A Comprehensive Regression Analysis Approach into Investigating Predictive Models for Tracheostomy Outcomes

Victoria Grase

2023-11-12

Abstract

This study endeavors to construct predictive models for the composite outcome of tracheostomy, with a specific emphasis on guiding indication criteria and optimal timing for tracheostomy placement. Leveraging various regression analysis techniques, including Lasso regression, logistic regression, and backward stepwise regression, the research aims to understand the intricate relationships among different variables influencing tracheostomy occurrences. Initial exploration of bivariate and univariate relationships reveals a class imbalance, prompting a meticulous approach to model development. The analysis encompasses a diverse dataset, encompassing patient demographics, clinical indicators, and relevant medical history. Lasso regression, renowned for its feature selection capability, is employed to identify the most influential predictors associated with tracheostomy. Logistic regression further refines the model by calculating the probability of tracheostomy occurrence based on the identified predictors. Addressing challenges posed by missing data, multiple imputation techniques are implemented to effectively handle incomplete information. This involves creating imputed datasets, conducting analyses on each, and comparing results to provide a comprehensive and robust prediction model. The developed models aim results show that logistic regression model seems to be the best predictor for tracheostomy placement timing. From the statistical analysis performed its certain measurements such as sensitivity, AUC, specificity, and observation of coefficients that further supports why logistic regression outperforms lasso.

Introduction

Bronchopulmonary Dysplasia (BPD) encompasses a spectrum of complications that affect premature infants, varying from mild to moderate and severe manifestations. This condition is intricately linked to the premature birth scenario, where infants are delivered before attaining the expected gestational maturity, leading to structural damage in their lungs. Each year, over 10,000 neonates experience the impact of severe Bronchopulmonary Dysplasia (sBPD), necessitating the reliance on ventilators to support lung function. Notably, 75% of these infants are discharged from medical facilities with ventilator support, administered through tracheostomy, ensuring continued assistance for daily living and the preservation of vital functions in the future.

Despite tracheostomy being a prevalent surgical intervention providing support to patients with sBPD, it is not without inherent risks, particularly an elevated vulnerability to death and infection. Consequently, within clinical settings, the crucial need arises to differentiate among sBPD patients, identifying those who genuinely necessitate tracheostomy intervention. The primary objective of this investigation is to construct a robust regression model tailored to predict the composite outcome of tracheostomy and death, thereby offering valuable guidance on the indication criteria and optimal timing for tracheostomy placement. A distinctive aspect of this model development lies in its focus on statistical variables collected at both 36 and 44 weeks of gestation, providing a comprehensive understanding of factors influencing the likelihood of tracheostomy. The variables of particular interest encompass birth-related factors such as weight, gestational age, the administration of prenatal steroids for lung development, maternal race, gender (acknowledging potential gender-based differences in outcomes), and the presence of chorioamnionitis, an infection of the

amniotic fluid. Additionally, variables recorded at 36 and 44 weeks, including weight and the eventual need for tracheostomy, contribute crucial insights into the predictive dynamics.

In addition of this investigation, it is essential to include additional variables that contribute significantly to the predictive accuracy of the model. Factors such as the respiratory support level and the presence of pulmonary hypertension, known risk factors, are integral to capturing the complexity of tracheostomy outcomes. Furthermore, recognizing the diversity in medical practices across different centers, the variable of medical center becomes pivotal. Some medical centers serve as referral centers with specialized practices, often being academic tertiary centers. To look at these these distinctions, an approach is to incorporate the medical center as a categorical covariate in the model, considering it as an interaction term to account for center-specific patterns. Additionally, the inclusion of surfactant as a variable, a substance aiding in maintaining lung inflation and developing around 34 weeks, further refines the model's capacity to predict tracheostomy outcomes. The nature of these variables ensures that the regression model not only addresses the core predictors but also the complexity associated with medical practices, respiratory support, and additional risk factors important for the tracheostomy outcome prediction.

The report gives detailed description and interpretation of two distinct models, lasso and binomial logistic. These models serve as analytical models show relationships between the identified key variables and predictions for diverse subset of infants. This exploration extends to the application of a sensitivity analysis, leveraging various methods such as cross-validation, F-score, and Brier score. The sensitivity analysis aims to examine the robustness and reliability of the models in capturing variations of the data. The comprehensive evaluation gives an understanding of how different variables interplay in influencing the composite outcome of tracheostomy. And our analysis gives a recommendation for the best robust model of that would be best to use in real world examples by looking at completed data through imputation. This method offers meaningful insight into the intricate dynamics surrounding the indication criteria and timing of tracheostomy placement.

Re-formatting and Missing Data of Study First, my aim is to enhance the data integrity by reformatting it, addressing any outliers or missing values that could potentially introduce bias or skew the results. To achieve this, I applied factorization to variables including center, race, ethnicity, delivery method, prenatal steroid administration, chorioamnionitis, gender, and surfactant status. Additionally, ventilation support level and medication for pulmonary hypertension were factored at both 36 and 44 weeks to assess baseline characteristics concerning the binomial outcome of Tracheostomy. During this process, I identified and removed duplicate entries for a specific patient, ensuring the dataset's consistency. Furthermore, an outlier affecting gestational age was identified and subsequently excluded to maintain the statistical robustness of the analyses.

Next, I looked at missing data patterns in variables across the dataset provided. Among the variables examined, Inspired Oxygen at 44 weeks exhibits the highest degree of missingness, with 44.98% of the data absent, followed closely by Peak Inspiratory Pressure (p_delta.44) and infant weight at 44 weeks, both with a missingness rate of 44.98% and 44.78%, respectively. Similarly, Positive End Exploratory Pressure (peep_cm_h2o_modified.44), the binary variable indicating the administration of any surfactant (any_surf), and the modified ventilation support level at 44 weeks display substantial missing data, each at a rate exceeding 40%. Notably, Medication for Pulmonary Hypertension at 44 weeks (med_ph.44), Composite Prenatal Steroids (com_prenat_ster), and Positive End Exploratory Pressure at 36 weeks (p_delta.36) exhibit relatively lower missingness rates ranging from 12.45% to 19.38%.As you can see from the table there seems to be a lot of missing values for the 44 week time period. It's essential to acknowledge a limitation in the form of higher missing data for variables at 44 weeks compared to those at 36 weeks. This discrepancy can be attributed to discharges occurring between the 36 and 44 weeks, introducing challenges in data collection during this interval. This limitation underscores the need for careful consideration and potential adjustment for missing data when interpreting and generalizing the study's findings.

Study Characteristics and Population

The dataset comprises information from 996 participants encompassing a total of 28 variables, offering a comprehensive representation of demographics, gestational age, birthweight, prenatal steroid administration (facilitating lung development), maternal race/ethnicity, gender (acknowledging gender-based differences),

Table 1: Table 1: Variables Missing

Variable name	# Missing	% Missing
inspired_oxygen.44	448	44.9799197
p_delta.44	448	44.9799197
weight_today.44	446	44.7791165
peep_cm_h2o_modified.44	446	44.7791165
any_surf	433	43.4738956
ventilation_support_level_modified.44	424	42.5702811
med_ph.44	424	42.5702811
com_prenat_ster	193	19.3775100
p_delta.36	128	12.8514056
hosp_dc_ga	124	12.4497992
peep_cm_h2o_modified.36	117	11.7469880
weight_today.36	92	9.2369478
inspired_oxygen.36	92	9.2369478
blength	78	7.8313253
birth_hc	77	7.7309237
mat_chorio	62	6.2248996
mat_ethn	57	5.7228916
mat_race	56	5.6224900
prenat_ster	35	3.5140562
ventilation_support_level.36	30	3.0120482
med_ph.36	30	3.0120482
sga	15	1.5060241
center	10	1.0040161
gender	4	0.4016064
del_method	3	0.3012048
Death	2	0.2008032
record_id	0	0.0000000
bw	0	0.0000000
ga	0	0.0000000
Trach	0	0.0000000

Table 2: Summary of Demographics and Infant Delivery

Characteristic	N Missing	N =996
center	10 (1.0%)	NA
1	NA	55 (5.6%)
2	NA	630 (64%)
3	NA	57 (5.8%)
4	NA	60 (6.1%)
5	NA	40 (4.1%)
7	NA	32 (3.2%)
12	NA	69 (7.0%)
16	NA	38 (3.9%)
20	NA	4 (0.4%)
21	NA	1 (0.1%)
mat_race	56 (5.6%)	NA
0	NA	538 (57%)
1	NA	290 (31%)
2	NA	112 (12%)
mat_ethn	57 (5.7%)	NA
1	NA	74 (7.9%)
2	NA	865 (92%)
del_method	3 (0.3%)	NA
1	NA	285 (29%)
2	NA	708 (71%)
gender	4 (0.4%)	NA
Female	NA	408 (41%)
Male	NA	584 (59%)
sga	15 (1.5%)	NA
Not SGA	NA	778 (79%)
SGA	NA	203 (21%)
any_surf	433 (43%)	461 (82%)
Trach	0 (0%)	NA
0	NA	850 (85%)
1	NA	146 (15%)
Death	2 (0.2%)	54 (5.4%)

and the presence of chorioamnionitis (an infection of the amniotic fluid). Additionally, the study captures measurements at both 36 and 44 weeks for critical variables such as Peak Inspiratory Pressure, Positive End Exploratory Pressure, Medication for Pulmonary Hypertension, weight, ventilation support level, and fraction of inspired oxygen. Within this wealth of data, binary potential outcomes were recorded, presenting opportunities for evaluating the goal of determining optimal tracheostomy placement timing, with death and the presence of tracheostomy being key outcomes of interest. To streamline the analysis and enhance interpretability, the focus was deliberately narrowed down to the binomial outcome variable "Trach" to discern the best-fitting model for achieving the study's goals.

Table 3: Summary of Infant Information

Characteristic	N Missing	N =996
prenat_ster	35 (3.5%)	835 (87%)
com_prenat_ster	193 (19%)	610 (76%)
mat_chorio	62 (6.2%)	160 (17%)
weight_today.36	92 (9.2%)	2,130 (1,856, 2,400)
ventilation_support_level.36	30 (3.0%)	NA
0	NA	117 (12%)
1	NA	589 (61%)
2	NA	260 (27%)
inspired_oxygen.36	92 (9.2%)	0.30 (0.24, 0.38)
p_delta.36	128 (13%)	0 (0, 8)
peep_cm_h2o_modified.36	117 (12%)	7 (6, 8)
med_ph.36	30 (3.0%)	NA
0	NA	900 (93%)
1	NA	66 (6.8%)
weight_today.44	446 (45%)	3,700 (3,241, 4,115)
ventilation_support_level_modified.44	424 (43%)	NA
0	NA	269 (47%)
1	NA	146 (26%)
2	NA	157 (27%)
inspired_oxygen.44	448 (45%)	0.29 (0.25, 0.36)
p_delta.44	448 (45%)	0 (0, 11)
peep_cm_h2o_modified.44	446 (45%)	5 (0, 8)
med_ph.44	424 (43%)	NA
0	NA	473 (83%)
1	NA	99 (17%)
hosp_dc_ga	124 (12%)	46 (42, 54)

Table 4: ** Average Birth and Infants Weights and Heights**

Variable	**Overall**, N = 996	**0**, N = 850	**1**, N = 146
___gender___			
Female	408 (41%)	348 (41%)	60 (41%)
Male	584 (59%)	498 (59%)	86 (59%)
Missing	4	4	0
___Average Birth Weight {g}___			
Mean	806	814	761
___Average Height of Infant {cm}___			
Mean	32	33	32
Missing	78	48	30
___Average Circumference of Infant Head {cm}___			
Mean	23.19	23.22	22.99
Missing	77	46	31
___Average Weight at 36 Weeks {g}___			
Mean	2,121	2,132	2,024
Missing	92	38	54
___Average Weight at 44 Weeks {g}___			
Mean	3,646	3,667	3,550
Missing	446	399	47

Table 5: **Average of Oxygen, Pressure, and Gestiational Age**

Variable	**Overall**, N = 996	**0**, N = 850	**1**, N = 1
___Fraction of Inspired Oxygen at 36 weeks___			
Mean	0.34	0.32	0.49
Missing	92	37	55
___Fraction of Inspired Oxygen at 44 weeks___			
Mean	0.34	0.32	0.44
Missing	448	398	50
___Peak Inspiratory Pressure (cmH2O) at 36 weeks___			
Mean	5	4	15
Missing	128	63	65
___Peak Inspiratory Pressure (cmH2O) at 44 weeks___			
Mean	8	5	21
Missing	448	397	51
___Hospital Discharge Gestational Age___			
Mean	53	49	80
Missing	124	86	38
___Positive and exploratory pressure (cm H2O) at 36 weeks___			
Mean	6	6	8
Missing	117	59	58
___Positive and exploratory pressure (cm H2O) at 44 weeks___			
Mean	4	3	9
Missing	446	396	50

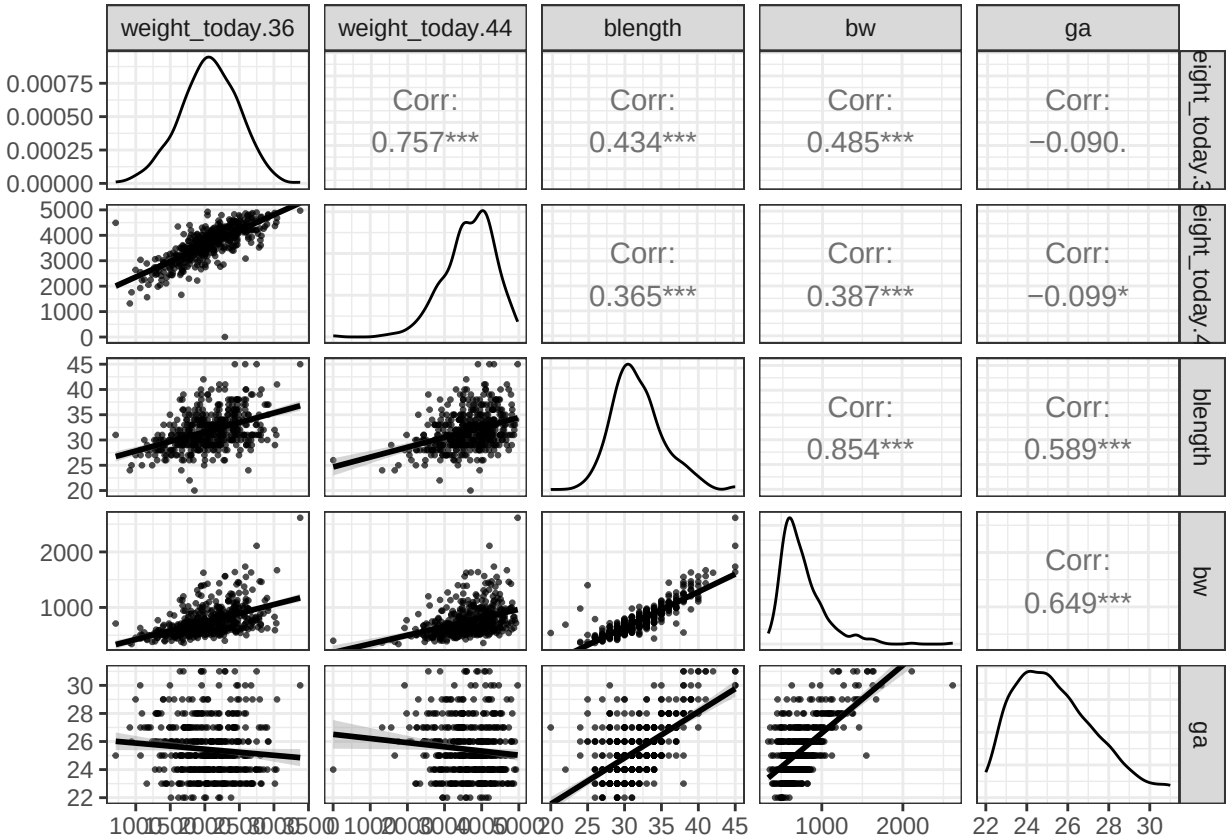
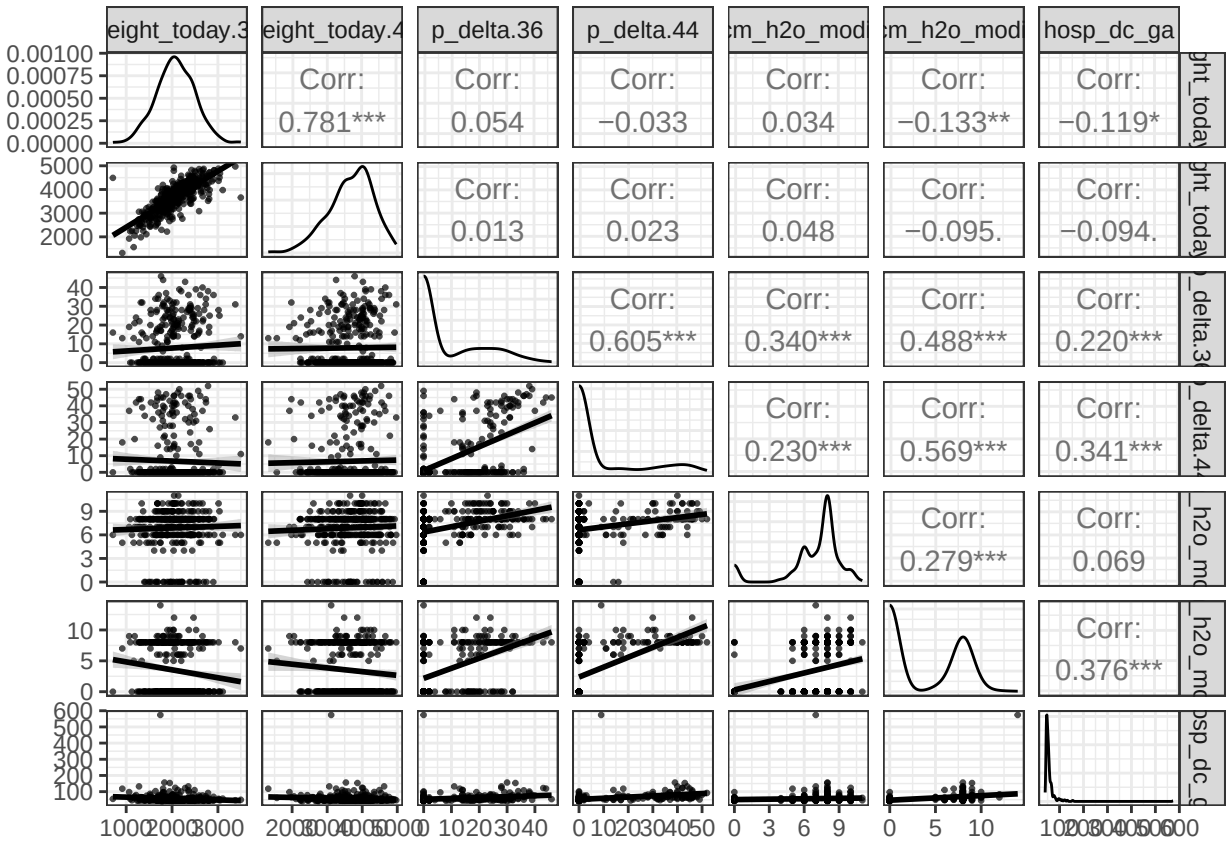


Table 6: ** Summary of Steroid Status, Chorioamnionitis,Severity, Ventilation**

	Overall, N = 996	**0**, N = 850	**1**, N = 146
___Average #___			
Female	408 (41%)	348 (41%)	60 (41%)
Male	584 (59%)	498 (59%)	86 (59%)
Missing	4	4	0
___Prenatal Steroids___	835 (87%)	712 (86%)	123 (94%)
Missing	35	20	15
___Delivery Method___			
1	285 (29%)	254 (30%)	31 (21%)
2	708 (71%)	594 (70%)	114 (79%)
Missing	3	2	1
___Ventilation support level at 36 weeks___			
0	117 (12%)	111 (13%)	6 (4.7%)
1	589 (61%)	560 (67%)	29 (23%)
2	260 (27%)	168 (20%)	92 (72%)
Missing	30	11	19
___Ventilation support level at 44 weeks___			
0	269 (47%)	262 (57%)	7 (6.3%)
1	146 (26%)	128 (28%)	18 (16%)
2	157 (27%)	71 (15%)	86 (77%)
Missing	424	389	35
___Completed Prenatal Steroids___	610 (76%)	524 (76%)	86 (76%)
Missing	193	160	33
___Maternal Chorioamnionitis___	160 (17%)	138 (17%)	22 (16%)
Missing	62	50	12
___Medication for Pulmonary Hypertension at 36 weeks___			
0	900 (93%)	798 (95%)	102 (80%)
1	66 (6.8%)	41 (4.9%)	25 (20%)
Missing	30	11	19
___Medication for Pulmonary Hypertension at 44 weeks___			
0	473 (83%)	413 (90%)	60 (54%)
1	99 (17%)	48 (10%)	51 (46%)
Missing	424	389	35
___Severity of BPD___			
1	263 (29%)	256 (31%)	7 (7.5%)
2	232 (25%)	226 (28%)	6 (6.5%)
3	415 (46%)	335 (41%)	80 (86%)
Missing	86	33	53


eight_today.3
eight_today.4
p_delta.36
p_delta.44
m_h2o_mod
m_h2o_mod
hosp_dc_ga
Corr:
0.781***
Corr:
0.054
Corr:
−0.033
Corr:
0.034
Corr:
−0.133**
Corr:
−0.119*
Corr:
0.013
Corr:
0.023
Corr:
0.048
Corr:
−0.095.
Corr:
−0.094.
Corr:
0.605***
Corr:
0.340***
Corr:
0.488***
Corr:
0.220***
Corr:
0.230***
Corr:
0.569***
Corr:
0.341***
Corr:
0.279***
Corr:
0.069
Corr:
0.376***

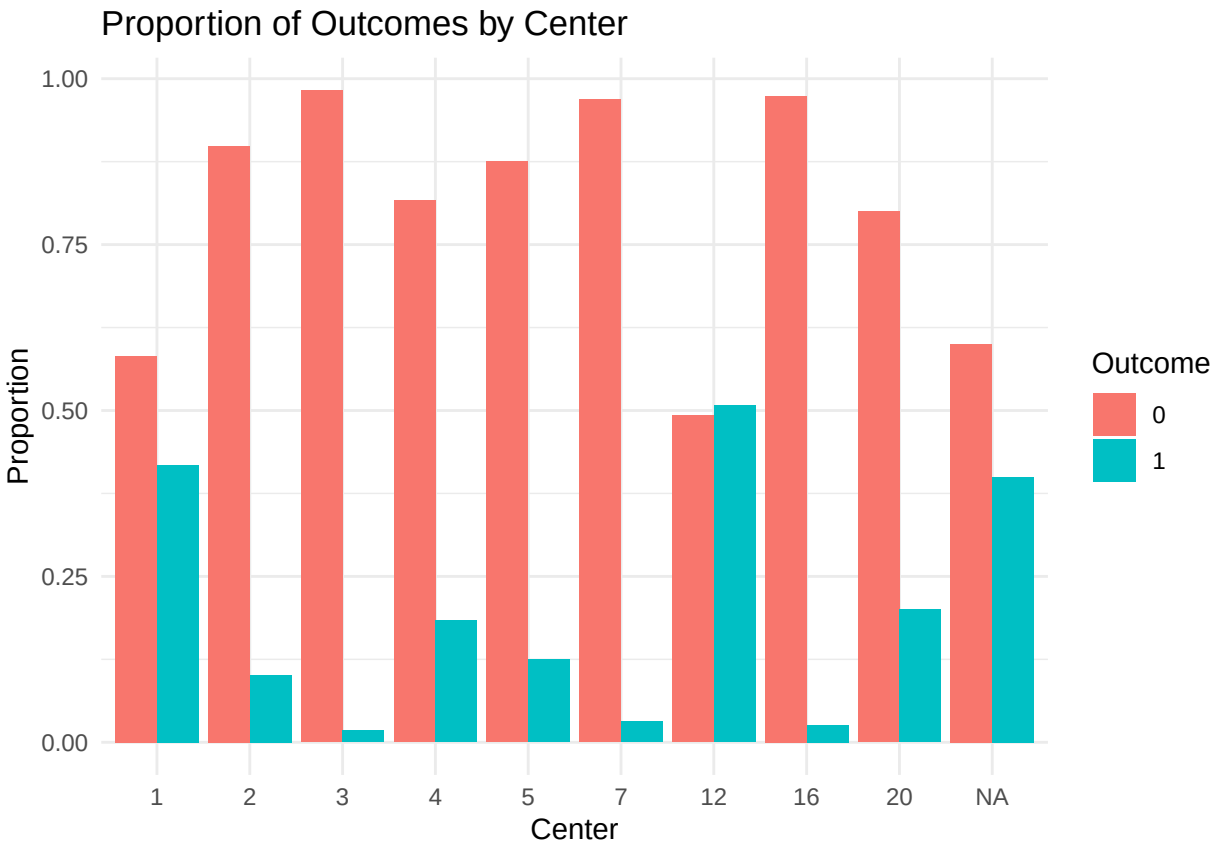

Proportion of Outcomes by Center
1.00
0.75
0.50
0.25
0.00
Proportion
1
2
3
4
5
7
12
16
20
NA
Center
Outcome
0
1

Table 7: Summary Statistics by Center

Characteristic	**1**, N = 55	**2**, N = 630	**3**, N = 57	**4**, N = 60	**5**, N = 40	**7**, N = 32	**12**, N = 69	**16**, N = 38	**20**, N = 5	**21**, N = 0	**p-value**
mat_race											
0	20 (53%)	391 (62%)	31 (60%)	40 (69%)	12 (30%)	1 (20%)	16 (23%)	25 (66%)	0 (0%)	0 (NA%)	
1	18 (47%)	192 (30%)	10 (19%)	16 (28%)	26 (65%)	3 (60%)	16 (23%)	5 (13%)	1 (25%)	0 (NA%)	
2	0 (0%)	47 (7.5%)	11 (21%)	2 (3.4%)	2 (5.0%)	1 (20%)	37 (54%)	8 (21%)	3 (75%)	0 (NA%)	
bw	652 (538, 790)	770 (611, 967)	720 (580, 854)	785 (635, 968)	593 (515, 666)	695 (540, 863)	730 (590, 920)	788 (650, 1,076)	980 (590, 1,250)	NA (NA, NA)	<0.001
ga	25 (24, 27)	26 (24, 27)	26 (24, 27)	25 (24, 27)	24 (23, 25)	25 (23, 27)	26 (25, 27)	26 (24, 28)	25 (24, 26)	NA (NA, NA)	<0.001
blength	31 (29, 33)	32 (30, 35)	32 (31, 34)	33 (31, 35)	29 (28, 31)	32 (29, 36)	33 (31, 34)	33 (31, 37)	30 (29, 38)	NA (NA, NA)	<0.001
birth_hc	22.00 (21.00, 23.00)	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)	24.00 (21.00, 25.80)	NA (NA, NA)	<0.001
del_method											
1	15 (27%)	177 (28%)	17 (30%)	18 (30%)	14 (35%)	10 (31%)	18 (27%)	14 (37%)	1 (20%)	0 (NA%)	
2	40 (73%)	453 (72%)	40 (70%)	42 (70%)	26 (65%)	22 (69%)	49 (73%)	24 (63%)	4 (80%)	0 (NA%)	
prenat_ster	46 (90%)	544 (86%)	47 (87%)	47 (80%)	37 (93%)	26 (87%)	41 (89%)	33 (89%)	4 (80%)	0 (NA%)	
com_prenat_ster	29 (73%)	415 (79%)	41 (87%)	25 (58%)	27 (73%)	15 (65%)	26 (60%)	26 (79%)	1 (33%)	0 (NA%)	
mat_chorio	14 (48%)	105 (17%)	4 (12%)	8 (14%)	14 (36%)	3 (9.7%)	6 (8.7%)	2 (6.1%)	1 (25%)	0 (NA%)	
gender											
Female	21 (38%)	249 (40%)	22 (39%)	28 (47%)	17 (43%)	16 (50%)	28 (41%)	20 (53%)	2 (40%)	0 (NA%)	
Male	34 (62%)	379 (60%)	34 (61%)	32 (53%)	23 (58%)	16 (50%)	41 (59%)	18 (47%)	3 (60%)	0 (NA%)	
sga											
Not SGA	33 (61%)	503 (81%)	41 (76%)	54 (92%)	32 (80%)	24 (75%)	52 (75%)	31 (82%)	3 (60%)	0 (NA%)	
SGA	21 (39%)	117 (19%)	13 (24%)	5 (8.5%)	8 (20%)	8 (25%)	17 (25%)	7 (18%)	2 (40%)	0 (NA%)	
any_surf	31 (91%)	265 (79%)	54 (96%)	11 (69%)	33 (94%)	3 (50%)	48 (84%)	9 (56%)	2 (100%)	0 (NA%)	
weight_today.36	2,115 (1,835, 2,358)	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)	2,375 (2,270, 2,595)	NA (NA, NA)	0.013
ventilation_support_level.36											
0	7 (13%)	50 (8.1%)	5 (8.9%)	8 (13%)	0 (0%)	22 (69%)	1 (2.0%)	22 (58%)	1 (20%)	0 (NA%)	
1	19 (35%)	425 (68%)	35 (63%)	34 (57%)	31 (78%)	8 (25%)	17 (34%)	14 (37%)	3 (60%)	0 (NA%)	
2	29 (53%)	146 (24%)	16 (29%)	18 (30%)	9 (23%)	2 (6.3%)	32 (64%)	2 (5.3%)	1 (20%)	0 (NA%)	
inspired_oxygen.36	0.35 (0.28, 0.50)	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)	<0.001
p_delta.36	3 (0, 14)	0 (0, 0)	0 (0, 15)	4 (0, 9)	0 (0, 9)	0 (0, 0)	12 (0, 15)	0 (0, 0)	12 (6, 16)	NA (NA, NA)	<0.001
peep_cm_h2o_modified.36	8 (7, 9)	7 (6, 8)	8 (7, 10)	6 (6, 7)	9 (8, 10)	0 (0, 5)	6 (6, 7)	0 (0, 8)	6 (5, 7)	NA (NA, NA)	<0.001
med_ph.36											
0	42 (76%)	596 (96%)	53 (95%)	49 (82%)	37 (93%)	30 (94%)	46 (92%)	34 (89%)	4 (80%)	0 (NA%)	
1	13 (24%)	25 (4.0%)	3 (5.4%)	11 (18%)	3 (7.5%)	2 (6.3%)	4 (8.0%)	4 (11%)	1 (20%)	0 (NA%)	
weight_today.44	3,700 (3,315, 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)	3,505 (2,850, 3,945)	NA (NA, NA)	0.015
ventilation_support_level_modified.44											
0	9 (18%)	198 (51%)	12 (60%)	0 (NA%)	19 (61%)	10 (83%)	12 (26%)	5 (100%)	2 (40%)	0 (NA%)	
1	14 (27%)	97 (25%)	7 (35%)	0 (NA%)	9 (29%)	0 (0%)	13 (28%)	0 (0%)	2 (40%)	0 (NA%)	
2	28 (55%)	96 (25%)	1 (5.0%)	0 (NA%)	3 (9.7%)	2 (17%)	22 (47%)	0 (0%)	1 (20%)	0 (NA%)	
inspired_oxygen.44	0.32 (0.25, 0.40)	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)	0.30 (0.26, 0.41)	NA (NA, NA)	0.040
p_delta.44	12 (0, 17)	0 (0, 2)	0 (0, 0)	NA (NA, NA)	0 (0, 0)	0 (0, 0)	0 (0, 17)	0 (0, 0)	0 (0, 11)	NA (NA, NA)	<0.001
peep_cm_h2o_modified.44	9 (6, 12)	0 (0, 8)	0 (0, 6)	NA (NA, NA)	0 (0, 8)	0 (0, 0)	6 (0, 8)	0 (0, 0)	5 (0, 7)	NA (NA, NA)	<0.001
med_ph.44											
0	25 (49%)	350 (90%)	19 (95%)	0 (NA%)	26 (84%)	8 (67%)	30 (64%)	4 (80%)	4 (80%)	0 (NA%)	
1	26 (51%)	41 (10%)	1 (5.0%)	0 (NA%)	5 (16%)	4 (33%)	17 (36%)	1 (20%)	1 (20%)	0 (NA%)	
hosp_dc_ga	60 (60, 60)	47 (42, 55)	44 (41, 45)	NA (NA, NA)	52 (48, 54)	43 (39, 47)	51 (47, 59)	40 (39, 43)	60 (44, 66)	NA (NA, NA)	<0.001
Trach											
0	32 (58%)	566 (90%)	56 (98%)	49 (82%)	35 (88%)	31 (97%)	34 (49%)	37 (97%)	4 (80%)	0 (NA%)	
1	23 (42%)	64 (10%)	1 (1.8%)	11 (18%)	5 (13%)	1 (3.1%)	35 (51%)	1 (2.6%)	1 (20%)	0 (NA%)	
Death	7 (13%)	29 (4.6%)	1 (1.8%)	1 (1.7%)	2 (5.0%)	0 (0%)	14 (20%)	0 (0%)	0 (0%)	0 (NA%)	
severity											
1	8 (19%)	177 (30%)	10 (18%)	10 (18%)	4 (10%)	23 (74%)	5 (12%)	23 (61%)	2 (67%)	0 (NA%)	
2	6 (14%)	177 (30%)	15 (27%)	9 (16%)	8 (20%)	1 (3.2%)	6 (15%)	10 (26%)	0 (0%)	0 (NA%)	
3	29 (67%)	241 (41%)	30 (55%)	38 (67%)	28 (70%)	7 (23%)	30 (73%)	5 (13%)	1 (33%)	0 (NA%)	
mother_ethn											
Hispanic or Latino	4 (12%)	24 (3.8%)	14 (25%)	5 (8.6%)	8 (20%)	1 (20%)	8 (12%)	6 (16%)	2 (50%)	0 (NA%)	
Not Hispanic or Latino	30 (88%)	606 (96%)	41 (75%)	53 (91%)	32 (80%)	4 (80%)	61 (88%)	32 (84%)	2 (50%)	0 (NA%)	
mother_race											
American Indian or Alaskan Native	20 (53%)	391 (62%)	31 (60%)	40 (69%)	12 (30%)	1 (20%)	16 (23%)	25 (66%)	0 (0%)	0 (NA%)	
Asian	18 (47%)	192 (30%)	10 (19%)	16 (28%)	26 (65%)	3 (60%)	16 (23%)	5 (13%)	1 (25%)	0 (NA%)	
Other	0 (0%)	47 (7.5%)	11 (21%)	2 (3.4%)	2 (5.0%)	1 (20%)	37 (54%)	8 (21%)	3 (75%)	0 (NA%)	

[1] n (%); Median (IQR)
[2] Kruskal-Wallis rank sum test

Table 3 contains variables related to oxygen levels, pressure, and gestational age, comparing the overall cohort (N = 996) with subgroups based on tracheostomy placement (Trach 0, N = 850; Trach 1, N = 146). The Fraction of Inspired Oxygen (FiO2) at 36 weeks demonstrated a mean of 0.34 in the overall cohort, with a slight decrease in the tracheostomy group (0.32) and a notable increase in the FiO2 for infants with tracheostomy (0.49). Similarly, at 44 weeks, the FiO2 means remained consistent, but the tracheostomy group exhibited a slightly lower mean (0.44). For Peak Inspiratory Pressure at 36 and 44 weeks, the overall averages increased from 5 to 8 cmH2O, and 15 to 21 cmH2O, respectively. The tracheostomy group consistently demonstrated higher pressures at both time points compared to the non-tracheostomy group. Hospital Discharge Gestational Age showcased a mean of 53 weeks for the overall cohort, while the tracheostomy and non-tracheostomy groups had mean values of 49 and 80 weeks, respectively.

Table 4 provides an insightful overview of the maternal demographics. The ethnic distribution reveals that 741 mothers identified as Hispanic or Latino, while 8651 mothers identified as Not Hispanic or Latino. When examining maternal race a notable difference in representation between Hispanic and non-Hispanic individuals within each category. Furthermore, the severity distribution indicates the classification of mothers based on severity levels, with 27% categorized as severity level 1, 26% as severity level 2, and 46% as severity level 3 (most sever). The analysis suggests variations in severity levels between Hispanic and non-Hispanic individuals, highlighting potential disparities in maternal health outcomes.

Table 5 provides a comprehensive overview of key factors related to tracheostomy placement in a cohort of 996 infants. Among them, 850 infants did not require a tracheostomy, while 146 infants did. The gender distribution showed a predominance of males in both groups. Prenatal steroid administration was prevalent, with 87% of infants receiving steroids in the non-tracheostomy group and 94% in the tracheostomy group. Delivery methods varied, with 71% of infants in the non-tracheostomy group delivered via c-section, compared

to 79% in the tracheostomy group. Ventilation support levels at 36 weeks revealed a higher proportion of infants requiring support in the tracheostomy group. This trend continued at 44 weeks, indicating a significant association between tracheostomy placement and the need for prolonged ventilation. Additionally, medication for pulmonary hypertension at 36 and 44 weeks had a higher prevalence in the tracheostomy group. The severity of (BPD) also exhibited distinct patterns, with a higher proportion of severe cases in the tracheostomy group.

Methods

Initially, a carefully chosen subset of variables essential for the analysis is identified and renamed for clarity. To address missing values, the Multiple Imputation by Chained Equations (MICE) method is employed. This method involves generating five imputed datasets to ensure robustness and account for uncertainty in the imputation process. A strategic decision is made to split the data, excluding observations for the 44-week time point due to extensive missing data in each variable resulting from patient discharges between 36 and 44 weeks.

Afterwards, the imputation process focuses on the 36-week data, where each variable exhibits less than 20% missing values. To enhance reproducibility, a random seed is set. The imputed datasets are stored in a list for further analysis, and the mice::complete function is utilized to obtain the completed datasets for each imputation iteration. A visual representation of the imputed datasets is presented through a plot, offering insights into the distribution and patterns of imputed values. The imputation methods used are documented and can be inspected through the imp$method attribute. Ultimately, the imputed datasets are consolidated into a dataframe named "completed_data" using the complete function. This step aggregates the imputed values into a single dataset, ensuring readiness for subsequent analyses. The imputation process plays a crucial role in fortifying the dataset's completeness and reliability, laying a strong foundation for subsequent regression analyses, specifically lasso and logistic regression models.

Model Descriptions

The incorporation of the Lasso regularization in logistic regression adds a layer of sophistication to the modeling process. The Lasso's ability to induce sparsity in the model is particularly valuable in scenarios with a multitude of potential predictors. In the context of predicting tracheostomy placement, where various clinical, demographic, and physiological factors may influence the outcome, the Lasso helps in automatic variable selection. By shrinking some coefficients to exactly zero, it identifies the most influential predictors while disregarding less significant ones. This not only streamlines the model but also enhances its interpretability, providing clinicians with a focused set of factors that contribute significantly to the likelihood of tracheostomy placement in infants with BPD. Additionally, the Lasso's regularization properties aid in preventing overfitting, improving the model's generalizability to new data and making it a robust tool for predicting clinical outcomes in this medical context.

The lasso model performs a 10-fold cross-validated lasso logistic regression. This function takes a dataset (df) as input and returns a list of lasso coefficients for two models: one without interactions and another with main interactions. Cross-validation to select the optimal regularization parameter (lambda) is included in this process. The dataset is preprocessed by removing the "death" variable and creating matrix forms for ordered and non-ordered variables.The code applies the function to five imputed datasets, obtaining lasso coefficients for both models (resulting coefficients are combined into two matrices). The average and variance of these coefficients are then calculated for each model.Finally, the code uses the obtained average coefficients to calculate lasso scores on a long-format imputed dataset. It then fits logistic regression models using the lasso scores and predicts probabilities for tracheostomy outcomes. The commented-out section provides an additional step for two-way interactions, which can be activated based on the analysis requirements.

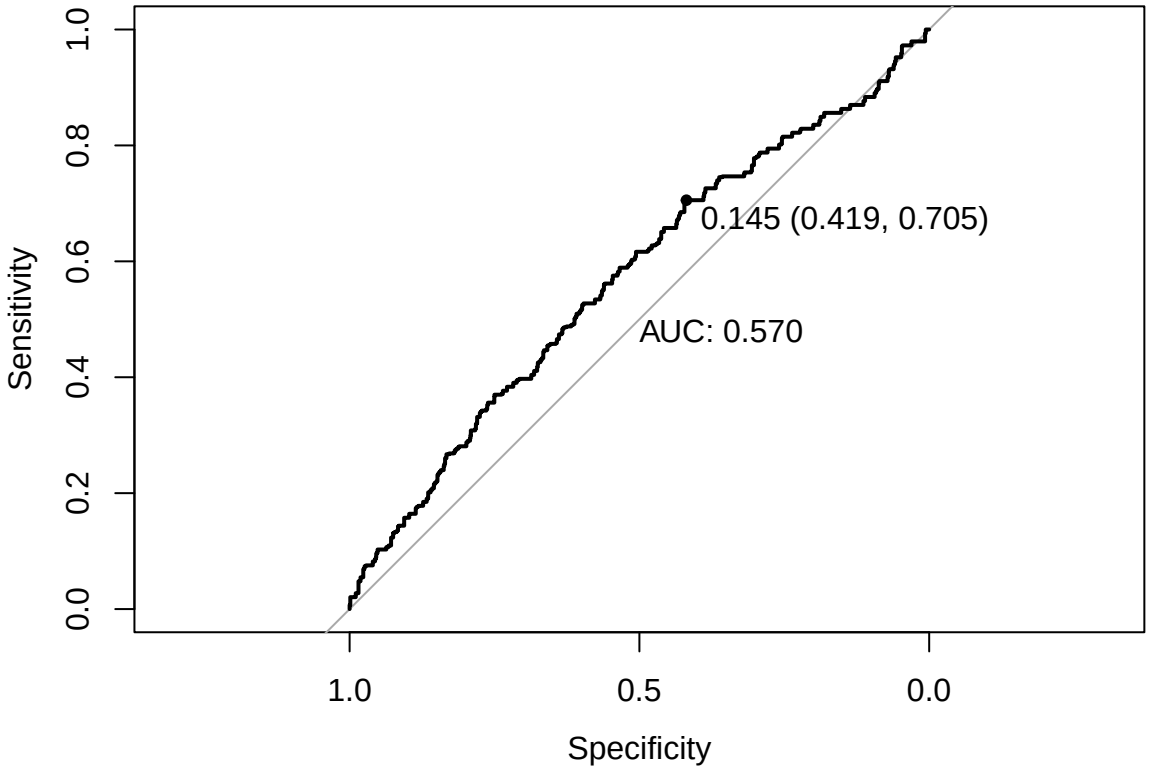
The choice of employing a logistic regression model for predicting tracheostomy placement in the context of Bronchopulmonary Dysplasia (BPD) is rooted in the specific characteristics of the dataset and the nature of the prediction task. Logistic regression is a well-suited statistical method for binary outcomes, making it particularly relevant for scenarios where we are interested in predicting the likelihood of an event occurring or not occurring. In this case, the binary outcome variable is tracheostomy placement (Trach), representing whether a patient requires this surgical intervention or not.

The logistic function, commonly known as the sigmoid function, is the fundamental component of logistic regression. It transforms linear combinations of predictor variables into probabilities, with a range between 0 and 1. This aligns perfectly with the nature of binary outcomes, where the goal is to estimate the probability of an event happening. For the present analysis, the logistic regression model is applied to the dataset (df) to predict the occurrence of tracheostomy placement based on the given set of predictor variables.Careful

consideration is given to the balance between model complexity and generalizability to ensure accurate predictions while avoiding overfitting to the specific characteristics of the training data.

The function first preprocesses the data by removing any irrelevant variables, specifically excluding the variable representing death. It then generates a model matrix from the data, considering linear predictors. The logistic regression models are fitted using the glm function with a binomial family. The coefficients of these models are extracted to understand the influence of the predictor variables on the probability of the binary outcome.The function returns a list containing the coefficients of both models and the fitted models themselves. This process is repeated for each imputed dataset, resulting in separate sets of coefficients for each dataset. The average coefficients and their variances are then calculated for a model with main interactions. Finally, the function applies the average coefficients to the long-format imputed data. The logistic model is then used to predict probabilities for the binary outcome.

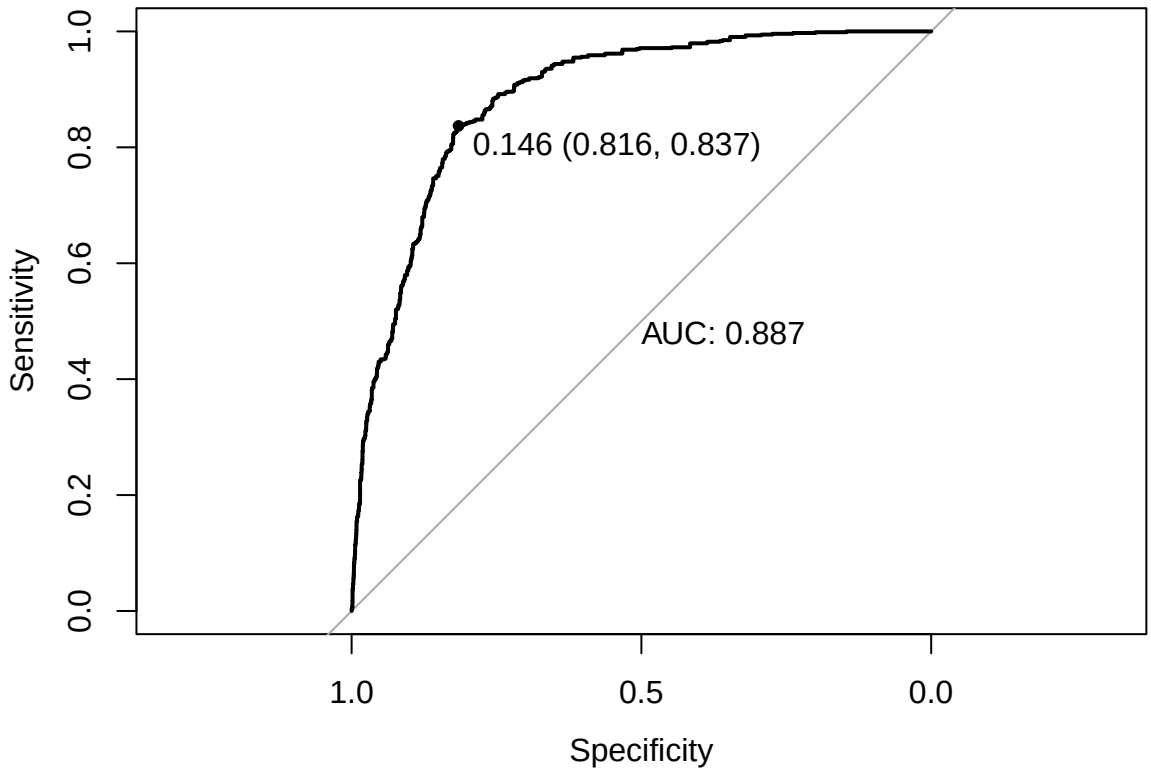
Statistical Analysis and Results




```
##     threshold specificity sensitivity
## 517 0.1391789   0.3190588   0.7506849
```

Table 8: Performance Metrics for Models

Metric	Lasso	Logistic
Sensitivity	0.7055	0.8370
Specificity	0.4188	0.8155
AUC	0.5703	0.8874
Best Threshold	0.1451	0.1460




```
##      threshold specificity sensitivity
## 1421 0.1922015   0.8541176   0.7506849
```

In summary, the Logistic Regression model with an AUC of 0.887 is considered better than the Lasso model with an AUC of 0.570. A higher AUC indicates that the Logistic Regression model has a better ability to distinguish between positive and negative instances.In addition a higher AUC value is generally associated with improved predictive performance. It suggests that the Logistic Regression model has a better balance between sensitivity and specificity, making it more reliable in making accurate predictions across different thresholds.The AUC summarizes the model's performance across various classification thresholds, providing a comprehensive view of its ability to trade off true positive rate and false positive rate. The higher AUC for Logistic Regression indicates a better performance across a range of decision thresholds compared to Lasso.

The Precision (PPV): The Logistic Regression model outperforms the Lasso model significantly in terms of Precision, indicating a higher proportion of true positive predictions among the positive predictions made by the model. The Negative Predictive Value (NPV): The Logistic Regression model also exhibits a higher Negative Predictive Value, indicating a better ability to correctly identify true negatives among the total negative predictions.Accuracy (Acc): The Logistic Regression model again demonstrates superior performance in terms of overall Accuracy, suggesting a higher proportion of correct predictions compared to the Lasso

Table 9: Coefficient Estimates for Models

Lasso	Logistic
-4.73	-4.59
0.00	-1.41
-0.41	-3.42
-1.02	-0.55
0.00	-1.43
0.00	-2.16
-0.16	0.14
1.07	-2.27
-0.25	-1.10
0.00	0.00
0.00	0.39
0.01	0.77
0.00	1.13
0.47	0.00
0.00	-0.02
0.01	0.01
0.00	-0.02
0.00	0.10
0.02	1.48
0.73	0.17
0.00	-0.37
0.00	0.00
0.00	0.08
0.00	0.18
0.00	0.00
0.00	-0.42
1.27	1.33
3.29	3.83
0.00	-0.01
0.03	0.08
0.17	0.34
-0.43	-1.27

Table 10: Zero Coefficients for Models

Lasso	Logistic
16	4

model.

In evaluating the performance metrics of the two models, Logistic regression emerges as the superior choice for predicting tracheostomy placement in infants with Bronchopulmonary Dysplasia. The Logistic model exhibited notable strengths across key metrics compared to the Lasso model. With a higher sensitivity of 0.8370, the Logistic model demonstrated a superior ability to correctly identify positive cases. Additionally, the Logistic model displayed a substantially higher specificity of 0.8155, indicating enhanced accuracy in predicting true negative cases. The Area Under the Curve (AUC), a comprehensive measure of discriminatory ability, further favored the Logistic model with a higher value of 0.8874 compared to the Lasso model's 0.5703. While both models had similar threshold values, the combination of superior sensitivity and specificity positions the Logistic regression model as the more effective and reliable choice for predicting tracheostomy placement in this specific medical context.

From coefficients table we can observe large differences between the estimated coefficients across both models. Most of the non-zero coefficients for the model came from the Lasso model. Analyzing the coefficients table reveals substantial disparities in the estimated coefficients between the two models, with a notable observation of distinctive patterns. The Lasso regression model predominantly contributes to the identification of non-zero coefficients, showcasing a more extensive utilization of predictors in the model formulation. This implies that the Lasso model retains a larger set of features with non-zero coefficients, suggesting a broader inclusion of variables in its predictive framework. In contrast, the Logistic regression model, while likely incorporating fewer predictors with non-zero coefficients, emphasizes a more selective and refined set of variables.

Discussion

The comparison between the Lasso and Logistic Regression models reveals compelling evidence supporting the superiority of the Logistic Regression model in predicting the outcome of interest of what regression would be best to help predict the time of the tracesotomy placement for infants. Several key performance metrics and the Area Under the Receiver Operating Characteristic curve (AUC-ROC) were evaluated to assess the models' effectiveness. In real-world scenarios, such as clinical settings or decision-making processes, where accurate predictions are critical, the Logistic Regression model's reliability in distinguishing between positive and negative instances positions it as a valuable tool. The interpretability of Logistic Regression coefficients further facilitates the translation of model insights into actionable knowledge, contributing to its generalizability and practical utility in addressing similar classification tasks across diverse contexts. As with any predictive model, ongoing validation and testing on new data will be essential to ensure its continued effectiveness in real-world applications.

Limitations

The analysis and study of tracheostomy placement for infants faces several limitations that impact the comprehensiveness of its findings. One notable constraint is the absence of precise information regarding the timing of tracheostomy placement, hindering timelines critical for clinical decision-making. This limitation obscures potential insights into developmental stages where tracheostomies might be most efficacious. Another limitation arises from the exclusion of infants with genetic anomalies, particularly heart disease, a significant contributor to infant mortality. By not considering this subset, the study might overlook essential factors influencing both the occurrence of tracheostomy placement and broader health outcomes.

Furthermore,there is a methodological limitation due to the lack of separation in the analysis between 36 and 44 weeks, preventing identification of developmental changes and/or patterns during these gestational periods. This can obscure fine differences in predictive variables and tracheostomy outcomes at distinct gestational ages. The study may fail to capture the broader spectrum of mortality risks, hindering intervention strategies for infants in a specific population. Incorporating a broader consideration of mortality data would enhance the study's utility and provide a more holistic perspective on the complexities surrounding both tracheostomy placement and mortality in infants.

References:

"Predicting the need for tracheostomy in infants with severe bronchopulmonarydysplasia" (Robin McKinney, MD,Jon Levin, MD)-2023

Code Appendix

```
library(knitr)
library(kableExtra)
library(mvtnorm)
library(glmnet)
library(leaps)
library(broom)
library(tinytex)
library(dplyr)
library(tidyr)
library(ggplot2)
library(gtsummary)
library(tidyverse)
library(GGally)
library(knitr)
library(mice)
library(naniar)
library(HDSinRdata) # version 0.1.0
library(readxl)
knitr::opts_chunk$set(echo = FALSE,
                      message = FALSE,
                      warning = FALSE)
#summary(cars)
project2<-read.csv("C:/Users/CAU Student/Documents/GitHub/2550_Project2/project2.csv")
project2_codebook<-read_excel("C:/Users/CAU Student/Documents/GitHub/2550_Project2/project2_codebook.xls

# Find duplicate rows
duplicates <- project2[duplicated(project2) | duplicated(project2, fromLast = TRUE), ]

# Display duplicate rows
#print(duplicates)
row_to_remove <-c(790,791,792)
project2<-project2[-row_to_remove, ]

#change numeric to factors
project2[,c(2:4,9:15,17,21,23,27, 29:30)] <- lapply(project2[,c(2:4,9:15,17,21,23,27,29:30)], factor)

overall_miss<-overall_missing<-miss_var_summary(project2)
knitr::kable(overall_miss, caption = "Table 1: Variables Missing", col.names = c("Variable name", "# Mis

#Look at missing data for each parent
pct_na_r <- rowSums(is.na(project2)) / ncol(project2) * 100
row_na <- data.frame(record_id = project2$record_id, pct_na = pct_na_r)
row_na <- row_na[row_na$pct_na > 35,]
#print(row_na)

#Drop Race, Censored, and other variables forr multiple imputatation
#Create severity of BPD variable indicator at week 36
project2<-project2%>% mutate(severity = case_when(
  (ventilation_support_level.36 == 0 |
     (ventilation_support_level.36 == 1  & inspired_oxygen.36 < 0.22)) ~ 'Mild',
  ((ventilation_support_level.36 == 2 & inspired_oxygen.36 <= 0.21) |
```

```
     (ventilation_support_level.36 == 1 & inspired_oxygen.36 < 0.30 & inspired_oxygen.36 >= 0.22)) ~ 'Mo
   ((ventilation_support_level.36 == 2 & inspired_oxygen.36 > 0.21) |
      (ventilation_support_level.36 == 1 & inspired_oxygen.36 >= 0.30)) ~ 'Severe'))
# which(is.na(project2$severity))
# sum(is.na(project2$severity))

baseline_char<-project2%>%
  select(center, mat_race,mat_ethn,del_method,gender,sga, any_surf,Trach,Death) %>%
  tbl_summary(missing = "no") %>%
  add_n(statistic = "{n_miss} ({p_miss}%)") %>%
  modify_header(n = "N Missing") %>%
  knitr::kable(caption = "Summary of Demographics and Infant Delivery", col.names = c("Characteristic",
  kableExtra::kable_styling(latex_options = c("striped","scale_down_table"),stripe_color = "gray!15")
baseline_char


baseline_info<-project2%>%
  select(prenat_ster,com_prenat_ster,mat_chorio,weight_today.36,ventilation_support_level.36,inspired_o
  tbl_summary(missing = "no",) %>%
  add_n(statistic = "{n_miss} ({p_miss}%)") %>%
  modify_header(n = "N Missing") %>%
  knitr::kable(caption = "Summary of Infant Information", col.names = c("Characteristic", "N Missing",
  kableExtra::kable_styling(latex_options = c("striped","scale_down_table"),stripe_color = "gray!15")
baseline_info
#Summary of variables stratified by Tracheostomy
table1<-project2 %>%
  select(gender, bw,blength,birth_hc,weight_today.36, weight_today.44,Trach) %>%
    tbl_summary(
      by = Trach,
      type = all_continuous() ~ "continuous2",
      statistic = all_continuous() ~ c("{mean}"),
      missing_text = "Missing",
      label = list(bw = "Average Birth Weight {g}",
                   blength="Average Height of Infant {cm}",
                   birth_hc="Average Circumference of Infant Head {cm}",
                  weight_today.36 = "Average Weight at 36 Weeks {g}",
                  weight_today.44 = "Average Weight at 44 Weeks {g}")) %>%
      add_overall() %>%
      modify_header(label ~ "**Variable**") %>%
      modify_caption("** Average Birth and Infants Weights and Heights**") %>%
      modify_spanning_header(c("stat_1", "stat_2") ~ "**Trach**") %>%
      bold_labels()
tbl_butchered1<-table1%>%
  tbl_butcher()
tbl_butchered1

table2<-project2 %>%
  select(Trach,inspired_oxygen.36,inspired_oxygen.44,p_delta.36,p_delta.44,hosp_dc_ga,peep_cm_h2o_modifi
    tbl_summary(
      by = Trach,
      type = all_continuous() ~ "continuous2",
      statistic = all_continuous() ~ c("{mean}"),
      missing_text = "Missing",
```

```
      label = list(inspired_oxygen.36 = "Fraction of Inspired Oxygen at 36 weeks",
                   inspired_oxygen.44= "Fraction of Inspired Oxygen at 44 weeks",
                   p_delta.36="Peak Inspiratory Pressure (cmH2O) at 36 weeks",
                   p_delta.44="Peak Inspiratory Pressure (cmH2O) at 44 weeks",
                   peep_cm_h2o_modified.36="Positive and exploratory pressure (cm H2O) at 36 weeks",peep
                   hosp_dc_ga="Hospital Discharge Gestational Age")) %>%
      add_overall() %>%
      modify_header(label ~ "**Variable**") %>%
      modify_caption("**Average of Oxygen, Pressure, and Gestiational Age**") %>%
      modify_spanning_header(c("stat_1", "stat_2") ~ "**Trach**") %>%
      bold_labels()

tbl_butchered2<-table2%>%
  tbl_butcher()
tbl_butchered2

# name the values for the mother's race and the mother's ethnicity
project2$mother_ethn <- ifelse(project2$mat_ethn == 1,
                               "Hispanic or Latino","Not Hispanic or Latino")
project2$mother_race<-
  case_when(
    project2$mat_race == 0 ~ "American Indian or Alaskan Native",
    project2$mat_race == 1 ~ "Asian",
    project2$mat_race == 3 ~ "Black or African American",
    project2$mat_race == 4 ~ "Native Hawaiian or Other Pacific Islander",
    project2$mat_race == 5 ~ "White",
    project2$mat_race == 2 ~ "Other")

project2$severity<-
  case_when(
    project2$severity == "Mild"~ 1,
    project2$severity == "Moderate"~ 2,
    project2$severity == "Severe"~ 3)
project2$severity<-as.factor(project2$severity)

# table displaying the summary of the mother's race and ethnicity
table3<-project2 %>%
  select(mother_race, mother_ethn,severity) %>%
  tbl_summary(
    by = mother_ethn, missing_text = "Missing",
    label = list(mother_race = "Race of Mother")) %>%
  add_overall() %>%
  modify_header(label ~ " ") %>%
  modify_caption(" Maternal Demographics") %>%
  modify_spanning_header(c("stat_1", "stat_2") ~ "**Ethnicity of Mother**") %>%
  bold_labels()

tbl_butchered3<-table3%>%
  tbl_butcher()


# average tracheostomy at discharge and death before discharge
table4<-project2 %>%
```

```
  select(Trach, gender, prenat_ster,del_method,ventilation_support_level.36,ventilation_support_level_m
  tbl_summary(
    by = Trach,
    type = all_continuous() ~ "continuous2",
    statistic = all_continuous() ~ c("{median}"),
    missing_text = "Missing",
    label = list(prenat_ster = "Prenatal Steroids",
                 del_method= "Delivery Method",
                 ventilation_support_level.36="Ventilation support level at 36 weeks",
                 ventilation_support_level_modified.44="Ventilation support level at 44 weeks",
                 com_prenat_ster = "Completed Prenatal Steroids",
                 mat_chorio = "Maternal Chorioamnionitis",
                 gender = "Average #",
                 severity= "Severity of BPD",
                 med_ph.36="Medication for Pulmonary Hypertension at 36 weeks",
                 med_ph.44="Medication for Pulmonary Hypertension at 44 weeks",
                 center= "Medical Center")) %>%
  add_overall() %>%
  modify_header(label ~ " ") %>%
  modify_caption("** Summary of Steroid Status, Chorioamnionitis,Severity, Ventilation**") %>%
  modify_spanning_header(c("stat_1", "stat_2") ~ "**Trach**") %>%
  bold_labels()

tbl_butchered4<-table4%>%
  tbl_butcher()
tbl_butchered4
plt1<-ggpairs(project2 %>% select(weight_today.36,weight_today.44,blength,bw,ga)%>%na.omit(),
lower = list(continuous = wrap("smooth", alpha = 0.7, size=0.5)))+
theme_bw()
plt1

plt2<-ggpairs(project2 %>% select(weight_today.36,weight_today.44,p_delta.36,p_delta.44,peep_cm_h2o_mod
lower = list(continuous = wrap("smooth", alpha = 0.7, size=0.5)))+
theme_bw()
plt2
project2$center[project2$center == 21] = 20

project2 %>%
  group_by(center, Trach) %>%
  summarise(count = n()) %>%
  mutate(freq = count / sum(count)) %>%
  ggplot(aes(x = center, y = freq, fill = as.factor(Trach))) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Center", y = "Proportion", fill = "Outcome", title = "Proportion of Outcomes by Center") +
  theme_minimal()


project2 %>%
tbl_summary(include = -c(record_id, mat_ethn),
by = center,
missing = "no")%>%
add_p() %>%
as_kable_extra(booktabs = TRUE,
```

```
caption = "Summary Statistics by Center") %>%
kableExtra::kable_styling(font_size = 9,
latex_options = c("repeat_header", "HOLD_position", "scale_down"))

# Select variables and rename
# project2<- project2 %>% select(c(center,bw,ga,blength,birth_hc,del_method,prenat_ster,com_prenat_ster
#Create imputation for variables at 44 weeks including weight, Fraction of Inspired Oxygen,Peak Inspira
#
# #Perform multiple imputation (generating 5 imputed datasets)
# # Set a random seed for reproducibility
# project2<- project2[,-c(15,28,30,31,32,33)]
# trache_mice<- mice(project2, m = 5, print = FALSE, seed = 10)
# # Store each imputed data set
# trache_imputed<- list()
# for (i in 1:5){
#   trache_imputed[[i]]<-complete(trache_mice,i)
# }
# #saveRDS(trache_mice,file="trache_mice_final_1.RDS")
# #trache_mice<- readRDS("trache_mice_final_1.RDS")
# #plot(trache_mice)
# #trache_mice$method
# #View the imputed datasets in the
# completed_data<-complete(trache_mice)
#columns_to_remove <-c(30,31,32)
#completed_data<-completed_data[,-columns_to_remove ]

#overall_missing1<-miss_var_summary(completed_data)
#overall_missing1

# Subset the dataframe to only include the specified variables
wk44<- c("inspired_oxygen.44", "p_delta.44", "weight_today.44",
               "peep_cm_h2o_modified.44",
               "ventilation_support_level_modified.44", "med_ph.44")

week44vars<-project2[wk44]

# Number of NAs in each row for the specified variables
na_count_per_row <- rowSums(is.na(week44vars))


project2$center[project2$center == 21] = 20

# Keep rows where not all of the specified variables are NA
wk36<-  project2 %>% dplyr::select(-c(wk44,mother_ethn,mother_race,hosp_dc_ga))
project2 <- project2[na_count_per_row != ncol(week44vars), ]
mice_mod<- mice(project2, m = 5, meth = 'pmm', seed = 123)
mice_mod_36<- mice(wk36, m = 5, meth = 'pmm', seed = 123)

# List of Imputed Set
p2_imp = vector("list",5)
for (i in 1:5){
  imp = complete(mice_mod,i)
  p2_imp[[i]] = cbind(imp)
```

```
}

# 36 Week Data
data_36 = vector("list",5)
for (i in 1:5){
  imp = complete(mice_mod_36,i)
  data_36[[i]] = cbind(imp)
}
lasso <- function(df) {
  #Runs 10-fold CV for lasso and returns corresponding coefficients
  #@param df, data set
  #@return coef, coefficients for minimum cv error

   #grid range for lambda
  grid<- 10^ seq (10 , -2, length = 100)
  df<- imp[,c(-1,-24)] #remove death

  # Matrix form for ordered variables
  x.ord.2<- model.matrix(Trach~.^2, data =df)[,-c(22)]
  x.ord <- model.matrix(Trach~., data = df)[,-c(22)]
  y.ord <- df$Trach
#   # Assuming you have a data frame or matrix named 'data'
# # Subsample 'y' to match the number of rows in 'x'
#   #set.seed(123)  # for reproducibility
#
  # Generate folds
  k <- 10
  set.seed(1) # consistent seeds between imputed data sets
  folds <- sample(1:k, nrow(df), replace=TRUE)

  # Lasso model without interactions
  lasso_mod1<-cv.glmnet(x.ord, y.ord, nfolds = 10, foldid = folds, lambda = grid,
                        alpha = 1, family = "binomial")
  lasso_mod_bl<-glmnet(x.ord, y.ord, nfolds = 10, foldid = folds,
                        alpha = 1,lambda=lasso_mod1$lambda.min, family = "binomial")
  #Lasso model with interactions
  lasso_mod2<-cv.glmnet(x.ord.2, y.ord, nfolds = 10, foldid = folds,lambda = grid,
                        alpha = 1, family = "binomial")
  lasso_mod_bl2<-glmnet(x.ord.2, y.ord, nfolds = 10,foldid = folds,
                        alpha = 1,lambda =lasso_mod2$lambda.min, family = "binomial")

   # Get coefficients
  coef1<- coef(lasso_mod_bl)
  coef2<- coef(lasso_mod_bl2)

  #return best model
  lasso_models_coef<- list()
  lasso_models_coef<- list(coef1, coef2, lasso_mod_bl,lasso_mod_bl2)
  return(lasso_models_coef)
}

# Find average lasso coefficients ovetr imputed datasets
lasso_coef1 <- lasso(imp[[1]])
```

```r
lasso_coef2 <- lasso(imp[[2]])
lasso_coef3 <- lasso(imp[[3]])
lasso_coef4 <- lasso(imp[[4]])
lasso_coef5 <- lasso(imp[[5]])

lasso_coef_all1<- cbind(lasso_coef1[[1]], lasso_coef2[[1]], lasso_coef3[[1]],lasso_coef4[[1]], lasso_coe
lasso_coef_all2<- cbind(lasso_coef1[[2]], lasso_coef2[[2]], lasso_coef3[[2]],lasso_coef4[[2]], lasso_coe
#
#for the model with  main interactions
avg_coefs_lasso1<- apply(lasso_coef_all1, 1, mean)
var_coefs_lasso1<- apply(lasso_coef_all1, 1, var)
#for the model with two-way interactions
avg_coefs_lasso2<- apply(lasso_coef_all2, 1, mean)
var_coefs_lasso2<- apply(lasso_coef_all2, 1, var)
#
# Find predicted probabilities on long imputed data (no rounding applied in this case!)
trach_df_long<-complete(mice_mod_36,action="long")
subset_long<- trach_df_long[,-c(1,2,3,26)]
x_vars <- model.matrix(Trach~. , subset_long)
subset_long$lasso_score<- x_vars%*% (avg_coefs_lasso1)
mod_lasso<- glm(Trach~lasso_score, data = subset_long, family = "binomial")
predict_probs_lasso<- predict(mod_lasso, type="response")
#
#
#for two way interactions
# x_vars2<-model.matrix(Trach~(.)^2, subset_long[,-30])
# subset_long$lasso_score2<- x_vars2 %*% (avg_coefs_lasso2)
# mod_lasso2<- glm(Trach~lasso_score2, data = subset_long, family = "binomial")
# predict_probs_lasso2 <- predict(mod_lasso2, type="response")
#Binomial logistic regression
logistic <- function(df) {
  # Assuming 'imp' is a data frame with columns, and you want to use 'df' instead of 'imp'
  df <- imp[, -c(1, 24)]  # Remove columns 1 and 25

  # Create model matrices for linear and quadratic terms
  x.ord <- model.matrix(Trach ~ ., data = df)[, -c(1)]      # Linear terms
  x.ord.2 <- model.matrix(Trach ~ .^2, data = df)[, -c(1)]   # Quadratic terms
  y.ord <- df$Trach

  # Fit logistic models
  logistic_mod <- glm(y.ord ~ x.ord, family = "binomial")
  logistic_mod1 <- glm(y.ord ~ x.ord.2, family = "binomial")

  # Get coefficients
  coef1_log <- coef(logistic_mod)
  coef2_log <- coef(logistic_mod1)

  # Return the models and coefficients
  logistic_models_coef <- list(coef1_log, coef2_log, logistic_mod, logistic_mod1)
  return(logistic_models_coef)
}

# Assuming 'imp' is a list of data frames
```

```r
# Find average logistic coefficients over imputed datasets
log_coef1<-logistic(imp[[1]])
log_coef2<- logistic(imp[[2]])
log_coef3<- logistic(imp[[3]])
log_coef4<- logistic(imp[[4]])
log_coef5<- logistic(imp[[5]])
log_coef_all1<- cbind(log_coef1[[1]], log_coef2[[1]], log_coef3[[1]],log_coef4[[1]], log_coef5[[1]])
log_coef_all2<- cbind(log_coef1[[2]], log_coef2[[2]], log_coef3[[2]],log_coef4[[2]], log_coef5[[2]])
#for the model with  main interactions
avg_coefs_log1<- apply(log_coef_all1, 1, mean)
var_coefs_log1<- apply(log_coef_all1, 1, var)
#for the model with two-way interactions
# avg_coefs_log2<- apply(log_coef_all2, 1, mean)
# var_coefs_log2<- apply(log_coef_all2, 1, var)

# Find predicted probabilities on long imputed data (no rounding applied in this case!)
trach_df_long<- complete(mice_mod_36,action="long")
subset_long1<- trach_df_long[,-c(1,2,3,26)]
x_vars_log<- model.matrix(Trach~ .,subset_long1)
avg_coefs_log1[is.na(avg_coefs_log1)] <- 0
subset_long1$logistic_log<-x_vars_log %*% avg_coefs_log1
mod_logistic<- glm(Trach~logistic_log, data = subset_long1, family = "binomial")
predict_probs_logistic<- predict(mod_logistic, type="response")
library(pROC)
lasso_plot<-ggplot() +
  geom_histogram(aes(x=predict_probs_lasso, fill=as.factor(mod_lasso$y)),
                 bins=30) +
  scale_fill_discrete(name="Tracheostomy Presence") +
  labs(x="Predicted Probabilities Lasso", y="Count")
roc_mod_lasso<- roc(predictor=predict_probs_lasso,
                response=as.factor(mod_lasso$y),
                levels = c(0,1), direction = "<")
plot(roc_mod_lasso, print.auc=TRUE, print.thres = TRUE)
roc_vals_lasso<- coords(roc=roc_mod_lasso, x = "all")
#head(roc_vals_lasso)
roc_vals_lasso[roc_vals_lasso$sensitivity > 0.75, ] %>% tail(n=1)

pred_ys <- ifelse(predict_probs_lasso > 0.139, 1, 0)
pred_ys <- factor(pred_ys, levels = c("0", "1"))
tab_outcome <- table(mod_lasso$y, pred_ys)
#tab_outcome
#sens <- tab_outcome[2,2]/(tab_outcome[2,1]+tab_outcome[2,2])
#spec <- tab_outcome[1,1]/(tab_outcome[1,1]+tab_outcome[1,2])
ppv <- tab_outcome[2,2]/(tab_outcome[1,2]+tab_outcome[2,2])
npv <- tab_outcome[1,1]/(tab_outcome[1,1]+tab_outcome[2,1])
acc <- (tab_outcome[1,1]+tab_outcome[2,2])/sum(tab_outcome)
vals <- data.frame(Measures = c("PPV", "NPV", "Acc"),
           Values = round(c(ppv, npv, acc),3))


log_plot<-ggplot() +
  geom_histogram(aes(x=predict_probs_logistic, fill=as.factor(mod_logistic$y)),
                 bins=30) +
```

```
  scale_fill_discrete(name="Tracheostomy Presence") +
  labs(x="Predicted Probabilities Logistic", y="Count")
roc_mod_logistic<- roc(predictor=predict_probs_logistic,
               response=as.factor(mod_logistic$y),
               levels = c(0,1), direction = "<")
plot(roc_mod_logistic, print.auc=TRUE, print.thres = TRUE)
roc_vals_logistic<- coords(roc=roc_mod_logistic, x = "all")
#head(roc_vals_logistic)
roc_vals_logistic[roc_vals_logistic$sensitivity > 0.75, ] %>% tail(n=1)
pred_ys_log<- ifelse(predict_probs_logistic > 0.190, 1, 0)
pred_ys_log<- factor(pred_ys_log, levels = c("0", "1"))
tab_outcome_log<-table(mod_logistic$y, pred_ys_log)
#tab_outcome_log
#sens <- tab_outcome[2,2]/(tab_outcome[2,1]+tab_outcome[2,2])
#spec <- tab_outcome[1,1]/(tab_outcome[1,1]+tab_outcome[1,2])
ppv_log<- tab_outcome_log[2,2]/(tab_outcome_log[1,2]+tab_outcome_log[2,2])
npv_log<- tab_outcome_log[1,1]/(tab_outcome_log[1,1]+tab_outcome_log[2,1])
acc_log<- (tab_outcome_log[1,1]+tab_outcome_log[2,2])/sum(tab_outcome_log)
vals_log<- data.frame(Measures = c("PPV", "NPV", "Acc"), Values = round(c(ppv_log, npv_log, acc_log),3))


roc_vals_lass<- pROC::coords(roc=roc_mod_lasso, x = "best")
roc_vals_log<- pROC::coords(roc=roc_mod_logistic, x = "best")
library(pROC)



df_metrics <- data.frame(
  Metric = c("Sensitivity", "Specificity", "AUC", "Best Threshold"),
  Lasso = c(
    roc_vals_lass$sensitivity,
    roc_vals_lass$specificity,
    auc(roc_mod_lasso),
    (roc_vals_lass$threshold)
  ),
  Logistic = c(
    roc_vals_log$sensitivity,
    roc_vals_log$specificity,
    auc(roc_mod_logistic),
    (roc_vals_log$threshold)
  )
)


df_metrics %>% knitr::kable(caption = "Performance Metrics for Models", digits = 4)
#TABLES OF COEFFICIENTS

lass1 <- avg_coefs_lasso1
log<- avg_coefs_log1

coefficients_df <- data.frame(
  Lasso = c(lass1, rep(NA, max(length(lass1), length(log)))),
  Logistic = c(log, rep(NA, max(length(lass1), length(log))))
```

```
)

coefficients_df <- round(coefficients_df[1:length(log),], 2)

# Print the coefficients table
coefficients_df %>% knitr::kable(caption = "Coefficient Estimates for Models")

# Create a data frame with zero coefficients count
df2 <- coefficients_df %>%
  summarise_all(~sum(. == 0))

# Print the zero coefficients table
table0<- df2 %>%
  knitr::kable(col.names = c("Lasso", "Logistic"), caption = "Zero Coefficients for Models")%>%
  kable_styling(latex_options = c("striped"), full_width = FALSE)
table0




# bckwd_plot<-ggplot() +
#   geom_histogram(aes(x=predict_probs_backwardstep, fill=as.factor(mod_logistic$y)),
#                  bins=30) +
#   scale_fill_discrete(name="Tracheostomy Presence") +
#   labs(x="Predicted Probabilities Bakwards Step", y="Count")
# roc_mod_back<- roc(predictor=predict_probs_backwardstep,
#                response=as.factor(mod_logistic$y),
#                levels = c(0,1), direction = "<")
# plot(roc_mod_back, print.auc=TRUE, print.thres = TRUE)
# roc_vals_back<- coords(roc=roc_mod_back, x = "all")
# head(roc_vals_logistic)
# roc_vals_logistic[roc_vals_logistic$sensitivity > 0.75, ] %>% tail(n=1)
# pred_ys_back<- ifelse(predict_probs_backwardstep > 0.289, 1, 0)
# pred_ys_back<- factor(pred_ys_log, levels = c("0", "1"))
# tab_outcome_log<-table(mod_logistic$y, pred_ys_log)
# tab_outcome_log
# #sens <- tab_outcome[2,2]/(tab_outcome[2,1]+tab_outcome[2,2])
# #spec <- tab_outcome[1,1]/(tab_outcome[1,1]+tab_outcome[1,2])
# ppv_back<- tab_outcome_log[2,2]/(tab_outcome_log[1,2]+tab_outcome_log[2,2])
# npv_back<- tab_outcome_log[1,1]/(tab_outcome_log[1,1]+tab_outcome_log[2,1])
# acc_back<- (tab_outcome_log[1,1]+tab_outcome_log[2,2])/sum(tab_outcome_log)
# vals_back<- data.frame(Measures = c("PPV", "NPV", "Acc"), Values = round(c(ppv_back, npv_back, acc_ba
```