0.174
Peak Inspiratory Pressure 44wks	-	-
Positive and exploratory pressure 44wks	0.114	0.105
Pulmonary Hypertension Med 44wks	1.052	0.798
Non-invasiveSupport.36:Inspired Oxygen.36	-	0.683
InvasiveSupport.36:InspiredOxygen.36	-	-
Non-invasiveSupport.36:PulmHyperMed.36	-	1.147
InvasiveSupport.36:PulmHyperMed.36	-	-0.159
Non-invasiveSupport.44:Inspired Oxygen.44	-	-
InvasiveSupport.44:InspiredOxygen.44	-	0.665
Non-invasiveSupport.44:PulmHyperMed.44	-	0.916
InvasiveSupport.44:PulmHyperMed.44	-	-

The coefficients of 0.697 and 1.111 for invasive respiratory support at 36 weeks indicate a strong association between such support and the likelihood of tracheostomy or death. Similarly, coefficients of 1.521 and 1.386 for high levels of inspired oxygen at 36 weeks are strongly associated with the outcome, indicating the severity of respiratory issues. The interaction term for Non-invasive Support at 36 weeks and Inspired Oxygen at 36 weeks (0.683) suggests that the combined effect of non-invasive support and higher inspired oxygen levels increases the odds of tracheostomy and death. Furthermore, the interaction of Invasive Support at 44 weeks and Inspired Oxygen at 44 weeks (0.665) implies that the combined effect of invasive support and oxygen levels positively affects the likelihood of tracheostomy and death. We observe that the interaction involving non-invasive support evolves into one involving invasive support with increased inspired oxygen levels over

time. This progression may be attributed to infants who initially required non-invasive support eventually needing invasive support, compounded by the effects of inspired oxygen.

Model Evaluation

We evaluate the performance of the 36 and 44 week models based on both discrimination and calibration. Discrimination refers to a model's ability to correctly differentiate between positive and negative outcomes. In this context, we use **Sensitivity**, **Specificity**, and the **Area Under the Curve (AUC)** from the ROC-AUC plot to assess the model's ability to distinguish between composite outcomes of 1 and 0. Calibration, meanwhile, measures how well the predicted probabilities of an event match the observed outcomes. We assess the performance of these two models using the **Brier Score**.

For each model, we first manually calculated the predicted outcome values for each imputed testing set by using the coefficients calculated from the trained models, multiplied by the predictor values from the imputed testing set. Then, we used the formula $\frac{1}{1+e^{-(\beta_0+\beta_1+\cdots)}}$ to calculate the predicted probabilities. Finally, we transformed these probabilities into binary outcomes using the optimal threshold.

Table 6: Model Evaluation Measures

	36 Week Model	36 Week Model with Interactions	44 Week Model	44 Week Model with Interactions
AUC	0.886	0.888	0.920	0.920
Sensitivity	0.826	0.822	0.944	0.874
Specificity	0.855	0.859	0.763	0.811
Threshold	0.300	0.248	0.142	0.178
BrierScore	0.100	0.095	0.086	0.083

We calculated evaluation metrics for each imputed testing set and average them across the five imputed testing sets.

Among the four models, they share similar AUC (Area Under the Curve) scores. The 44-week models, both with and without interactions, have slightly higher AUC scores compared to the 36-week models. With an AUC score of 0.92 for both 44-week models, there is a 92% chance that these models will correctly distinguish between patients with and without tracheotomy and death. Meanwhile, the 36-week models, both with and without interactions, have an AUC score of around 0.89, indicating an 89% chance of accurately distinguishing between these patient outcomes.

The 44-week models exhibit higher sensitivity compared to the 36-week models, meaning they more frequently correctly identify patients with active tracheotomy and death. Conversely, the 36-week models demonstrate higher specificity, indicating a greater accuracy in correctly classifying patients without tracheotomy and death.

All models have very small thresholds. The 44-week models have a slightly lower threshold compared to the 36-week models, suggesting that they are set to be more sensitive in detecting positive cases, which could potentially lead to an increase in false positives.

Furthermore, the 44-week models have smaller Brier scores compared to the 36-week models, suggesting that they provide more accurate predictions regarding patients' tracheotomy and death outcomes. Notably, the 44-week model with interactions has the smallest Brier score, indicating its superior predictive accuracy.

Conclusion

Overall, all four models demonstrate similarly high performance in predicting the need for tracheostomy in infants with severe bronchopulmonary dysplasia. The 44-week models have a higher AUC score, indicating better performance in distinguishing between the two outcomes. They also have lower Brier scores, suggesting more accurate probabilistic predictions. The 44-week models are more precise in predicting positive cases,

while the 36-week models excel in predicting negative cases. In healthcare contexts, where correctly identifying true positives is often crucial due to the potential consequences of missing a serious condition, the 44-week models may be more appropriate. Early detection can lead to timely interventions, reducing the risk of complications and improving patient outcomes. However, the choice of the best predictive model should depend on several factors related to the specific needs of researchers and practitioners. It's worth noting that the 36-week models contain less information than the 44-week models, which may impact their predictive capabilities.

This project has several limitations. First, the data contains a substantial amount of missing values, which were filled using a multiple imputation process. While this approach helps in handling missing data, it may introduce biases or inaccuracies. The imputed values are estimates and may not perfectly represent the true missing values, potentially affecting the robustness of the findings. Second, we did not consider any interaction terms in our models. Finally, we did not use mixed effects models, which allow for random effects to account for variability between different centers. This is important because individuals' measurements might be influenced by center-specific characteristics. Future work can be done by considering the mixed effects for model improvement.

Appendix

Table 7: Frequency of Zero Coefficients Across Five Imputed Sets for 36 Week Models

	36 Week Model	36 Week Model with Interactions
(Intercept)	0	0
mat_ethn2	1	1
bw	3	3
ga	3	1
blength	5	5
birth_hc	3	3
del_method2	0	0
prenat_sterYes	0	0
com_prenat_sterYes	5	5
mat_chorioYes	2	2
genderMale	4	5
sgaSGA	2	3
any_surfYes	3	3
weight_today.36	0	0
ventilation_support_level.361	5	5
ventilation_support_level.362	0	0
inspired_oxygen.36	0	0
p_delta.36	3	3
peep_cm_h2o_modified.36	1	4
med_ph.361	0	3
ventilation_support_level.361:inspired_oxygen.36	-	3
ventilation_support_level.362:inspired_oxygen.36	-	5
ventilation_support_level.361:med_ph.361	-	0
ventilation_support_level.362:med_ph.361	-	5

Table 8: Frequency of Zero Coefficients in 44 Week Models Across Five Imputed Sets

	44 Week Model	44 Week Model with Interactions
(Intercept)	0	0
mat_ethn2	1	1
bw	2	3
ga	5	5
blength	5	5
birth_hc	0	1
del_method2	0	0
prenat_sterYes	0	0
com_prenat_sterYes	3	3
mat_chorioYes	5	5
genderMale	5	5
sgaSGA	5	4
any_surfYes	3	3
weight_today.36	2	3
ventilation_support_level.361	5	5
ventilation_support_level.362	0	0
inspired_oxygen.36	0	0
p_delta.36	1	1
peep_cm_h2o_modified.36	1	3
med_ph.361	5	4
weight_today.44	0	0
ventilation_support_level_modified.441	5	5
ventilation_support_level_modified.442	0	0
inspired_oxygen.44	2	3
p_delta.44	1	1
peep_cm_h2o_modified.44	0	0
med_ph.441	0	0
ventilation_support_level.361:inspired_oxygen.36	-	2
ventilation_support_level.362:inspired_oxygen.36	-	5
ventilation_support_level.361:med_ph.361	-	0
ventilation_support_level.362:med_ph.361	-	2
ventilation_support_level_modified.441:inspired_oxygen.44	-	5
ventilation_support_level_modified.442:inspired_oxygen.44	-	2
ventilation_support_level_modified.441:med_ph.441	-	0
ventilation_support_level_modified.442:med_ph.441	-	5

References

[1] Jensen, E. A., & Schmidt, B. (2014). Epidemiology of bronchopulmonary dysplasia. Birth Defects Research Part A: Clinical and Molecular Teratology, 100(3), 145-157.

[2]Mckinney, R., & Levin, J. (2023).Predicting the need for tracheostomy in infants with severe bronchopulmonary dysplasia. (Presentation)

Code Appendix:

```
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(message = F)
knitr::opts_chunk$set(warning = F)
library(mice)# missing data pattern and multiple imputation
library(gtsummary) #summary statistics
library(naniar)
library(tidyverse)
library(knitr)
library(corrplot) # corrlation heat map
library(RColorBrewer)
library(caret) # outcome stratification
library(glmnet)
library(L0Learn)
library(pROC)
library(DescTools)
library(gt)
library(kableExtra)
#load data
tracheo <- read.csv("/Users/kerryqwq/Downloads/project2.csv")
#factor the categorical variables
tracheo$mat_ethn <- as.factor(tracheo$mat_ethn)
tracheo$del_method <- as.factor(tracheo$del_method)
tracheo$prenat_ster <- as.factor(tracheo$prenat_ster)
tracheo$com_prenat_ster <- as.factor(tracheo$com_prenat_ster)
tracheo$mat_chorio <- as.factor(tracheo$mat_chorio)
tracheo$gender <- as.factor(tracheo$gender)
tracheo$sga <- as.factor(tracheo$sga)
tracheo$any_surf <- as.factor(tracheo$any_surf)
tracheo$ventilation_support_level.36 <- as.factor(tracheo$ventilation_support_level.36)
tracheo$ventilation_support_level_modified.44 <- as.factor(tracheo$ventilation_support_level_modified.4
tracheo$Trach <- as.factor(tracheo$Trach)
tracheo$Death <- as.factor(tracheo$Death)
tracheo$med_ph.36 <- as.factor(tracheo$med_ph.36)
tracheo$med_ph.44 <- as.factor(tracheo$med_ph.44)
# Data Preprocessing
tracheo <- tracheo %>%
  mutate(center = ifelse(is.na(center), 1, center)) %>%
  mutate(outcome = case_when(tracheo$Trach ==1 & tracheo$Death == "Yes" ~ 1,
  tracheo$Trach ==1 | tracheo$Death == "Yes" ~ 1,
  tracheo$Trach ==0 & tracheo$Death == "No" ~ 0))
tracheo <- tracheo[!duplicated(tracheo$record_id),]
tracheo <- dplyr::select(tracheo, -c("mat_race"))
tracheo[which(tracheo$center==21),]$record_id <- 1000001
tracheo[which(tracheo$center==21),]$center <- 1

tracheo <- tracheo %>%
  arrange(record_id)

tracheo$outcome <- as.factor(tracheo$outcome)
tracheo$center <- as.factor(tracheo$center)
# Summary statistics table
tracheo %>%
```

```r
  dplyr::select(!record_id) %>%
  tbl_summary(
    by = outcome, # stratify by center
    type = all_continuous() ~ "continuous2", # ensure all continuous variables are treated as such
    missing = "no",
    statistic = list(
      all_continuous() ~ "{mean} ({sd})",  # for continuous variables, display mean (SD), missing value
      all_categorical() ~ "{n} ({p}%)" )) %>%
  add_n(statistic = "{n}") %>%
  as_gt() %>%
  tab_options(
    table.font.size = "x-small"  # Options include "xx-small", "x-small", "small", etc.
  ) %>%
  tab_caption("Summary Statistics across Different Variables")
# missingness in tracheotomy dataset
missing_names <- colnames(tracheo)[colSums(is.na(tracheo)) > 0]
tracheo_missing <- tracheo[,c(missing_names)]
tracheo_miss <- miss_var_summary(tracheo_missing)
tracheo_miss %>%
kable(booktabs = T, escape = T, caption = "Missingness in Tracheostomy Dataset", align = "c", digits =
# missing data table by observations
center1 <- apply(tracheo[tracheo$center==1, !(names(tracheo) %in% c("record_id", "center"))], 2, functi
center2 <- apply(tracheo[tracheo$center==2, !(names(tracheo) %in% c("record_id", "center"))], 2, functi
center3 <- apply(tracheo[tracheo$center==3, !(names(tracheo) %in% c("record_id", "center"))], 2, functi
center4 <- apply(tracheo[tracheo$center==4, !(names(tracheo) %in% c("record_id", "center"))], 2, functi
center7 <- apply(tracheo[tracheo$center==7, !(names(tracheo) %in% c("record_id", "center"))], 2, functi
center12 <- apply(tracheo[tracheo$center==12, !(names(tracheo) %in% c("record_id", "center"))], 2, func
center16 <- apply(tracheo[tracheo$center==16, !(names(tracheo) %in% c("record_id", "center"))], 2, func
center20 <- apply(tracheo[tracheo$center==20, !(names(tracheo) %in% c("record_id", "center"))], 2, func

center_missing <- data.frame(center1, center2, center3, center4, center7, center12, center16, center20)
colnames(center_missing) <- c("Center 1", "Center 2", "Center 3", "Center 4", "Center 7", "Center 12",

center_missing %>%
kable(booktabs = T, escape = T, caption = "Missing Percentage Across Different Centers", align = "c", d
kable_styling(latex_options = c("HOLD_position", "scale_down"))
# Count NA values for each patient
na_count_per_patient <- rowSums(is.na(tracheo))

# Calculate the percentage of NAs for each patient
total_columns <- ncol(tracheo) - 1  # Subtract 1 to exclude the 'record_id' column from the total
percentage_na_per_patient <- (na_count_per_patient / total_columns) * 100

# Create a new data frame with the 'record_id', 'na_count', and 'percentage_na'
na_data <- data.frame(record_id = tracheo$record_id,
                      center = tracheo$center,
                      na_count = na_count_per_patient,
                      percentage_na = percentage_na_per_patient)

na_data %>%
  arrange(desc(na_count)) %>%
  head(5)
```

```r
#Correlation heat map
numeric_columns <- sapply(tracheo, is.numeric)
tracheo_cont <- tracheo[, numeric_columns]
M <-cor(tracheo_cont, use = "complete.obs")
corrplot(M, type="upper", order="hclust",
         col=brewer.pal(n=8, name="RdYlBu"))
# Multiple imputation for 36wks
set.seed(1)
tracheo_36wks <- subset(tracheo, select = -c(record_id, Death, Trach, hosp_dc_ga, weight_today.44, vent

# Create train 70% and test 30% by stratifying outcome proportion
#ignore <- sample(c(TRUE, FALSE), size = 996, replace = TRUE, prob = c(0.3, 0.7))
partition_36wks <- createDataPartition(y = tracheo_36wks$outcome, p = 0.7, list = TRUE)
ignore_36wks <- rep(TRUE, nrow(tracheo_36wks))
ignore_36wks[partition_36wks$Resample1] <- FALSE

tracheo_mice_out_36wks <- mice(select(tracheo_36wks,-center), m=5, seed = 1, ignore = ignore_36wks, pri
#
imp.train_36wks <- filter(tracheo_mice_out_36wks, !ignore_36wks)
imp.test_36wks <- filter(tracheo_mice_out_36wks, ignore_36wks)

# Create train and test imputed sets
tracheo_imp_train_36wks <- vector("list",5)
tracheo_imp_test_36wks <- vector("list",5)
for (i in 1:5) {
  tracheo_imp_train_36wks[[i]] <- mice::complete(imp.train_36wks,i)
  tracheo_imp_test_36wks[[i]] <- mice::complete(imp.test_36wks,i)
  tracheo_imp_train_36wks[[i]]$center <- tracheo$center[which(!ignore_36wks)]
  tracheo_imp_test_36wks[[i]]$center <- tracheo$center[which(ignore_36wks)]
}
# Multiple imputation for 44wks
set.seed(1)
tracheo_sub <- subset(tracheo, select = -c(record_id, Death, Trach, hosp_dc_ga))

# Create train 70% and test 30% by stratifying outcome proportion
# ignore <- sample(c(TRUE, FALSE), size = 996, replace = TRUE, prob = c(0.3, 0.7))
partition <- createDataPartition(y = tracheo_sub$outcome, p = 0.7, list = TRUE)
ignore <- rep(TRUE, nrow(tracheo_sub))
ignore[partition$Resample1] <- FALSE

tracheo_mice_out <- mice(select(tracheo_sub,-center), m=5, seed = 1, ignore = ignore, print=F)

imp.train <- filter(tracheo_mice_out, !ignore)
imp.test <- filter(tracheo_mice_out, ignore)

# Create train and test imputed sets
tracheo_imp_train <- vector("list",5)
tracheo_imp_test <- vector("list",5)
for (i in 1:5) {
  tracheo_imp_train[[i]] <- mice::complete(imp.train,i)
  tracheo_imp_test[[i]] <- mice::complete(imp.test,i)
  tracheo_imp_train[[i]]$center <- tracheo$center[which(!ignore)]
  tracheo_imp_test[[i]]$center <- tracheo$center[which(ignore)]
```

```
}
#######################################################
#### Lasso Model####
#######################################################

lasso <- function(df, modeltype, include.interact) {
  #' Runs 10-fold CV for lasso and returns corresponding coefficients
  #' @param df, data set
  #' @return coef, coefficients for minimum cv error

  # Matrix form for ordered variables

  df <- df %>%
    select(-center)

  if (modeltype == 44 & include.interact==TRUE) {
  x.ord <- model.matrix(outcome~.+ ventilation_support_level.36*inspired_oxygen.36 + ventilation_suppor
  } else if(modeltype == 36 & include.interact==TRUE){
    x.ord <- model.matrix(outcome~.+ ventilation_support_level.36*inspired_oxygen.36 + ventilation_suppo
  }
  else{
  x.ord <- model.matrix(outcome~., data = df)[,-1]
  }
  y.ord <- df$outcome

  # Generate folds
  k <- 10
  set.seed(1) # consistent seeds between imputed data sets
  folds <- sample(1:k, nrow(df), replace=TRUE)

  # Lasso model
  lasso_mod <- cv.glmnet(x.ord, y.ord, nfolds = 10, foldid = folds,
                         alpha = 1, family = "binomial")

  lasso_ml <- glmnet(x.ord, y.ord, lambda = lasso_mod$lambda.min, alpha = 1, family = "binomial")
  # Get coefficients
  coef <- coef(lasso_ml, lambda = lasso_ml$lambda.min)
  return(coef)
}
# average coefficients across five imputed datasets for 36wks without interactions
lasso_coef1_36wks <- lasso(tracheo_imp_train_36wks[[1]], modeltype = 36, include.interact = FALSE)
lasso_coef2_36wks <- lasso(tracheo_imp_train_36wks[[2]], modeltype = 36, include.interact = FALSE)
lasso_coef3_36wks <- lasso(tracheo_imp_train_36wks[[3]], modeltype = 36, include.interact = FALSE)
lasso_coef4_36wks <- lasso(tracheo_imp_train_36wks[[4]], modeltype = 36, include.interact = FALSE)
lasso_coef5_36wks <- lasso(tracheo_imp_train_36wks[[5]], modeltype = 36, include.interact = FALSE)

lasso_coef_36wks <- cbind(lasso_coef1_36wks, lasso_coef2_36wks, lasso_coef3_36wks,
                   lasso_coef4_36wks, lasso_coef5_36wks)
avg_coefs_lasso_36wks <- apply(lasso_coef_36wks, 1, mean)
lasso1_zero_names_36wks <- rownames(lasso_coef1_36wks)[as.numeric(lasso_coef1_36wks)== 0]
lasso2_zero_names_36wks <- rownames(lasso_coef2_36wks)[as.numeric(lasso_coef2_36wks)== 0]
lasso3_zero_names_36wks <- rownames(lasso_coef3_36wks)[as.numeric(lasso_coef3_36wks)== 0]
lasso4_zero_names_36wks <- rownames(lasso_coef4_36wks)[as.numeric(lasso_coef4_36wks)== 0]
```

```r
lasso5_zero_names_36wks <- rownames(lasso_coef5_36wks)[as.numeric(lasso_coef5_36wks)== 0]

# create the frequency of zero coefficients table across five imputed sets
all_zero_names_36wks <- c(lasso1_zero_names_36wks, lasso2_zero_names_36wks, lasso3_zero_names_36wks, la
allzerostable_36wks <- table(all_zero_names_36wks)
# 36wks model without interactions
lasso_model_36wks_no_int <-avg_coefs_lasso_36wks[!names(avg_coefs_lasso_36wks) %in% c("blength", "com_p
# average coefficients across five imputed datasets for 36wks with interactions
lasso_int_coef1_36wks <- lasso(tracheo_imp_train_36wks[[1]], modeltype = 36, include.interact = TRUE)
lasso_int_coef2_36wks <- lasso(tracheo_imp_train_36wks[[2]], modeltype = 36, include.interact = TRUE)
lasso_int_coef3_36wks <- lasso(tracheo_imp_train_36wks[[3]], modeltype = 36, include.interact = TRUE)
lasso_int_coef4_36wks <- lasso(tracheo_imp_train_36wks[[4]], modeltype = 36, include.interact = TRUE)
lasso_int_coef5_36wks <- lasso(tracheo_imp_train_36wks[[5]], modeltype = 36, include.interact = TRUE)

lasso_int_coef_36wks <- cbind(lasso_int_coef1_36wks, lasso_int_coef2_36wks, lasso_int_coef3_36wks,
                    lasso_int_coef4_36wks, lasso_int_coef5_36wks)
avg_coefs_lasso_int_36wks <- apply(lasso_int_coef_36wks, 1, mean)
lasso1_zero_names_36wks <- rownames(lasso_int_coef1_36wks)[as.numeric(lasso_int_coef1_36wks)== 0]
lasso2_zero_names_36wks <- rownames(lasso_int_coef2_36wks)[as.numeric(lasso_int_coef2_36wks)== 0]
lasso3_zero_names_36wks <- rownames(lasso_int_coef3_36wks)[as.numeric(lasso_int_coef3_36wks)== 0]
lasso4_zero_names_36wks <- rownames(lasso_int_coef4_36wks)[as.numeric(lasso_int_coef4_36wks)== 0]
lasso5_zero_names_36wks <- rownames(lasso_int_coef5_36wks)[as.numeric(lasso_int_coef5_36wks)== 0]

# Create the frequency of zero coefficients table across five imputed sets
all_zero_names_36wks <- c(lasso1_zero_names_36wks, lasso2_zero_names_36wks, lasso3_zero_names_36wks, las
allzerostable_36wks <- table(all_zero_names_36wks)
# 36wks model with interactions
lasso_model_36wks_int <- avg_coefs_lasso_int_36wks[!names(avg_coefs_lasso_int_36wks) %in% c("blength", "
# average coefficients across five imputed datasets for 44wks without interactions
lasso_coef1 <- lasso(tracheo_imp_train[[1]], modeltype = 44, include.interact = FALSE)
lasso_coef2 <- lasso(tracheo_imp_train[[2]], modeltype = 44, include.interact = FALSE)
lasso_coef3 <- lasso(tracheo_imp_train[[3]], modeltype = 44, include.interact = FALSE)
lasso_coef4 <- lasso(tracheo_imp_train[[4]], modeltype = 44, include.interact = FALSE)
lasso_coef5 <- lasso(tracheo_imp_train[[5]], modeltype = 44, include.interact = FALSE)

lasso_coef <- cbind(lasso_coef1, lasso_coef2, lasso_coef3,
                    lasso_coef4, lasso_coef5)
avg_coefs_lasso <- apply(lasso_coef, 1, mean)
lasso1_zero_names <- rownames(lasso_coef1)[as.numeric(lasso_coef1)== 0]
lasso2_zero_names <- rownames(lasso_coef2)[as.numeric(lasso_coef2)== 0]
lasso3_zero_names <- rownames(lasso_coef3)[as.numeric(lasso_coef3)== 0]
lasso4_zero_names <- rownames(lasso_coef4)[as.numeric(lasso_coef4)== 0]
lasso5_zero_names <- rownames(lasso_coef5)[as.numeric(lasso_coef5)== 0]

# Create the frequency of zero coefficients table across five imputed sets
all_zero_names <- c(lasso1_zero_names, lasso2_zero_names, lasso3_zero_names, lasso4_zero_names, lasso5_z
allzerostable <- table(all_zero_names)
# 44wks model without interactions
lasso_model_no_int <- avg_coefs_lasso[!names(avg_coefs_lasso) %in% c("blength", "ga", "genderMale", "mat
# average coefficients across five imputed datasets for 44wks with interactions
lasso_int_coef1 <- lasso(tracheo_imp_train[[1]], modeltype = 44, include.interact = TRUE)
lasso_int_coef2 <- lasso(tracheo_imp_train[[2]], modeltype = 44, include.interact = TRUE)
lasso_int_coef3 <- lasso(tracheo_imp_train[[3]], modeltype = 44, include.interact = TRUE)
```

```r
lasso_int_coef4 <- lasso(tracheo_imp_train[[4]], modeltype = 44, include.interact = TRUE)
lasso_int_coef5 <- lasso(tracheo_imp_train[[5]], modeltype = 44, include.interact = TRUE)

lasso_int_coef <- cbind(lasso_int_coef1, lasso_int_coef2, lasso_int_coef3,
                        lasso_int_coef4, lasso_int_coef5)
avg_coefs_lasso_int <- apply(lasso_int_coef, 1, mean)
lasso1_zero_names <- rownames(lasso_int_coef1)[as.numeric(lasso_int_coef1)== 0]
lasso2_zero_names <- rownames(lasso_int_coef2)[as.numeric(lasso_int_coef2)== 0]
lasso3_zero_names <- rownames(lasso_int_coef3)[as.numeric(lasso_int_coef3)== 0]
lasso4_zero_names <- rownames(lasso_int_coef4)[as.numeric(lasso_int_coef4)== 0]
lasso5_zero_names <- rownames(lasso_int_coef5)[as.numeric(lasso_int_coef5)== 0]

all_zero_names <- c(lasso1_zero_names, lasso2_zero_names, lasso3_zero_names, lasso4_zero_names, lasso5_
allzerostable <- table(all_zero_names)
# 44wks model with interactions
lasso_model_int <- avg_coefs_lasso_int[!names(avg_coefs_lasso_int) %in% c("blength", "ga", "genderMale"
bestsubset <- function(df, include.interact) {
  if (include.interact==TRUE) {
  x.ord <- model.matrix(outcome~.+ ventilation_support_level.36*inspired_oxygen.36 + ventilation_suppor
  } else{
  x.ord <- model.matrix(outcome~., data = df)[,-1]
  }
  y.ord <- df$outcome
  cv_subset <- L0Learn.cvfit(
  x.ord,
  y.ord,
  seed = 1,
  loss = "Logistic",
  penalty = "L0",
  nFolds = 10,
  intercept = TRUE)
  best_cvmeans <- which.min(cv_subset$cvMeans[[1]])
  coef <- c(cv_subset$fit$a0[[1]][best_cvmeans], cv_subset$fit$beta[[1]][,best_cvmeans])
  return(coef)
}
subset_coef1 <-bestsubset(tracheo_imp_train[[1]], include.interact = FALSE)
subset_coef2 <-bestsubset(tracheo_imp_train[[2]], include.interact = FALSE)
subset_coef3 <-bestsubset(tracheo_imp_train[[3]], include.interact = FALSE)
subset_coef4 <-bestsubset(tracheo_imp_train[[4]], include.interact = FALSE)
subset_coef5 <-bestsubset(tracheo_imp_train[[5]], include.interact = FALSE)

subset_coef <- cbind(subset_coef1, subset_coef2, subset_coef3, subset_coef4, subset_coef5)
avg_coefs_subset <- apply(subset_coef, 1, mean)

subset1_zero_names <- names(avg_coefs_lasso)[which(subset_coef1 == 0)]
subset2_zero_names <- names(avg_coefs_lasso)[which(subset_coef2 == 0)]
subset3_zero_names <- names(avg_coefs_lasso)[which(subset_coef2 == 0)]
subset4_zero_names <- names(avg_coefs_lasso)[which(subset_coef3 == 0)]
subset5_zero_names <- names(avg_coefs_lasso)[which(subset_coef4 == 0)]

all_zero_names <- c(subset1_zero_names, subset2_zero_names, subset3_zero_names, subset4_zero_names, subs
names(avg_coefs_subset) <- names(avg_coefs_lasso)
allzerostable <- table(all_zero_names)
```

```r
subset_int_coef1 <-bestsubset(tracheo_imp_train[[1]], include.interact = TRUE)
subset_int_coef2 <-bestsubset(tracheo_imp_train[[2]], include.interact = TRUE)
subset_int_coef3 <-bestsubset(tracheo_imp_train[[3]], include.interact = TRUE)
subset_int_coef4 <-bestsubset(tracheo_imp_train[[4]], include.interact = TRUE)
subset_int_coef5 <-bestsubset(tracheo_imp_train[[5]], include.interact = TRUE)

subset_int_coef <- cbind(subset_int_coef1, subset_int_coef2, subset_int_coef3, subset_int_coef4, subset_
avg_coefs_subset_int <- apply(subset_int_coef, 1, mean)

subset1_zero_names <- names(avg_coefs_lasso_int)[which(subset_int_coef1 == 0)]
subset2_zero_names <- names(avg_coefs_lasso_int)[which(subset_int_coef2 == 0)]
subset3_zero_names <- names(avg_coefs_lasso_int)[which(subset_int_coef2 == 0)]
subset4_zero_names <- names(avg_coefs_lasso_int)[which(subset_int_coef3 == 0)]
subset5_zero_names <- names(avg_coefs_lasso_int)[which(subset_int_coef4 == 0)]

all_zero_names <- c(subset1_zero_names, subset2_zero_names, subset3_zero_names, subset4_zero_names, subs
names(avg_coefs_subset_int) <- names(avg_coefs_lasso_int)
allzerostable <- table(all_zero_names)
# Final coefficients table for 36 week models
week36_int <- c("blength", "com_prenat_sterYes", "genderMale", "ventilation_support_level.361", "ventila
avg_coefs_36wks_int_table <- avg_coefs_lasso_int_36wks
avg_coefs_36wks_int_table[names(avg_coefs_36wks_int_table) %in% week36_int] <- 0


week36_no_int <- c("blength", "com_prenat_sterYes", "ventilation_support_level.361", "bw", "ga", "birth_
avg_coefs_36wks_no_int_table <- avg_coefs_lasso_36wks
avg_coefs_36wks_no_int_table[names(avg_coefs_36wks_no_int_table) %in% week36_no_int] <- 0
avg_coefs_36wks_no_int_table["ventilation_support_level.361:inspired_oxygen.36"] <- 0
avg_coefs_36wks_no_int_table["ventilation_support_level.362:inspired_oxygen.36"] <- 0
avg_coefs_36wks_no_int_table["ventilation_support_level.361:med_ph.361"] <- 0
avg_coefs_36wks_no_int_table["ventilation_support_level.362:med_ph.361"] <- 0

coeff_df <- round(data.frame(coefs_lasso_no_int = avg_coefs_36wks_no_int_table, coefs_36wks_int = avg_co
coeff_df[coeff_df == 0] <- "-"
rownames(coeff_df) <- c("(Intercept)", "Maternal Ethnicity", "Birth weight (g)", "Obstetrical gestationa
colnames(coeff_df) <- c("36 week model", "36 week model with interactions")

coeff_df %>%
kable(caption = "36 Week models Regression Coefficient Comparisons", align = "l", booktabs = T, escape =
kable_styling(font_size = 8, latex_options = c("HOLD_position", "scale_down"))
# Final coefficients table for 44 week models
week44_int <- c("blength", "ga", "genderMale", "mat_chorioYes", "ventilation_support_level_modified.441"
avg_coefs_44wks_int_table <- avg_coefs_lasso_int
avg_coefs_44wks_int_table[names(avg_coefs_44wks_int_table) %in% week44_int] <- 0


week44_no_int <- c("blength", "ga", "genderMale", "mat_chorioYes", "med_ph.361", "sgaSGA", "ventilation_
avg_coefs_44wks_no_int_table <- avg_coefs_lasso
avg_coefs_44wks_no_int_table[names(avg_coefs_44wks_no_int_table) %in% week44_no_int] <- 0
avg_coefs_44wks_no_int_table["ventilation_support_level.361:inspired_oxygen.36"] <- 0
avg_coefs_44wks_no_int_table["ventilation_support_level.362:inspired_oxygen.36"] <- 0
avg_coefs_44wks_no_int_table["ventilation_support_level.361:med_ph.361"] <- 0
avg_coefs_44wks_no_int_table["ventilation_support_level.362:med_ph.361"] <- 0
```

```r
avg_coefs_44wks_no_int_table["ventilation_support_level_modified.441:inspired_oxygen.44"] <- 0
avg_coefs_44wks_no_int_table["ventilation_support_level_modified.442:inspired_oxygen.44"] <- 0
avg_coefs_44wks_no_int_table["ventilation_support_level_modified.441:med_ph.441"] <- 0
avg_coefs_44wks_no_int_table["ventilation_support_level_modified.442:med_ph.441"] <- 0

coeff_df <- round(data.frame(coefs_44wks_no_int = avg_coefs_44wks_no_int_table, coefs_44wks_int = avg_co
coeff_df[coeff_df == 0] <- "-"
rownames(coeff_df) <- c("(Intercept)", "Maternal Ethnicity", "Birth weight (g)", "Obstetrical gestationa
colnames(coeff_df) <- c("44 week model", "44 week model with interactions")

coeff_df %>%
kable(caption = "44 Week models Regression Coefficient Comparisons", align = "l", booktabs = T, escape
kable_styling(font_size = 8, latex_options = c("HOLD_position", "scale_down"))
# Performance Metrics Calculation for 36 Week Model without Interactions
auc_lasso_36wks <- numeric(5)
sensitivity_lasso_36wks <- numeric(5)
specificity_lasso_36wks <- numeric(5)
threshold_lasso_36wks <- numeric(5)
BrierScore_lasso_36wks <- numeric(5)

for (i in 1:5) {
x_test <- model.matrix(outcome~., data=tracheo_imp_test_36wks[[i]])
y_test <- tracheo_imp_test_36wks[[i]]$outcome

qq <- x_test[,colnames(x_test) %in% names(lasso_model_36wks_no_int)]
qq <- as.matrix(qq)
dimnames(qq) <- NULL
pred <- qq %*% c(as.numeric(lasso_model_36wks_no_int))
predprob_lasso <- 1 / (1 + exp(-pred))

roc_info <- roc(predictor=as.numeric(predprob_lasso),
                response=y_test)
roc_coord <- coords(roc = roc_info, x="best")
auc_lasso_36wks[i] <- auc(roc_info)
sensitivity_lasso_36wks[i] <- roc_coord$sensitivity
specificity_lasso_36wks[i] <- roc_coord$specificity
threshold_lasso_36wks[i] <- roc_coord$threshold
BrierScore_lasso_36wks[i] <- BrierScore(as.numeric(as.character(y_test)), predprob_lasso)
}

auc_lasso_36wks <-  mean(auc_lasso_36wks)
sensitivity_lasso_36wks <- mean(sensitivity_lasso_36wks)
specificity_lasso_36wks <- mean(specificity_lasso_36wks)
threshold_lasso_36wks <- mean(threshold_lasso_36wks)
BrierScore_lasso_36wks <- mean(BrierScore_lasso_36wks)
# Performance Metrics Calculation for 36 Week Model with Interactions
auc_lasso_36wks_int <- numeric(5)
sensitivity_lasso_36wks_int <- numeric(5)
specificity_lasso_36wks_int <- numeric(5)
threshold_lasso_36wks_int <- numeric(5)
BrierScore_lasso_36wks_int <- numeric(5)

for (i in 1:5) {
```

```r
x_test <- model.matrix(outcome~.+ ventilation_support_level.36*inspired_oxygen.36 + ventilation_support_
y_test <- tracheo_imp_test_36wks[[i]]$outcome

qq <- x_test[,colnames(x_test) %in% names(lasso_model_36wks_int)]
qq <- as.matrix(qq)
dimnames(qq) <- NULL
pred <- qq %*% c(as.numeric(lasso_model_36wks_int))
predprob_lasso <- 1 / (1 + exp(-pred))

roc_info <- roc(predictor=as.numeric(predprob_lasso),
                response=y_test)
roc_coord <- coords(roc = roc_info, x="best")
auc_lasso_36wks_int[i] <- auc(roc_info)
sensitivity_lasso_36wks_int[i] <- roc_coord$sensitivity
specificity_lasso_36wks_int[i] <- roc_coord$specificity
threshold_lasso_36wks_int[i] <- roc_coord$threshold
BrierScore_lasso_36wks_int[i] <- BrierScore(as.numeric(as.character(y_test)), predprob_lasso)
}
auc_lasso_36wks_int <-  mean(auc_lasso_36wks_int)
sensitivity_lasso_36wks_int <- mean(sensitivity_lasso_36wks_int)
specificity_lasso_36wks_int <- mean(specificity_lasso_36wks_int)
threshold_lasso_36wks_int <- mean(threshold_lasso_36wks_int)
BrierScore_lasso_36wks_int <- mean(BrierScore_lasso_36wks_int)
# Performance Metrics Calculation for 44 Week Model without Interactions
auc_lasso <- numeric(5)
sensitivity_lasso <- numeric(5)
specificity_lasso <- numeric(5)
threshold_lasso <- numeric(5)
BrierScore_lasso <- numeric(5)

for (i in 1:5) {
x_test <- model.matrix(outcome~., data=tracheo_imp_test[[i]])
y_test <- tracheo_imp_test[[i]]$outcome

qq <- x_test[,colnames(x_test) %in% names(lasso_model_no_int)]
qq <- as.matrix(qq)
dimnames(qq) <- NULL
pred <- qq %*% c(as.numeric(lasso_model_no_int))
predprob_lasso <- 1 / (1 + exp(-pred))

roc_info <- roc(predictor=as.numeric(predprob_lasso),
                response=y_test)
roc_coord <- coords(roc = roc_info, x="best")
auc_lasso[i] <- auc(roc_info)
sensitivity_lasso[i] <- roc_coord$sensitivity
specificity_lasso[i] <- roc_coord$specificity
threshold_lasso[i] <- roc_coord$threshold
BrierScore_lasso[i] <- BrierScore(as.numeric(as.character(y_test)), predprob_lasso)
}

auc_lasso <-  mean(auc_lasso)
sensitivity_lasso <- mean(sensitivity_lasso)
specificity_lasso <- mean(specificity_lasso)
```

```r
threshold_lasso <- mean(threshold_lasso)
BrierScore_lasso <- mean(BrierScore_lasso)
# Performance Metrics Calculation for 44 Week Model with Interactions
auc_lasso_int <- numeric(5)
sensitivity_lasso_int <- numeric(5)
specificity_lasso_int <- numeric(5)
threshold_lasso_int <- numeric(5)
BrierScore_lasso_int <- numeric(5)

for (i in 1:5) {
x_test <- model.matrix(outcome~.+ ventilation_support_level.36*inspired_oxygen.36 + ventilation_support
y_test <- tracheo_imp_test[[i]]$outcome

qq <- x_test[,colnames(x_test) %in% names(lasso_model_int)]
qq <- as.matrix(qq)
dimnames(qq) <- NULL
pred <- qq %*% c(as.numeric(lasso_model_int))
predprob_lasso <- 1 / (1 + exp(-pred))

roc_info <- roc(predictor=as.numeric(predprob_lasso),
                response=y_test)
roc_coord <- coords(roc = roc_info, x="best")
auc_lasso_int[i] <- auc(roc_info)
sensitivity_lasso_int[i] <- roc_coord$sensitivity
specificity_lasso_int[i] <- roc_coord$specificity
threshold_lasso_int[i] <- roc_coord$threshold
BrierScore_lasso_int[i] <- BrierScore(as.numeric(as.character(y_test)), predprob_lasso)
}

auc_lasso_int <- mean(auc_lasso_int)
sensitivity_lasso_int <- mean(sensitivity_lasso_int)
specificity_lasso_int <- mean(specificity_lasso_int)
threshold_lasso_int <- mean(threshold_lasso_int)
BrierScore_lasso_int <- mean(BrierScore_lasso_int)
# Best subset code (REFERENCE ONLY)
auc_subsetlogistic <- numeric(5)
sensitivity_subsetlogistic <- numeric(5)
specificity_subsetlogistic <- numeric(5)
threshold_subsetlogistic <- numeric(5)
BrierScore_subsetlogistic <- numeric(5)

for (i in 1:5) {
x_test <- model.matrix(outcome~., data=tracheo_imp_test[[i]])
y_test <- tracheo_imp_test[[i]]$outcome

qq <- x_test[,colnames(x_test) %in% names(subsetlogistic_no_int)]
qq <- as.matrix(qq)
dimnames(qq) <- NULL
pred <- qq %*% c(as.numeric(subsetlogistic_no_int))
predprob_subsetlogistic <- 1 / (1 + exp(-pred))

roc_info <- roc(predictor=as.numeric(predprob_subsetlogistic),
                response=y_test)
```

```r
roc_coord <- coords(roc = roc_info, x="best")
auc_subsetlogistic[i] <- auc(roc_info)
sensitivity_subsetlogistic[i] <- roc_coord$sensitivity
specificity_subsetlogistic[i] <- roc_coord$specificity
threshold_subsetlogistic[i] <- roc_coord$threshold
BrierScore_subsetlogistic[i] <- BrierScore(as.numeric(as.character(y_test)), predprob_subsetlogistic)
}

auc_subsetlogistic <-  mean(auc_subsetlogistic)
sensitivity_subsetlogistic <- mean(sensitivity_subsetlogistic)
specificity_subsetlogistic <- mean(specificity_subsetlogistic)
threshold_subsetlogistic <- mean(threshold_subsetlogistic)
BrierScore_subsetlogistic <- mean(BrierScore_subsetlogistic)

auc_subsetlogistic_int <- numeric(5)
sensitivity_subsetlogistic_int <- numeric(5)
specificity_subsetlogistic_int <- numeric(5)
threshold_subsetlogistic_int <- numeric(5)
BrierScore_subsetlogistic_int <- numeric(5)

for (i in 1:5) {
x_test <- model.matrix(outcome~.+ ventilation_support_level.36*inspired_oxygen.36 + ventilation_support_
y_test <- tracheo_imp_test[[i]]$outcome

qq <- x_test[,colnames(x_test) %in% names(subsetlogistic_int)]
qq <- as.matrix(qq)
dimnames(qq) <- NULL
pred <- qq %*% c(as.numeric(subsetlogistic_int))
predprob_subsetlogistic <- 1 / (1 + exp(-pred))

roc_info <- roc(predictor=as.numeric(predprob_subsetlogistic),
                response=y_test)
roc_coord <- coords(roc = roc_info, x="best")
auc_subsetlogistic_int[i] <- auc(roc_info)
sensitivity_subsetlogistic_int[i] <- roc_coord$sensitivity
specificity_subsetlogistic_int[i] <- roc_coord$specificity
threshold_subsetlogistic_int[i] <- roc_coord$threshold
BrierScore_subsetlogistic_int[i] <- BrierScore(as.numeric(as.character(y_test)), predprob_subsetlogistic
}

auc_subsetlogistic_int <-  mean(auc_subsetlogistic_int)
sensitivity_subsetlogistic_int <- mean(sensitivity_subsetlogistic_int)
specificity_subsetlogistic_int <- mean(specificity_subsetlogistic_int)
threshold_subsetlogistic_int <- mean(threshold_subsetlogistic_int)
BrierScore_subsetlogistic_int <- mean(BrierScore_subsetlogistic_int)
# Model evaluation measures table
metrics_table <- data.frame(Measures = c("AUC", "Sensitivity", "Specificity","Threshold", "BrierScore")
                            weeks36 = round(c(auc_lasso_36wks, sensitivity_lasso_36wks, specificity_las
                            weeks36_with_interaction = round(c(auc_lasso_36wks_int, sensitivity_lasso_3
                            weeks44 = round(c(auc_lasso, sensitivity_lasso, specificity_lasso, threshol
                            weeks44_with_interaction = round(c(auc_lasso_int, sensitivity_lasso_int, sp
          )
colnames(metrics_table) <- c(" ", "36 Week Model", "36 Week Model with Interactions", "44 Week Model",
```

```
kable(metrics_table, caption = "Model Evaluation Measures", align = "l",digits = 3)  %>%
  kable_styling(latex_options = c("HOLD_position", "scale_down"))
#Frequency of Zero Coefficients Across Five Imputed Sets for 36 Week Models
t1 <- rowSums(as.matrix(lasso_coef_36wks)==0)
t1["ventilation_support_level.361:inspired_oxygen.36"] <- "-"
t1["ventilation_support_level.362:inspired_oxygen.36"] <- "-"
t1["ventilation_support_level.361:med_ph.361"] <- "-"
t1["ventilation_support_level.362:med_ph.361"] <- "-"
zero_freq <- data.frame(lasso = t1, week36 = rowSums(as.matrix(lasso_int_coef_36wks)==0))
colnames(zero_freq) <- c("36 Week Model", "36 Week Model with Interactions")
zero_freq %>%
kable(booktabs = T, escape = T, caption = "Frequency of Zero Coefficients Across Five Imputed Sets for 3
kable_styling(font_size = 8, latex_options = c("HOLD_position", "scale_down"))
#Frequency of Zero Coefficients in 44 Week Models Across Five Imputed Sets
t2 <- rowSums(as.matrix(lasso_coef)==0)
t2["ventilation_support_level.361:inspired_oxygen.36"] <- "-"
t2["ventilation_support_level.362:inspired_oxygen.36"] <- "-"
t2["ventilation_support_level.361:med_ph.361"] <- "-"
t2["ventilation_support_level.362:med_ph.361"] <- "-"
t2["ventilation_support_level_modified.441:inspired_oxygen.44"] <- "-"
t2["ventilation_support_level_modified.442:inspired_oxygen.44"] <- "-"
t2["ventilation_support_level_modified.441:med_ph.441"] <- "-"
t2["ventilation_support_level_modified.442:med_ph.441"] <- "-"
zero_freq <- data.frame(lasso = t2, week36 = rowSums(as.matrix(lasso_int_coef)==0))
colnames(zero_freq) <- c("44 Week Model", "44 Week Model with Interactions")
zero_freq %>%
kable(booktabs = T, escape = T,caption = "Frequency of Zero Coefficients in 44 Week Models Across Five
kable_styling(font_size = 8, latex_options = c("HOLD_position", "scale_down"))
```