PHP 2550 Project 2

Himashree Chandru

2023-11-14

Abstract

This study, conducted in collaboration with the BPD Collaborative Registry, aimed to develop a predictive model for tracheostomy placement in infants with severe bronchopulmonary dysplasia (sBPD). Analyzing data from diverse neonatal care centers, the study incorporated demographic, birth, and respiratory variables to capture the complexity of decision-making in neonatal care.

Preprocessing efforts addressed data inconsistencies, and multiple imputation techniques handled missing data, resulting in a robust dataset. Lasso models, chosen for their ability to handle complex relationships and promote variable selection, were employed to develop predictive models at 36 and 44 weeks postmenstrual age.

The evaluation of these models revealed consistent superiority in performance for the 44 Week Model. Metrics such as precision, recall, specificity, F score, and the Brier Score consistently favored the 44 Week Model, signifying its heightened ability to predict tracheostomy placement accurately. This model not only demonstrated superior predictive power but also achieved a more balanced trade-off between precision and recall.

In conclusion, this study provides valuable insights into the predictive modeling of tracheostomy placement, offering a nuanced understanding of factors influencing clinical decisions in neonatal care. The results hold potential for enhancing patient outcomes and informing counseling practices in the challenging realm of sBPD management.

Introduction

Severe bronchopulmonary dysplasia (sBPD) poses a complex challenge in neonatal care, necessitating nuanced decision-making regarding tracheostomy placement. The lack of clear indications and optimal timing prompted this collaboration with Dr. Chris Schmid from the Biostatistics Department. The main aim of this project is to develop a predictive model capable of predicting the likelihood of tracheostomy based on demographic variables, birth variables and respiratory variables measured at different time points (36 and 44 weeks PMA), such that it results in potentially improving patient outcomes and informing counseling practices.

Data Overview

The dataset originates from the BPD Collaborative Registry, a multi-center consortium of interdisciplinary BPD programs across the United States and Sweden. It includes infants with gestational age below 32 weeks and diagnosed with sBPD according to the 2001 NHLBI criteria. Standard demographic and clinical data are collected at four critical time points: birth, 36 weeks postmenstrual age (PMA), 44 weeks PMA, and discharge. The dataset, covering the period between January 1 and July 19, 2021, was queried for patients with sBPD and complete growth data. Ten BPD Collaborative centers contributed data meeting the study's inclusion criteria.

The data analyzed here contained different types of variables, and consisted of 996 observations of 30 variables. Patient ID and center number were included. The demographic variables included were maternal race and ethnicity (1 = Hispanic/Latino, 2 = Not Hispanic or Latino). The birth variables included were birth weight,

obstetrical gestational age, whether the infant was small for gestational age, birth length, birth head circumference, delivery method (1 = Vaginal Delivery, 2 = Cesarean Section), whether prenatal corticosteroids were administered, whether complete prenatal steroids were administered, whether maternal chorioamnionitis was present, gender, and whether the infant received any surfactant at any point in the first 72 hours. The weight of the infant, and respiratory support variables such as ventilation support (0 = No respiratory support or supplemental oxygen, 1 = Non-invasive positive pressure, 2 = Invasive positive pressure), fraction of inspired oxygen, peak inspiratory pressure, positive end exploratory pressure, and whether medication for pulmonary hypertension was administered, were recorded at both 36 and 44 weeks. In addition, data on the infant's gestational age at the time of discharge and whether tracheostomy (0 = No, 1 = Yes) or death had occurred at this point were recorded.

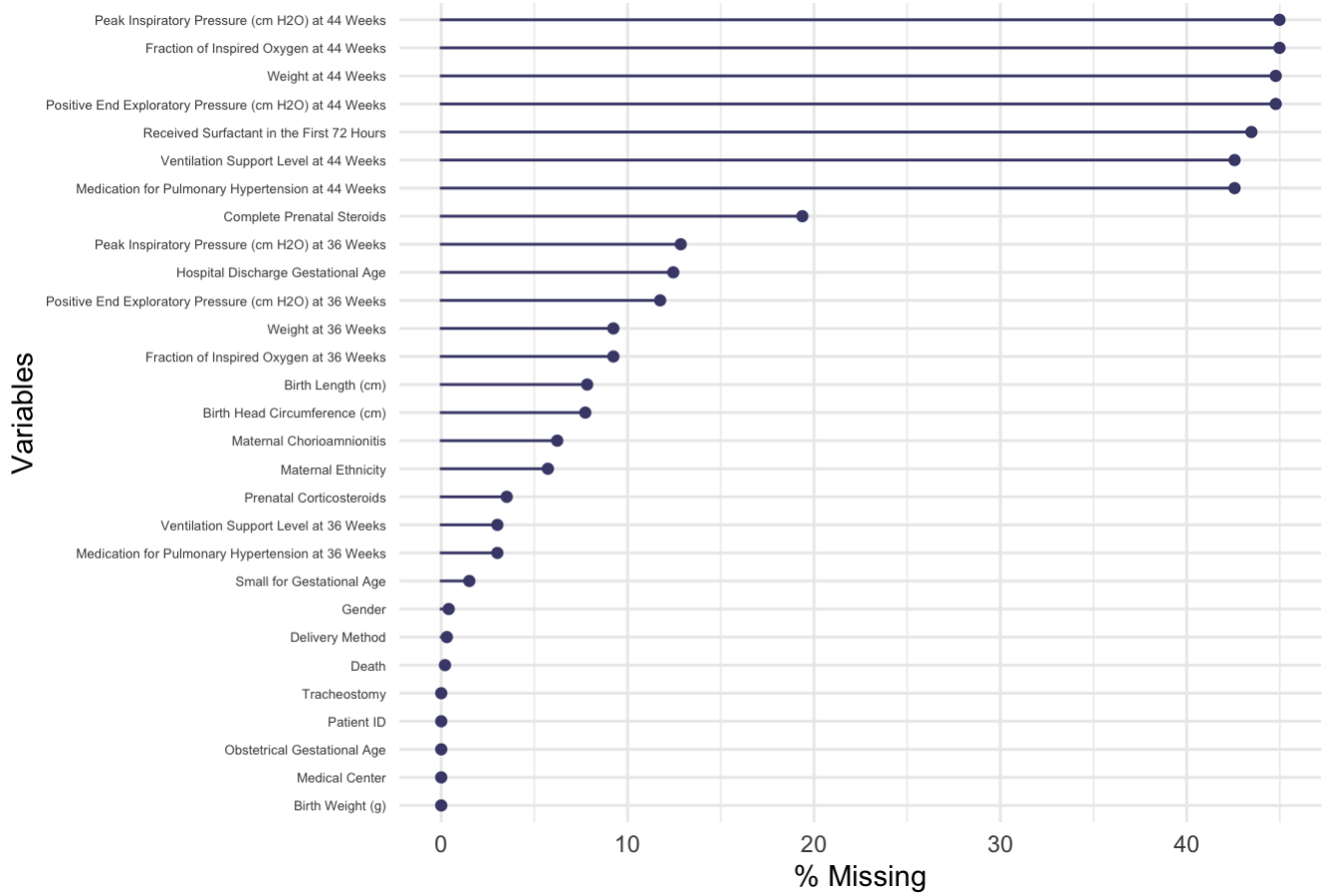
Exploratory analysis

Preprocessing

Due to coding inconsistencies in the maternal race variable, it was excluded from the analysis. The levels of the variable corresponding to whether the infant was small for gestational age were recoded to Yes/No from SGA/Not SGA. In addition, the observation corresponding to Patient ID 2000824 was duplicated 3 times, so these duplicates were removed. Any missing values for the center variable were filled in using the Patient ID number for that observation. The categorical variables were converted to factors to help with ease of analysis. Variable labels for all variables were also created.

Missing Data

Figure 1. Percent Missing Observations by Variable



From exploring the missing data (Figure 1), it can be seen that all of the 44 week variables have about 40-45% of their values missing. This may be due to a combination of reasons - different centers might collect different data, and the missingness might also be coming from the infants that were discharged before 44 weeks. Other variables with large percentages of missingness are the variable recording whether the infant received surfactant at any point in the first 72 hours (~45% missing), and the variable recording whether complete prenatal steroids were administered (~20%).

Descriptive Statistics

Overall descriptive statistics (Table 1) and descriptive statistics by center (Table 2) were computed.

Table 1. Overall Descriptive Statistics

Characteristic	**N = 996**[1]
Medical Center	
1	65 (6.5%)
2	630 (63%)
3	57 (5.7%)
4	60 (6.0%)
5	40 (4.0%)
7	32 (3.2%)
12	69 (6.9%)
16	38 (3.8%)
20	4 (0.4%)

[1] n (%); Median (IQR)

Table 1. Overall Descriptive Statistics

Characteristic	N = 996[1]
21	1 (0.1%)
Maternal Ethnicity	
1	74 (7.9%)
2	865 (92%)
Birth Weight (g)	745 (600, 936)
Obstetrical Gestational Age	25.00 (24.00, 27.00)
Birth Length (cm)	32.0 (30.0, 35.0)
Birth Head Circumference (cm)	23.00 (21.05, 24.78)
Delivery Method	
1	285 (29%)
2	708 (71%)
Prenatal Corticosteroids	835 (87%)
Complete Prenatal Steroids	610 (76%)
Maternal Chorioamnionitis	160 (17%)
Gender	
Female	408 (41%)
Male	584 (59%)
Small for Gestational Age	203 (21%)
Received Surfactant in the First 72 Hours	461 (82%)
Weight at 36 Weeks	2,130 (1,856, 2,400)
Ventilation Support Level at 36 Weeks	
0	117 (12%)
1	589 (61%)
2	260 (27%)
Fraction of Inspired Oxygen at 36 Weeks	0.30 (0.24, 0.38)
Peak Inspiratory Pressure (cm H2O) at 36 Weeks	0 (0, 8)
Positive End Exploratory Pressure (cm H2O) at 36 Weeks	7.0 (6.0, 8.0)
Medication for Pulmonary Hypertension at 36 Weeks	
0	900 (93%)
1	66 (6.8%)
Weight at 44 Weeks	3,700 (3,241, 4,115)
Ventilation Support Level at 44 Weeks	
0	269 (47%)
1	146 (26%)
2	157 (27%)
Fraction of Inspired Oxygen at 44 Weeks	0.29 (0.25, 0.36)
Peak Inspiratory Pressure (cm H2O) at 44 Weeks	0 (0, 11)
Positive End Exploratory Pressure (cm H2O) at 44 Weeks	5.0 (0.0, 8.0)
Medication for Pulmonary Hypertension at 44 Weeks	
0	473 (83%)
1	99 (17%)
Hospital Discharge Gestational Age	46 (42, 54)
Tracheostomy	
0	850 (85%)
1	146 (15%)
Death	54 (5.4%)

[1] n (%); Median (IQR)

Table 2. Descriptive Statistics by Center

Characteristic	**1**, N = 65[1]	**2**, N = 630[1]	**3**, N = 57[1]	**4**, N = 60[1]	**5**, N = 40[1]	**7**, N = 32[1]	**12**, N = 69[1]	**16**, N = 38[1]	**20**, N = 4[1]	**21**, N = 1[1]
Maternal Ethnicity										
1	6 (15%)	24 (3.8%)	14 (25%)	5 (8.6%)	8 (20%)	1 (20%)	8 (12%)	6 (16%)	1 (33%)	1 (100%)
2	34 (85%)	606 (96%)	41 (75%)	53 (91%)	32 (80%)	4 (80%)	61 (88%)	32 (84%)	2 (67%)	0 (0%)
Missing	25	0	2	2	0	27	0	0	1	0
Birth Weight (g)	652 (549, 785)	770 (611, 967)	720 (580, 854)	785 (635, 968)	593 (515, 666)	695 (540, 863)	730 (590, 920)	788 (650, 1,076)	1,115 (879, 1,325)	590 (590, 590)

[1] n (%); Median (IQR)

Table 2. Descriptive Statistics by Center

Characteristic	1, N = 65[1]	2, N = 630[1]	3, N = 57[1]	4, N = 60[1]	5, N = 40[1]	7, N = 32[1]	12, N = 69[1]	16, N = 38[1]	20, N = 4[1]	21, N = 1[1]
Obstetrical Gestational Age	25.00 (24.00, 27.00)	26.00 (24.00, 27.00)	26.00 (24.00, 27.00)	25.00 (24.00, 27.00)	24.00 (23.00, 25.00)	25.00 (23.00, 27.00)	26.00 (25.00, 27.00)	26.00 (24.25, 28.00)	25.50 (24.50, 26.75)	24.00 (24.00, 24.00)
Birth Length (cm)	31.0 (29.0, 33.0)	32.0 (30.0, 35.0)	32.0 (31.0, 34.0)	33.0 (30.5, 35.0)	29.0 (28.0, 30.5)	32.0 (29.0, 35.8)	33.0 (30.8, 34.0)	33.0 (31.0, 36.8)	34.0 (28.5, 38.0)	29.0 (29.0, 29.0)
Missing	9	24	0	1	1	6	37	0	0	0
Birth Head Circumference (cm)	22.25 (21.00, 23.50)	23.00 (21.50, 25.00)	23.20 (21.75, 25.00)	23.50 (22.00, 25.00)	21.00 (20.00, 22.00)	22.00 (20.53, 24.00)	23.00 (21.25, 24.38)	23.50 (21.81, 25.50)	24.90 (22.94, 25.98)	21.00 (21.00, 21.00)
Missing	9	29	0	2	0	6	31	0	0	0
Delivery Method										
1	16 (25%)	177 (28%)	17 (30%)	18 (30%)	14 (35%)	10 (31%)	18 (27%)	14 (37%)	1 (25%)	0 (0%)
2	48 (75%)	453 (72%)	40 (70%)	42 (70%)	26 (65%)	22 (69%)	49 (73%)	24 (63%)	3 (75%)	1 (100%)
Missing	1	0	0	0	0	0	2	0	0	0
Prenatal Corticosteroids	56 (92%)	544 (86%)	47 (87%)	47 (80%)	37 (93%)	26 (87%)	41 (89%)	33 (89%)	3 (75%)	1 (100%)
Missing	4	1	3	1	0	2	23	1	0	0
Complete Prenatal Steroids	34 (69%)	415 (79%)	41 (87%)	25 (58%)	27 (73%)	15 (65%)	26 (60%)	26 (79%)	1 (50%)	0 (0%)
Missing	16	105	10	17	3	9	26	5	2	0
Maternal Chorioamnionitis	17 (49%)	105 (17%)	4 (12%)	8 (14%)	14 (36%)	3 (9.7%)	6 (8.7%)	2 (6.1%)	1 (33%)	0 (0%)
Missing	30	0	23	1	1	1	0	5	1	0
Gender										
Female	26 (41%)	249 (40%)	22 (39%)	28 (47%)	17 (43%)	16 (50%)	28 (41%)	20 (53%)	1 (25%)	1 (100%)
Male	38 (59%)	379 (60%)	34 (61%)	32 (53%)	23 (58%)	16 (50%)	41 (59%)	18 (47%)	3 (75%)	0 (0%)
Missing	1	2	1	0	0	0	0	0	0	0
Small for Gestational Age	26 (41%)	117 (19%)	13 (24%)	5 (8.5%)	8 (20%)	8 (25%)	17 (25%)	7 (18%)	2 (50%)	0 (0%)
Missing	1	10	3	1	0	0	0	0	0	0
Received Surfactant in the First 72 Hours	36 (90%)	265 (79%)	54 (96%)	11 (69%)	33 (94%)	3 (50%)	48 (84%)	9 (56%)	2 (100%)	0 (NA%)
Missing	25	295	1	44	5	26	12	22	2	1
Weight at 36 Weeks	2,100 (1,843, 2,354)	2,150 (1,866, 2,405)	2,115 (1,850, 2,430)	2,100 (1,861, 2,408)	1,943 (1,723, 2,138)	2,200 (1,925, 2,420)	2,009 (1,793, 2,180)	2,273 (2,069, 2,472)	2,485 (2,349, 2,604)	1,339 (1,339, 1,339)
Missing	17	36	3	6	0	1	29	0	0	0
Ventilation Support Level at 36 Weeks										
0	8 (13%)	50 (8.1%)	5 (8.9%)	8 (13%)	0 (0%)	22 (69%)	1 (2.0%)	22 (58%)	1 (25%)	0 (0%)
1	22 (34%)	425 (68%)	35 (63%)	34 (57%)	31 (78%)	8 (25%)	17 (34%)	14 (37%)	2 (50%)	1 (100%)
2	34 (53%)	146 (24%)	16 (29%)	18 (30%)	9 (23%)	2 (6.3%)	32 (64%)	2 (5.3%)	1 (25%)	0 (0%)
Missing	1	9	1	0	0	0	19	0	0	0
Fraction of Inspired Oxygen at 36 Weeks	0.35 (0.30, 0.49)	0.27 (0.23, 0.35)	0.30 (0.25, 0.37)	0.40 (0.30, 0.50)	0.33 (0.26, 0.43)	0.35 (0.32, 0.38)	0.35 (0.27, 0.45)	0.35 (0.27, 0.39)	0.41 (0.31, 0.50)	NA (NA, NA)
Missing	18	36	2	3	0	1	29	0	2	1
Peak Inspiratory Pressure (cm H2O) at 36 Weeks	2 (0, 14)	0 (0, 0)	0 (0, 15)	4 (0, 9)	0 (0, 9)	0 (0, 0)	12 (0, 15)	0 (0, 0)	8 (4, 13)	16 (16, 16)
Missing	20	39	7	15	13	1	32	0	1	0
Positive End Exploratory Pressure (cm H2O) at 36 Weeks	8.0 (6.0, 9.0)	7.0 (6.0, 8.0)	8.0 (7.0, 9.8)	6.0 (6.0, 7.0)	9.0 (8.0, 10.0)	0.0 (0.0, 4.5)	6.0 (6.0, 7.0)	0.0 (0.0, 7.8)	5.5 (3.8, 6.3)	10.0 (10.0, 10.0)
Missing	24	41	11	6	0	1	34	0	0	0
Medication for Pulmonary Hypertension at 36 Weeks										

[1] n (%); Median (IQR)

Table 2. Descriptive Statistics by Center

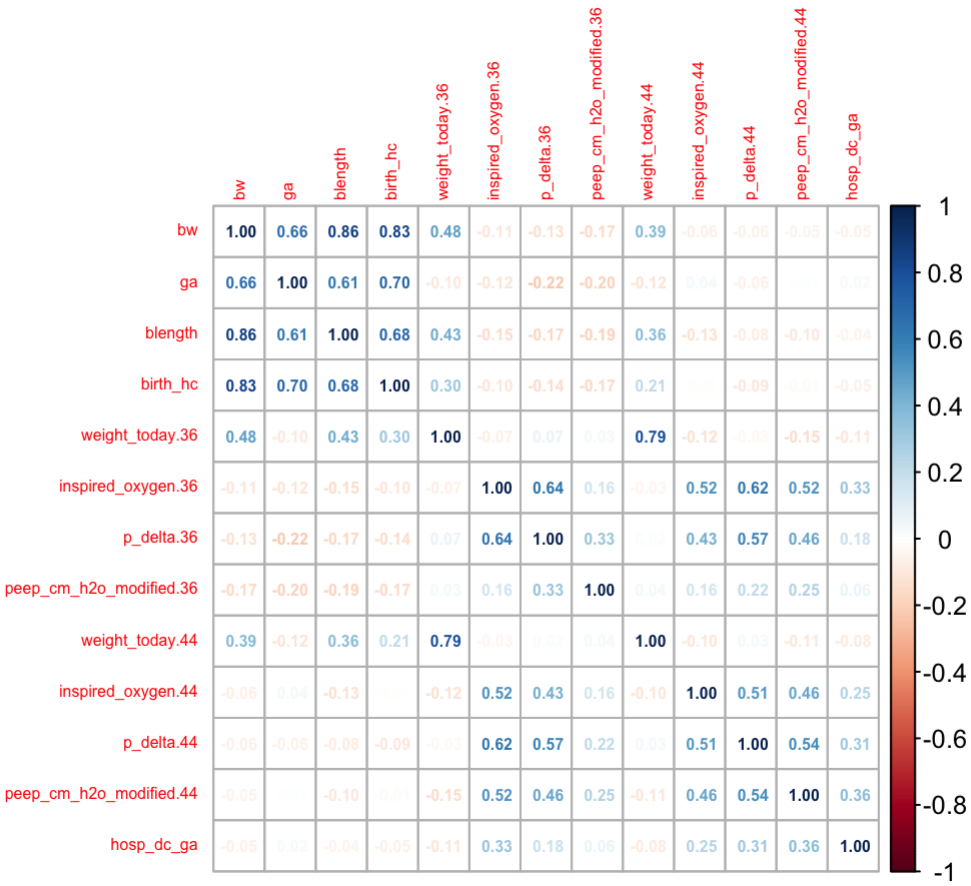
Characteristic	1, N = 65[1]	2, N = 630[1]	3, N = 57[1]	4, N = 60[1]	5, N = 40[1]	7, N = 32[1]	12, N = 69[1]	16, N = 38[1]	20, N = 4[1]	21, N = 1[1]
0	51 (80%)	596 (96%)	53 (95%)	49 (82%)	37 (93%)	30 (94%)	46 (92%)	34 (89%)	4 (100%)	0 (0%)
1	13 (20%)	25 (4.0%)	3 (5.4%)	11 (18%)	3 (7.5%)	2 (6.3%)	4 (8.0%)	4 (11%)	0 (0%)	1 (100%)
Missing	1	9	1	0	0	0	19	0	0	0
Weight at 44 Weeks	3,680 (3,338, 4,200)	3,768 (3,376, 4,140)	3,800 (3,323, 4,120)	NA (NA, NA)	3,372 (3,155, 3,998)	3,860 (3,375, 4,465)	3,270 (2,903, 3,815)	2,950 (2,468, 4,110)	3,725 (3,341, 4,060)	2,610 (2,610, 2,610)
Missing	7	252	38	60	9	21	26	33	0	0
Ventilation Support Level at 44 Weeks										
0	11 (18%)	198 (51%)	12 (60%)	0 (NA%)	19 (61%)	10 (83%)	12 (26%)	5 (100%)	2 (50%)	0 (0%)
1	18 (30%)	97 (25%)	7 (35%)	0 (NA%)	9 (29%)	0 (0%)	13 (28%)	0 (0%)	1 (25%)	1 (100%)
2	32 (52%)	96 (25%)	1 (5.0%)	0 (NA%)	3 (9.7%)	2 (17%)	22 (47%)	0 (0%)	1 (25%)	0 (0%)
Missing	4	239	37	60	9	20	22	33	0	0
Fraction of Inspired Oxygen at 44 Weeks	0.31 (0.25, 0.42)	0.28 (0.26, 0.35)	0.25 (0.23, 0.32)	NA (NA, NA)	0.27 (0.24, 0.33)	0.31 (0.25, 0.47)	0.31 (0.25, 0.51)	0.27 (0.24, 0.29)	0.36 (0.29, 0.42)	0.21 (0.21, 0.21)
Missing	9	253	38	60	9	21	25	33	0	0
Peak Inspiratory Pressure (cm H2O) at 44 Weeks	10 (0, 17)	0 (0, 2)	0 (0, 0)	NA (NA, NA)	0 (0, 0)	0 (0, 0)	0 (0, 17)	0 (0, 0)	6 (0, 13)	0 (0, 0)
Missing	7	248	38	60	14	22	26	33	0	0
Positive End Exploratory Pressure (cm H2O) at 44 Weeks	9.0 (6.0, 12.0)	0.0 (0.0, 8.0)	0.0 (0.0, 6.0)	NA (NA, NA)	0.0 (0.0, 8.0)	0.0 (0.0, 0.0)	6.0 (0.0, 8.0)	0.0 (0.0, 0.0)	3.5 (0.0, 7.0)	5.0 (5.0, 5.0)
Missing	7	250	39	60	9	20	28	33	0	0
Medication for Pulmonary Hypertension at 44 Weeks										
0	32 (52%)	350 (90%)	19 (95%)	0 (NA%)	26 (84%)	8 (67%)	30 (64%)	4 (80%)	4 (100%)	0 (0%)
1	29 (48%)	41 (10%)	1 (5.0%)	0 (NA%)	5 (16%)	4 (33%)	17 (36%)	1 (20%)	0 (0%)	1 (100%)
Missing	4	239	37	60	9	20	22	33	0	0
Hospital Discharge Gestational Age	60 (60, 60)	47 (42, 55)	44 (41, 45)	NA (NA, NA)	52 (48, 54)	43 (39, 47)	51 (47, 59)	40 (39, 43)	52 (44, 63)	66 (66, 66)
Missing	64	0	0	60	0	0	0	0	0	0
Tracheostomy										
0	38 (58%)	566 (90%)	56 (98%)	49 (82%)	35 (88%)	31 (97%)	34 (49%)	37 (97%)	4 (100%)	0 (0%)
1	27 (42%)	64 (10%)	1 (1.8%)	11 (18%)	5 (13%)	1 (3.1%)	35 (51%)	1 (2.6%)	0 (0%)	1 (100%)
Death	7 (11%)	29 (4.6%)	1 (1.8%)	1 (1.7%)	2 (5.0%)	0 (0%)	14 (20%)	0 (0%)	0 (0%)	0 (0%)
Missing	0	1	0	1	0	0	0	0	0	0

[1] n (%); Median (IQR)

Correlations

Correlations for all the continuous variables were also computed (Figure 2). Birth weight and birth length, birth weight and birth head circumference, gestational age and birth head circumference, weight at 36 weeks and weight at 44 weeks, seem to be most strongly positively correlated (correlation coefficient >= 0.7).

Figure 2. Correlations Between Continuous Variables



Methods

Multiple Imputation for Missing Data

The data was split into test (30%) and train data (70%). Upon removing the variables corresponding to death (tracheostomy is the outcome), patient ID (not relevant for imputation), whether the infant received surfactant at any point in the first 72 hours (high missingness and hard to impute with available information), and medical center (not relevant for imputation), multiple imputation was implemented to impute missing data, resulting in 10 imputed data sets (5 imputed data sets for the train data and 5 imputed data sets for the test data).

Model Building - Lasso model

In the development of predictive models for tracheostomy, two lasso models were created, encompassing all 21 predictors except the center variable. The exclusion of center aimed to enhance model generalizability, a critical consideration given the limited data from specific centers within the collaborative NICU network. These models were built using data gathered at 36 weeks PMA and 44 weeks PMA. No interactions were considered to aid in ease of interpretation.

The decision to employ lasso models was rooted in their distinctive attributes. The lasso's capacity to shrink coefficients to zero provides effective variable selection, streamlining the identification of influential predictors. Moreover, its ability to adeptly handle multicollinearity contributes to the development of models that are not only accurate but also simpler and more interpretable.

This emphasis on variable selection and multicollinearity management aligns with the overarching goal of not just predicting tracheostomy accurately but also ensuring the models are intelligible and clinically meaningful. The lasso model's tendency to shrink certain coefficients to zero inherently fosters model simplicity, facilitating the interpretation of key predictors and their impacts on the outcome. These characteristics collectively enhance the robustness, clarity, and practical relevance of the predictive models in the context of neonatal care.

Cross-validation was employed to determine the optimal value for lambda, the parameter regulating the extent of shrinkage in lasso regression. The goal was to identify the lambda value that minimized cross-validation error. Lasso models were individually fitted to each of the five imputed datasets for both the 36 and 44-week time points. The resulting coefficients from these models were then averaged to derive the final models.

The 36 week model is given by:

$$\log(\text{Odds of Tracheostomy}) = -7.5901 - 0.0657X_1 - 0.0003X_2 - 0.0216X_3 + 0.1287X_4 - 0.0095X_5 + 0.1840X_6 + 0.9040X_7 - 0.1057X_8 - 0.2747X_9 - 0.1891X_{10} - 0.0009X_{11} - 0.7525X_{12} + 1.3279X_{13} + 2.7442X_{14} - 0.0358X_{15} + 0.0852X_{16} - 0.0179X_{17} + 0.0376X_{18}$$

where X_1 = mother is not Hispanic/Latino, X_2 = birth weight in grams, X_3 = obstetrical gestational age, X_4 = birth length in cm, X_5 = birth head circumference, X_6 = Cesarean section delivery, X_7 = prenatal steroids were administered, X_8 = complete prenatal steroids were administered, X_9 = maternal chorioamnionitis was present, X_{10} = infant was small for gestational age, X_{11} = infant weight at 36 weeks, X_{12} = ventilation support through non-invasive positive pressure at 36 weeks, X_{13} = ventilation support through invasive positive pressure at 36 weeks, X_{14} = fraction of inspired oxygen at 36 weeks, X_{15} = peak inspiratory pressure (cm H_2O) at 36 weeks, X_{16} = positive end exploratory pressure (cm H_2O) at 36 weeks, X_{17} = medication for pulmonary hypertension was administered at 36 weeks, X_{18} = hospital discharge gestational age

Interpretation of the 36 week model coefficients

- The odds of getting a tracheostomy for infants born to a non-Hispanic/non-Latino mother is 6.36% lower than that for infants born to a Hispanic/Latino mother, provided all other variables are held constant.
- The odds of getting a tracheostomy decreases by 0.31% for every 1 gram increase in birth weight provided all other variables are held constant.
- The odds of getting a tracheostomy decreases by 2.14% for every 1 week increase in obstetrical gestational age provided all other variables are held constant.
- The odds of getting a tracheostomy increases by 13.73% for every 1 cm increase in birth length, provided all other variables are held constant.
- The odds of getting a tracheostomy decreases by 0.95% for every 1 cm increase in birth head circumference, provided all other variables are held constant.
- The odds of getting a tracheostomy for infants born through Cesarean section is 20% higher than that for infants born through vaginal delivery, provided all other variables are held constant.
- The odds of getting a tracheostomy for infants whose mothers received prenatal steroids was 146.94% higher than that for infants whose mothers did not receive them, provided all other variables are held constant.
- The odds of getting a tracheostomy for infants whose mothers received complete prenatal steroids was 10.03% lower than that for infants whose mothers did not receive them, provided all other variables are held constant.
- The odds of getting a tracheostomy for infants whose mothers had chorioamnionitis was 24.02% lower than that for infants whose mothers did not have it, provided all other variables are held constant.
- The odds of getting a tracheostomy for infants who were small for gestational age was 17.23% lower than that for infants who were not small for gestational age, provided all other variables are held constant.
- The odds of getting a tracheostomy decreases by 0.09% for every 1 gram increase in weight measured at 36 weeks, provided all other variables are held constant.

- The odds of getting a tracheostomy for infants who received ventilation support through non-invasive positive pressure at 36 weeks was 52.88% lower than that for infants who received no respiratory support or supplemental oxygen, provided all other variables are held constant.
- The odds of getting a tracheostomy for infants who received ventilation support through invasive positive pressure at 36 weeks was 277.32% higher than that for infants who received no respiratory support or supplemental oxygen, provided all other variables are held constant.
- The odds of getting a tracheostomy increases by 1555.28% for every unit increase in fraction of inspired oxygen at 36 weeks, provided all other variables are held constant.
- The odds of getting a tracheostomy decreases by 3.52% for every 1 cm increase in peak inspiratory pressure at 36 weeks, provided all other variables are held constant.
- The odds of getting a tracheostomy increases by 8.90% for every 1 cm increase in positive end exploratory pressure at 36 weeks, provided all other variables are held constant.
- The odds of getting a tracheostomy for infants who received medication for pulmonary hypertension at 36 weeks was 1.78% lower than that for infants who did not, provided all other variables are held constant.
- The odds of getting a tracheostomy increases by 3.83% for every 1 week increase in gestational age at discharge from hospital, provided all other variables are held constant.

The 44 week model is given by:

$$\log(\text{Odds of Tracheostomy}) = -4.2746 - 0.0694X_1 - 0.0004X_2 - 0.1140X_3 + 0.1362X_4 - 0.0007X_5 + 0.3392X_6 + 0.9517X_7 + 0.0199X_8 - 0.1461X_9 + 0.0431X_{10} - 0.3542X_{11} - 0.0008X_{12} - 1.0793X_{13} + 0.4656X_{14} + 0.1639X_{15} + 0.0027X_{16} - 0.2234X_{17} + 0.5307X_{18} + 0.0330X_{19}$$

where X_1 = mother is not Hispanic/Latino, X_2 = birth weight in grams, X_3 = obstetrical gestational age, X_4 = birth length in cm, X_5 = birth head circumference, X_6 = Cesarean section delivery, X_7 = prenatal steroids were administered, X_8 = complete prenatal steroids were administered, X_9 = maternal chorioamnionitis was present, X_{10} = infant was male, X_{11} = infant was small for gestational age, X_{12} = infant weight at 44 weeks, X_{13} = ventilation support through non-invasive positive pressure at 44 weeks, X_{14} = ventilation support through invasive positive pressure at 44 weeks, X_{15} = fraction of inspired oxygen at 44 weeks, X_{16} = peak inspiratory pressure (cm H_2O) at 44 weeks, X_{17} = positive end exploratory pressure (cm H_2O) at 44 weeks, X_{18} = medication for pulmonary hypertension was administered at 44 weeks, X_{19} = hospital discharge gestational age

Interpretation of the 44 week model coefficients

- The odds of getting a tracheostomy for infants born to a non-Hispanic/non-Latino mother is 6.71% lower than that for infants born to a Hispanic/Latino mother, provided all other variables are held constant.
- The odds of getting a tracheostomy decreases by 0.04% for every 1 gram increase in birth weight provided all other variables are held constant.
- The odds of getting a tracheostomy decreases by 10.77% for every 1 week increase in obstetrical gestational age provided all other variables are held constant.
- The odds of getting a tracheostomy increases by 14.59% for every 1 cm increase in birth length, provided all other variables are held constant.
- The odds of getting a tracheostomy increases by 0.07% for every 1 cm increase in birth head circumference, provided all other variables are held constant.
- The odds of getting a tracheostomy for infants born through Cesarean section is 40.37% higher than that for infants born through vaginal delivery, provided all other variables are held constant.
- The odds of getting a tracheostomy for infants whose mothers received prenatal steroids was 159.01% higher than that for infants whose mothers did not receive them, provided all other variables are held constant.

- The odds of getting a tracheostomy for infants whose mothers received complete prenatal steroids was 2.01% higher than that for infants whose mothers did not receive them, provided all other variables are held constant.
- The odds of getting a tracheostomy for infants whose mothers had chorioamnionitis was 13.59% lower than that for infants whose mothers did not have it, provided all other variables are held constant.
- The odds of getting a tracheostomy for male infants was 4.41% higher than that for female infants, provided all other variables are held constant.
- The odds of getting a tracheostomy for infants who were small for gestational age was 29.83% lower than that for infants who were not small for gestational age, provided all other variables are held constant.
- The odds of getting a tracheostomy decreases by 0.08% for every 1 gram increase in weight measured at 44 weeks, provided all other variables are held constant.
- The odds of getting a tracheostomy for infants who received ventilation support through non-invasive positive pressure at 44 weeks was 66.02% lower than that for infants who received no respiratory support or supplemental oxygen, provided all other variables are held constant.
- The odds of getting a tracheostomy for infants who received ventilation support through invasive positive pressure at 44 weeks was 59.29% higher than that for infants who received no respiratory support or supplemental oxygen, provided all other variables are held constant.
- The odds of getting a tracheostomy increases by 17.81% for every unit increase in fraction of inspired oxygen at 44 weeks, provided all other variables are held constant.
- The odds of getting a tracheostomy increases by 0.27% for every 1 cm increase in peak inspiratory pressure at 44 weeks, provided all other variables are held constant.
- The odds of getting a tracheostomy increases by 25.03% for every 1 cm increase in positive end exploratory pressure at 44 weeks, provided all other variables are held constant.
- The odds of getting a tracheostomy for infants who received medication for pulmonary hypertension at 44 weeks was 70% higher than that for infants who did not, provided all other variables are held constant.
- The odds of getting a tracheostomy increases by 3.36% for every 1 week increase in gestational age at discharge from hospital, provided all other variables are held constant.

Model Evaluation

Performance Measures for the Two Models

Table 3. Performance Measures for the Two Models

Measure	36 Week Model	44 Week Model
Accuracy	0.8742	0.8756
Precision	0.5938	0.5547
Recall	0.2000	0.3737
Specificity	0.9788	0.9535
F Score	0.2992	0.4465
AUC	0.8429	0.8490
Brier Score	0.0950	0.0881

As illustrated in Table 3, both models showcase similar accuracies, indicating comparable overall predictive performance. However, the 44 Week Model consistently outperforms the 36 Week Model across various metrics, including precision, recall, specificity, F score, AUC, and Brier Score. This signifies the 44 Week Model's heightened ability to predict tracheostomy, distinguish between classes, and achieve a more balanced trade-off between precision and recall.

The F score, a composite metric incorporating precision and recall, accentuates the 44 Week Model's superior performance in minimizing both false positives and false negatives. Furthermore, the AUC values affirm that both models possess robust discriminatory power, with a slightly higher AUC for the 44 Week Model.

Notably, the Brier Score, a measure of the mean squared difference between predicted probabilities and actual outcomes, provides insights into calibration. In this aspect, the 44 Week Model excels, exhibiting a lower Brier Score compared to the 36 Week Model. This suggests that the 44 Week Model achieves better alignment between predicted probabilities and observed outcomes, enhancing its overall reliability and predictive accuracy.

Limitations

Data

The dataset used is sourced from the BPD Collaborative Registry, which represents a specific set of interdisciplinary BPD programs across the United States and Sweden. The results and conclusions drawn from this dataset may not be entirely generalizable to other populations or healthcare settings. In addition, limited representation from specific centers might result in biased predictions, emphasizing the need for cautious interpretation. The dataset also contains a significant amount of missing data, particularly in variables recorded at 44 weeks. The missingness might be due to variations in data collection practices among different centers and could potentially introduce bias into the analysis.

Methods

Multiple imputation was used to address missing data, but imputation introduces assumptions about the nature of missingness. It assumes that the missing data are missing at random (MAR), and if this assumption is violated, the imputed values may introduce bias into the models. The decision to exclude the 'center' variable for model development, while enhancing generalizability, might overlook essential center-specific nuances influencing tracheostomy decisions. Additionally, the assumption of linearity in the logistic regression models could limit the capture of complex, non-linear relationships within the data, not to mention no interactions were included in the models developed.

Future Directions

Prospective validation studies are essential to affirm the model's utility in real-time clinical decision-making. In addition, comparison of models built using different approaches, as well as including center-specific predictors in the analysis would be interesting avenues to explore.

Conclusions

This analysis aimed at predicting tracheostomy placement in infants with severe bronchopulmonary dysplasia (sBPD) yields valuable insights for neonatal care. The comprehensive analysis of demographic, birth, and respiratory variables, drawn from diverse centers in the BPD Collaborative Registry, provides a nuanced

understanding of the factors influencing tracheostomy decisions.

Our exploratory analysis revealed complex relationships, emphasizing the importance of preprocessing steps in handling data inconsistencies and missing values. Multiple imputation techniques enhanced dataset completeness, contributing to the robustness of subsequent analyses.

The choice of lasso models for predictive modeling at 36 and 44 weeks postmenstrual age was guided by their ability to handle multicollinearity and promote effective variable selection. The exclusion of the center variable in these models aimed to enhance generalizability, considering variations in data availability across centers.

Model evaluation underscored the consistent superiority of the 44 Week Model. Across various metrics, including precision, recall, specificity, F score, AUC, and the Brier Score, this model exhibited heightened predictive accuracy. The F score highlighted its balanced precision and recall, essential for minimizing both false positives and false negatives.

The lower Brier Score for the 44 Week Model indicates superior calibration, aligning predicted probabilities with observed outcomes. This suggests the model's enhanced reliability and accuracy in predicting tracheostomy placement, crucial for informing clinical decisions in neonatal care.

In conclusion, this study contributes a robust predictive model tailored for neonatal tracheostomy decisions. The findings not only advance our understanding of sBPD management but also offer a practical tool for clinicians, potentially improving patient outcomes and counseling practices in this challenging medical landscape. Further research and validation efforts are encouraged to solidify the model's applicability across diverse clinical settings.

Code Appendix

```
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_knit$set(root.dir = "~/Downloads")

#packages
install.packages("Hmisc", repos = "http://cran.us.r-project.org")
install.packages("performance", repos = "http://cran.us.r-project.org")

library(tidyverse)
library(naniar)
library(mice)
library(gtsummary)
library(corrplot)
library(readr)
library(labelled)
library(Hmisc)
library(glmnet)
library(pROC)
library(lme4)
library(knitr)
library(kableExtra)
#laoding the data
df <- read_csv("project2.csv")
#changing the values of the sga variable and removing the mat_race variable
df <- df %>%
  mutate(sga = case_when(sga == "Not SGA" ~ "No",
         sga == "SGA" ~ "Yes")) %>%
  select(-mat_race)

length(unique(df$record_id)) #only 996 unique ids but we have 999 observations
df <- df[!duplicated(df$record_id),] #removing the duplicates

#getting the center number from the record id
df$center <- ifelse(nchar(as.character(df$record_id)) == 7, as.numeric(substr(df$record_
id, 1, 1)), as.numeric(substr(df$record_id, 1, 2)))

#Change variables to factors
df$center <- as.factor(df$center)
df$mat_ethn <- as.factor(df$mat_ethn)
df$del_method <- as.factor(df$del_method)
df$prenat_ster <- as.factor(df$prenat_ster)
df$com_prenat_ster <- as.factor(df$com_prenat_ster)
df$mat_chorio <- as.factor(df$mat_chorio)
df$gender <- as.factor(df$gender)
df$sga <- as.factor(df$sga)
df$any_surf <- as.factor(df$any_surf)
df$ventilation_support_level.36 <- as.factor(df$ventilation_support_level.36)
df$med_ph.36 <- as.factor(df$med_ph.36)
df$ventilation_support_level_modified.44 <- as.factor(df$ventilation_support_level_modif
ied.44)
df$med_ph.44 <- as.factor(df$med_ph.44)
df$Trach <- as.factor(df$Trach)
df$Death <- as.factor(df$Death)
```

```
#create variable labels
var_label(df) <- list(
  record_id = "Patient ID",
  center = "Medical Center",
  mat_ethn = "Maternal Ethnicity",
  bw = "Birth Weight (g)",
  ga = "Obstetrical Gestational Age",
  blength = "Birth Length (cm)",
  birth_hc = "Birth Head Circumference (cm)",
  del_method = "Delivery Method",
  prenat_ster = "Prenatal Corticosteroids",
  com_prenat_ster = "Complete Prenatal Steroids",
  mat_chorio = "Maternal Chorioamnionitis",
  gender = "Gender",
  sga = "Small for Gestational Age",
  any_surf = "Received Surfactant in the First 72 Hours",
  weight_today.36 = "Weight at 36 Weeks",
  ventilation_support_level.36 = "Ventilation Support Level at 36 Weeks",
  inspired_oxygen.36 = "Fraction of Inspired Oxygen at 36 Weeks",
  p_delta.36 = "Peak Inspiratory Pressure (cm H2O) at 36 Weeks",
  peep_cm_h2o_modified.36 = "Positive End Exploratory Pressure (cm H2O) at 36 Weeks",
  med_ph.36 = "Medication for Pulmonary Hypertension at 36 Weeks",
  weight_today.44 = "Weight at 44 Weeks",
  ventilation_support_level_modified.44 = "Ventilation Support Level at 44 Weeks",
  inspired_oxygen.44 = "Fraction of Inspired Oxygen at 44 Weeks",
  p_delta.44 = "Peak Inspiratory Pressure (cm H2O) at 44 Weeks",
  peep_cm_h2o_modified.44 = "Positive End Exploratory Pressure (cm H2O) at 44 Weeks",
  med_ph.44 = "Medication for Pulmonary Hypertension at 44 Weeks",
  hosp_dc_ga = "Hospital Discharge Gestational Age",
  Trach = "Tracheostomy",
  Death = "Death"
)
#missing data summary
df_miss <- df
colnames(df_miss) <- label(df)

gg_miss_var(df_miss, show_pct = TRUE) +
  theme(axis.text.y = element_text(size = 5)) +
  ggtitle("Figure 1. Percent Missing Observations by Variable") +
  theme(plot.title = element_text(size = 10))
theme_gtsummary_compact(set_theme=TRUE, font_size = 10)

#table of descriptive statistics - overall
df_miss <- df_miss[,-1]

df_miss %>%
  tbl_summary(missing = "no") %>%
  as_gt() %>%
  gt::tab_header(title = "Table 1. Overall Descriptive Statistics")

#table of descriptive statistics by center
```

```
df_miss %>%
  tbl_summary(by = "Medical Center", missing_text = "Missing") %>%
  as_gt() %>%
  gt::tab_header(title = "Table 2. Descriptive Statistics by Center")
#Correlation matrix for the continuous variables
numeric_df <- df %>%
  select(c("bw", "ga", "blength", "birth_hc", "weight_today.36",
                                "inspired_oxygen.36", "p_delta.36", "peep_cm_h2o_modifie
d.36",
                                "weight_today.44", "inspired_oxygen.44", "p_delta.44",
"peep_cm_h2o_modified.44", "hosp_dc_ga"))

correlation_matrix <- cor(numeric_df[complete.cases(numeric_df),])
corrplot(correlation_matrix, method = "number", number.cex = 0.5, title = "Figure 2. Cor
relations Between Continuous Variables", tl.cex = 0.5, mar=c(0,0,2,0))
set.seed(155)
#to create test and train data sets
ignore <- sample(c(TRUE, FALSE), 996, replace = TRUE, prob = c(0.3,0.7))

#data using tracheostomy as the outcome (removing death), and removing any_surf since it
has a lot of missing data and cannot be imputed with information from any other variable
s
df <- df %>%
  select(-c(Death, any_surf, record_id))

#train and test data sets
train_df <- df[!ignore, ]
test_df <- df[ignore, ]

#imputing 5 test and train datasets
imp.train <- mice(train_df[,-1], m = 5, print = FALSE, seed = 155)
imp.test <- mice.mids(imp.train, newdata = test_df[,-1])

df_train_imp_36 <- vector("list", length = 5)
df_test_imp_36 <- vector("list", length = 5)
df_train_imp_44 <- vector("list", length = 5)
df_test_imp_44 <- vector("list", length = 5)

for(i in 1:5){
  df_train_imp_36[[i]] <- mice::complete(imp.train, i) %>%
    select(-c(weight_today.44, ventilation_support_level_modified.44, inspired_oxygen.4
4, p_delta.44, peep_cm_h2o_modified.44, med_ph.44))
  df_train_imp_36[[i]]$center <- train_df$center

  df_test_imp_36[[i]] <- mice::complete(imp.test, i) %>%
    select(-c(weight_today.44, ventilation_support_level_modified.44, inspired_oxygen.4
4, p_delta.44, peep_cm_h2o_modified.44, med_ph.44))
  df_test_imp_36[[i]]$center <- test_df$center

  df_train_imp_44[[i]] <- mice::complete(imp.train, i) %>%
    select(-c(weight_today.36, ventilation_support_level.36, inspired_oxygen.36, p_delt
a.36, peep_cm_h2o_modified.36, med_ph.36))
```

```
  df_train_imp_44[[i]]$center <- train_df$center

  df_test_imp_44[[i]] <- mice::complete(imp.test, i) %>%
    select(-c(weight_today.36, ventilation_support_level.36, inspired_oxygen.36, p_delt
a.36, peep_cm_h2o_modified.36, med_ph.36))
  df_test_imp_44[[i]]$center <- test_df$center
}

#storing all the imputed data
df_train_imp_36_long <- mice::complete(imp.train, action = "long") %>%
  select(-c(weight_today.44, ventilation_support_level_modified.44, inspired_oxygen.44,
p_delta.44, peep_cm_h2o_modified.44, med_ph.44))
  df_train_imp_36_long$center <- rep(train_df$center, 5)

df_test_imp_36_long <- mice::complete(imp.test, action = "long") %>%
  select(-c(weight_today.44, ventilation_support_level_modified.44, inspired_oxygen.44,
p_delta.44, peep_cm_h2o_modified.44, med_ph.44))
  df_test_imp_36_long$center <- rep(test_df$center, 5)

df_train_imp_44_long <- mice::complete(imp.train, action = "long") %>%
  select(-c(weight_today.36, ventilation_support_level.36, inspired_oxygen.36, p_delta.3
6, peep_cm_h2o_modified.36, med_ph.36))
  df_train_imp_44_long$center <- rep(train_df$center, 5)

df_test_imp_44_long <- mice::complete(imp.test, action = "long") %>%
  select(-c(weight_today.36, ventilation_support_level.36, inspired_oxygen.36, p_delta.3
6, peep_cm_h2o_modified.36, med_ph.36))
  df_test_imp_44_long$center <- rep(test_df$center, 5)
lasso <- function(df) {
  #' 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
  x.ord <- model.matrix(Trach~., data = df[,-20])[,-1]
  y.ord <- df$Trach

  # 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 <- cv.glmnet(x.ord, y.ord, nfolds = 10, foldid = folds,
                            alpha = 1, family = "binomial")
  lasso_mod <- glmnet(x.ord, y.ord, nfolds = 10, alpha = 1, family = "binomial",
                      lambda = lasso_mod_cv$lambda.min)

  # Get coefficients
  coef <- coef(lasso_mod)
  return(coef)
}
```

```
# Find average lasso coefficients over imputed datasets - 36 week data
lasso_coef1_36 <- lasso(df_train_imp_36[[1]])
lasso_coef2_36 <- lasso(df_train_imp_36[[2]])
lasso_coef3_36 <- lasso(df_train_imp_36[[3]])
lasso_coef4_36 <- lasso(df_train_imp_36[[4]])
lasso_coef5_36 <- lasso(df_train_imp_36[[5]])
lasso_coef_36 <- cbind(lasso_coef1_36, lasso_coef2_36, lasso_coef3_36,
                    lasso_coef4_36, lasso_coef5_36)
avg_coefs_lasso_36 <- apply(lasso_coef_36, 1, mean)

# Find average lasso coefficients over imputed datasets - 44 week data
lasso_coef1_44 <- lasso(df_train_imp_44[[1]])
lasso_coef2_44 <- lasso(df_train_imp_44[[2]])
lasso_coef3_44 <- lasso(df_train_imp_44[[3]])
lasso_coef4_44 <- lasso(df_train_imp_44[[4]])
lasso_coef5_44 <- lasso(df_train_imp_44[[5]])
lasso_coef_44 <- cbind(lasso_coef1_44, lasso_coef2_44, lasso_coef3_44,
                    lasso_coef4_44, lasso_coef5_44)
avg_coefs_lasso_44 <- apply(lasso_coef_44, 1, mean)
# Find predicted probabilities and predicted classes on long imputed test data

#36 week model
x_vars_36 <- model.matrix(Trach~. , df_test_imp_36_long[,-c(2,3,22)])
df_test_imp_36_long$pred_probs <- as.vector(plogis(x_vars_36 %*% avg_coefs_lasso_36)) #p
redicted probabilities
df_test_imp_36_long$pred <- ifelse(df_test_imp_36_long$pred_probs > 0.5, 1, 0) #predicte
d classes

#44 week model
x_vars_44 <- model.matrix(Trach~. , df_test_imp_44_long[,-c(2,3,22)])
df_test_imp_44_long$pred_probs <- as.vector(plogis(x_vars_44 %*% avg_coefs_lasso_44)) #p
redicted probabilities
df_test_imp_44_long$pred <- ifelse(df_test_imp_44_long$pred_probs > 0.5, 1, 0) #predicte
d classes
#36 week model

#confusion matrix
confusion_matrix_36 <- table(Observed = df_test_imp_36_long$Trach, Predicted = df_test_i
mp_36_long$pred)

#accuracy
accuracy_36 <- sum(diag(confusion_matrix_36)) / sum(confusion_matrix_36)

#precision (positive predictive value)
precision_36 <- confusion_matrix_36[2, 2] / sum(confusion_matrix_36[, 2])

#recall (sensitivity / true positive rate)
recall_36 <- confusion_matrix_36[2, 2] / sum(confusion_matrix_36[2, ])

#specificity (true negative rate)
specificity_36 <- confusion_matrix_36[1, 1] / sum(confusion_matrix_36[1, ])
```

```
#f score
f_score_36 <- 2 * (precision_36 * recall_36) / (precision_36 + recall_36)

#roc and auc
roc_obj_36 <- roc(df_test_imp_36_long$Trach, df_test_imp_36_long$pred_probs, auc=TRUE)
auc_roc_36 <- auc(roc_obj_36)

#brier score
brier_score_36 <- mean((df_test_imp_36_long$pred_probs - (as.numeric(df_test_imp_36_long
$Trach)-1))^2)

#44 week model

#confusion matrix
confusion_matrix_44 <- table(Observed = df_test_imp_44_long$Trach, Predicted = df_test_i
mp_44_long$pred)

#accuracy
accuracy_44 <- sum(diag(confusion_matrix_44)) / sum(confusion_matrix_44)

#precision (positive predictive value)
precision_44 <- confusion_matrix_44[2, 2] / sum(confusion_matrix_44[, 2])

#recall (sensitivity / true positive rate)
recall_44 <- confusion_matrix_44[2, 2] / sum(confusion_matrix_44[2, ])

#specificity (true negative rate)
specificity_44 <- confusion_matrix_44[1, 1] / sum(confusion_matrix_44[1, ])

#f score
f_score_44 <- 2 * (precision_44 * recall_44) / (precision_44 + recall_44)

#roc and auc
roc_obj_44 <- roc(df_test_imp_44_long$Trach, df_test_imp_44_long$pred_probs)
auc_roc_44 <- auc(roc_obj_44)

#brier score
brier_score_44 <- mean((df_test_imp_44_long$pred_probs - (as.numeric(df_test_imp_44_long
$Trach)-1))^2)

perf_tab <- data.frame(
  Measure = c("Accuracy", "Precision", "Recall", "Specificity", "F Score", "AUC", "Brier
Score"),
  Week36_Model = c(
    sprintf("%.4f", accuracy_36),
    sprintf("%.4f", precision_36),
    sprintf("%.4f", recall_36),
    sprintf("%.4f", specificity_36),
    sprintf("%.4f", f_score_36),
    sprintf("%.4f", auc_roc_36),
    sprintf("%.4f", brier_score_36)
```

```
  ),
  Week44_Model = c(
    sprintf("%.4f", accuracy_44),
    sprintf("%.4f", precision_44),
    sprintf("%.4f", recall_44),
    sprintf("%.4f", specificity_44),
    sprintf("%.4f", f_score_44),
    sprintf("%.4f", auc_roc_44),
    sprintf("%.4f", brier_score_44)
  )
)

perf_tab %>%
  kbl(col.names = c("Measure", "36 Week Model", "44 Week Model"), caption = "Table 3. Pe
rformance Measures for the Two Models", booktabs = T) %>%
  kable_styling(latex_options = "hold_position")
```